Annals of Computer Science and Information Systems
Volume 2

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems

September 7–10, 2014. Warsaw, Poland

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

PTI

IEEE

Annals of Computer Science and Information Systems, Volume 2

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems

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

2014, Warszawa,
Polskie Towarzystwo
Informatyczne

2014, New York City,
Institute of Electrical and
Electronics Engineers

Annals of Computer Science and Information Systems, Volume 2

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems

ART: ISBN 978-83-60810-61-3, IEEE Catalog Number CFP1485N-ART
USB: ISBN 978-83-60810-57-6, IEEE Catalog Number CFP1485N-USB
WEB: ISBN 978-83-60810-58-3

ISSN 2300-5963
DOI 10.15439/978-83-60810-58-3

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

10662 Los Vaqueros Circle
Los Alamitos, CA 90720
USA

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

Cover:
Alisa Denisiuk,
Elbląg, Poland

Also in this series:

Volume 1: Position Papers of the 2013 Federated Conference on Computer Science and Information Systems (FedCSIS), **ISBN WEB: 978-83-60810-55-2, ISBN USB: 978-83-60810-56-9**
Volume 3: Position Papers of the 2014 Federated Conference on Computer Science and Information Systems, **ISBN WEB: 978-83-60810-60-6, ISBN USB: 978-83-60810-59-0**
Volume 4: Proceedings of the E2LP Workshop, **ISBN WEB: 978-83-60810-64-4, ISBN USB: 978-83-60810-63-7**

Dear Reader, it is our pleasure to present to you Proceedings of the 2014 Federated Conference on Computer Science and Information Systems (FedCSIS), which took place in Warsaw, Poland, on September 7–10, 2014. Each of the papers, found in this volume, was refereed by at least two referees and the acceptance rate of full (regular) papers was 34.88% (150 papers out of 430 submissions).

FedCSIS 2014 was organized by the Polish Information Processing Society (Mazowsze Chapter), Warsaw University of Technology, Wrocław University of Economics and Systems Research Institute Polish Academy of Sciences. FedCSIS was organized in technical cooperation with: IEEE Computer Society, IEEE Region 8, Computer Society Chapter Poland, Gdańsk Computer Society Chapter, Poland, Polish Chapter of the IEEE Computational Intelligence Society (CIS), ACM Special Interest Group on Applied Computing, International Federation for Information Processing (IFIP), European Alliance for Innovation (EAI), Łódź ACM Chapter, Informatics Europe, Asociación de Técnicos de Informática, Committee of the Computer Science of the Polish Academy of Sciences, Polish Society for Business Informatics, Polish Chamber of Information Technology and Telecommunications, Polish Chamber of Commerce for High Technology, Mazovia Cluster ICT and Eastern Cluster ICT Poland. Furthermore, the 9th International Symposium Advances in Artificial Intelligence and Applications (AAIA'14) was organized in technical cooperation with: International Rough Set Society, International Fuzzy Systems Association and Polish Neural Networks Society.

FedCSIS 2014 consisted of the following events:

- **AAIA'14—9th International Symposium Advances in Artificial Intelligence and Applications**
  - AIMA'14 – 4th International Workshop on Artificial Intelligence in Medical Applications
  - ASIR'14 – 4th International Workshop on Advances in Semantic Information Retrieval
  - CEIM'14 – 1st Complex Events and Information Modelling
  - WCO'14 – 7th Workshop on Computational Optimization
- **CSS—Computer Science & Systems**
  - CANA'14 – 7th Computer Aspects of Numerical Algorithms
  - MMAP'14 – 7th International Symposium on Multimedia Applications and Processing
  - ScoDiS-LaSCoG'14 – 2nd Workshop on Scalable Computing in Distributed Systems and 7th Workshop on Large Scale Computations on Grids
- **ECRM—Education, Curricula & Research Methods**
  - E2LP Workshop Application of Innovative Teaching Methods in Embedded Engineering
  - ISEC'14 – Information Systems Education & Curricula Workshop
- **iNetSApp—Innovative Network Systems and Applications**
  - EAIS'14 – Emerging Aspects in Information Security
  - SoFAST-WS'14 – 3rd International Symposium on Frontiers in Network Applications, Network Systems and Web Services
  - WSN'14 – 3rd International Conference on Wireless Sensor Networks
- **IT4MBS—Information Technology for Management, Business & Society**
  - ABICT'14 – 5th International Workshop on Advances in Business ICT
  - AITM'14 – 12th Conference on Advanced Information Technologies for Management
  - ISM'14 – 9th Conference on Information Systems Management
  - IT4L'14 – 3rd Workshop on Information Technologies for Logistics
  - KAM&AI4KM'14 – 20th Conference on Knowledge Acquisition and Management and 2nd Workshop on Artificial Intelligence for Knowledge Management
- **Joint Agent-oriented Workshops in Synergy (JAWS)**
  - MAS&M'14 – International Workshop on Multi-Agent Systems and Simulation
  - SEN-MAS'14 – 3rd International Workshop on Smart Energy Networks & Multi-Agent Systems
- **SSD&A—Software Systems Development & Applications**
  - ATSE'14—5th International Workshop Automating Test Case Design, Selection and Evaluation
  - MDASD'14 – 3rd Workshop on Model Driven Approaches in System Development

Furthermore, an AAIA'14 Data Mining Competition, focused on key risk factors for the Polish State Fire Service has been organized. This competition was an integral part of the 1st Complex Events and Information Modeling workshop, while its results constitute a separate section in these proceedings. Awards for the winners of the contest were sponsored by: Dituel Ltd. and F&K Consulting Engineers Ltd.

Each event constituting FedCSIS had its own Organizing and Program Committee. We would like to express our warmest gratitude to members of all of them for their hard work attracting and later refereeing 430 submissions.

FedCSIS 2014 was organized under the auspices of Prof. Lena Kolarska-Bobińska, Minister of Science and Higher Education, Dr Rafał Trzaskowski, Minister of Administration and Digitization, Prof. Michał Kleiber, President of the Polish Academy of Sciences, Major General Wiesław Leśniakiewicz, Chief Commandant of the State Fire Service, Prof. Hanna Gronkiewicz-Waltz, Mayor of the Capital City of Warsaw, Prof. Jan Szmidt, Rector of Warsaw University of Technology, Gen. Prof. Zygmunt Mierczyk, Rector of Military Technical Academy, and Prof. Andrzej Gospodarowicz, Rector of Wroclaw University of Economics.

FedCSIS was sponsored by the Ministry of Science and Higher Eduction, Intel, Orange Polska S.A. and Samsung.

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

Annals of Computer Science and Information Systems, Volume 2

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems (FedCSIS)

September 7–10, 2014. Warsaw, Poland

TABLE OF CONTENTS

4TH International Workshop on Artificial Intelligence in Medical Applications

4TH International Workshop on Advances in Semantic Information Retrieval

1ST Complex Events and Information Modelling

AAIA'14 Data Mining Competition at the Knowledge Pit

7TH Workshop on Computational Optimization

Computer Science & Systems

7TH Computer Aspects of Numerical Algorithms

7TH International Symposium on Multimedia Applications and Processing

2ND Workshop on Scalable Computing in Distributed Systems and 7TH Workshop on Large Scale Computations on Grids

Education, Curricula & Research Methods

3RD Information Systems Education & Curricula Workshop

Innovative Network Systems and Applications

1ST Workshop on Emerging Aspects in Information Security

3RD International Symposium on Frontiers in Network Applications, Network Systems and Web Services

3RD International Conference on Wireless Sensor Networks

Information Technology for Management, Business & Society

5TH International Workshop on Advances in Business ICT

12TH Conference on Advanced Information Technologies for Management

9TH Conference on Information Systems Management

3RD Workshop on Information Technologies for Logistics

20TH 20th Conference on Knowledge Acquisition and Management & 2ND Workshop on Artificial Intelligence for Knowledge Management

5TH International Workshop Automating Test Case Design, Selection and Evaluation

3RD Workshop on Model Driven Approaches in System Development

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 1–8

DOI: 10.15439/2014F503
ACSIS, Vol. 2

The Fog Computing Paradigm: Scenarios and Security Issues

Ivan Stojmenovic
SIT, Deakin University, Burwood, Australia
and
SEECS, University of Ottawa, Canada
Email: stojmenovic@gmail.com

Sheng Wen
School of Information Technology,
Deakin University,
220 Burwood Highway, Burwood, VIC, 3125, Australia
Email: wesheng@deakin.edu.au

***Abstract*—Fog Computing is a paradigm that extends Cloud computing and services to the edge of the network. Similar to Cloud, Fog provides data, compute, storage, and application services to end-users. In this article, we elaborate the motivation and advantages of Fog computing, and analyse its applications in a series of real scenarios, such as Smart Grid, smart traffic lights in vehicular networks and software defined networks. We discuss the state-of-the-art of Fog computing and similar work under the same umbrella. Security and privacy issues are further disclosed according to current Fog computing paradigm. As an example, we study a typical attack, man-in-the-middle attack, for the discussion of security in Fog computing. We investigate the stealthy features of this attack by examining its CPU and memory consumption on Fog device.**

***Index Terms*—Fog Computing, Cloud Computing, Internet of Things, Software Defined Networks.**

I. Introduction

CISCO recently delivered the vision of fog computing to enable applications on billions of connected devices, already connected in the Internet of Things (IoT), to run directly at the network edge [1]. Customers can develop, manage and run software applications on Cisco IOx framework of networked devices, including hardened routers, switches and IP video cameras. Cisco IOx brings the open source Linux and Cisco IOS network operating system together in a single networked device (initially in routers). The open application environment encourages more developers to bring their own applications and connectivity interfaces at the edge of the network. Regardless of Cisco's practices, we first answer the questions of what the Fog computing is and what are the differences between Fog and Cloud.

In Fog computing, services can be hosted at end devices such as set-top-boxes or access points. The infrastructure of this new distributed computing allows applications to run as close as possible to sensed actionable and massive data, coming out of people, processes and thing. Such Fog computing concept, actually a Cloud computing close to the 'ground', creates automated response that drives the value.

Both Cloud and Fog provide data, computation, storage and application services to end-users. However, Fog can be distinguished from Cloud by its proximity to end-users, the dense geographical distribution and its support for mobility [2]. We adopt a simple three level hierarchy as in Figure 1.

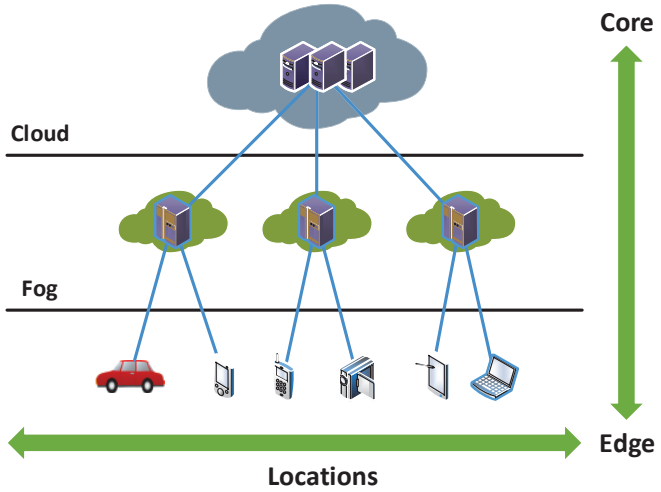


Fig. 1. Fog between edge and cloud.

In this framework, each smart thing is attached to one of Fog devices. Fog devices could be interconnected and each of them is linked to the Cloud.

In this article, we take a close look at the Fog computing paradigm. The goal of this research is to investigate Fog computing advantages for services in several domains, such as Smart Grid, wireless sensor networks, Internet of Things (IoT) and software defined networks (SDNs). We examine the state-of-the-art and disclose some general issues in Fog computing including security, privacy, trust, and service migration among Fog devices and between Fog and Cloud. We finally conclude this article with discussion of future work.

II. Why Do We Need Fog?

In the past few years, Cloud computing has provided many opportunities for enterprises by offering their customers a range of computing services. Current "pay-as-you-go" Cloud computing model becomes an efficient alternative to owning and managing private data centres for customers facing Web applications and batch processing [3]. Cloud computing frees the enterprises and their end users from the specification of many details, such as storage resources, computation limitation and network communication cost. However, this bliss becomes

978-83-60810-58-3/$25.00

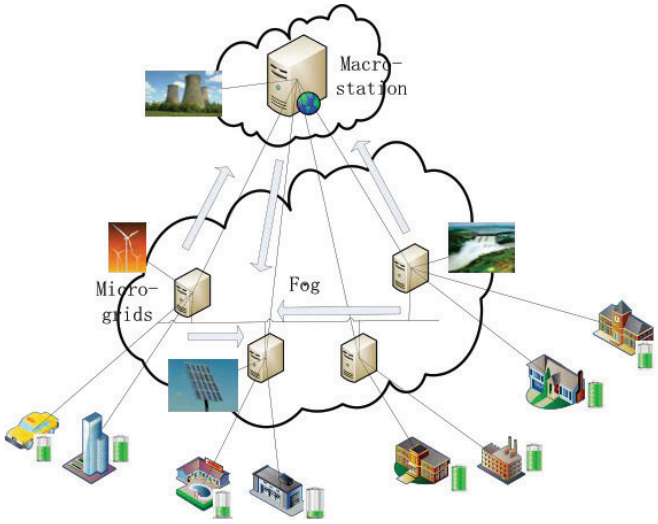


Fig. 2. Fog computing in smart grid.

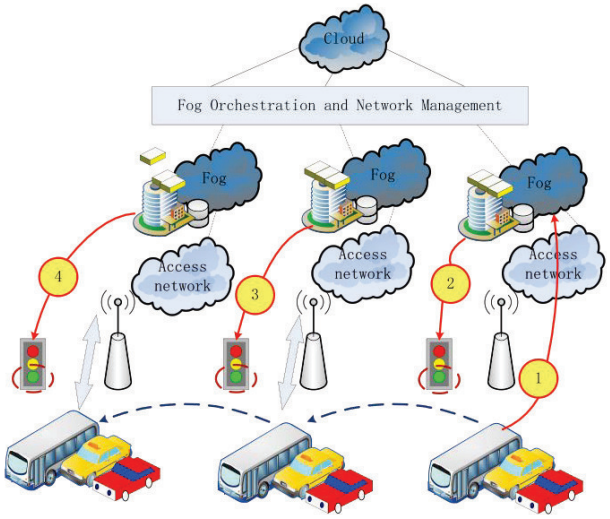


Fig. 3. Fog computing in smart traffic lights and connected vehicles.

a problem for latency-sensitive applications, which require nodes in the vicinity to meet their delay requirements [2]. When techniques and devices of IoT are getting more involved in people's life, current Cloud computing paradigm can hardly satisfy their requirements of mobility support, location awareness and low latency.

Fog computing is proposed to address the above problem [1]. As Fog computing is implemented at the edge of the network, it provides low latency, location awareness, and improves quality-of-services (QoS) for streaming and real time applications. Typical examples include industrial automation, transportation, and networks of sensors and actuators. Moreover, this new infrastructure supports heterogeneity as Fog devices include end-user devices, access points, edge routers and switches. The Fog paradigm is well positioned for real time big data analytics, supports densely distributed data collection points, and provides advantages in entertainment, advertising, personal computing and other applications.

III. What Can We Do With Fog?

We elaborate on the role of Fog computing in the following six motivating scenarios. The advantages of Fog computing satisfy the requirements of applications in these scenarios.

Smart Grid: Energy load balancing applications may run on network edge devices, such as smart meters and micro-grids [4]. Based on energy demand, availability and the lowest price, these devices automatically switch to alternative energies like solar and wind. As shown in Figure 2, Fog collectors at the edge process the data generated by grid sensors and devices, and issue control commands to the actuators [2]. They also filter the data to be consumed locally, and send the rest to the higher tiers for visualization, real-time reports and transactional analytics. Fog supports ephemeral storage at the lowest tier to semi-permanent storage at the highest tier. Global coverage is provided by the Cloud with business intelligence analytics.

Smart Traffic Lights and Connected Vehicles: Video camera that senses an ambulance flashing lights can automatically change street lights to open lanes for the vehicle to pass through traffic. Smart street lights interact locally with sensors and detect presence of pedestrian and bikers, and measure the distance and speed of approaching vehicles. As shown in Figure 3, intelligent lighting turns on once a sensor identifies movement and switches off as traffic passes. Neighbouring smart lights serving as Fog devices coordinate to create green traffic wave and send warning signals to approaching vehicles [2]. Wireless access points like WiFi, 3G, road-side units and smart traffic lights are deployed along the roads. Vehicles-to-Vehicle, vehicle to access points, and access points to access points interactions enrich the application of this scenario.

Wireless Sensor and Actuator Networks: Traditional wireless sensor networks fall short in applications that go beyond sensing and tracking, but require actuators to exert physical actions like opening, closing or even carrying sensors [2]. In this scenario, actuators serving as Fog devices can control the measurement process itself, the stability and the oscillatory behaviours by creating a closed-loop system. For example, in the scenario of self-maintaining trains, sensor monitoring on a train's ball-bearing can detect heat levels, allowing applications to send an automatic alert to the train operator to stop the train at next station for emergency maintenance and avoid potential derailment. In lifesaving air vents scenario, sensors on vents monitor air conditions flowing in and out of mines and automatically change air-flow if conditions become dangerous to miners.

Decentralized Smart Building Control: The applications of this scenario are facilitated by wireless sensors deployed to measure temperature, humidity, or levels of various gases in the building atmosphere. In this case, information can be exchanged among all sensors in a floor, and their readings can be combined to form reliable measurements. Sensors will use distributed decision making and activation at Fog devices to

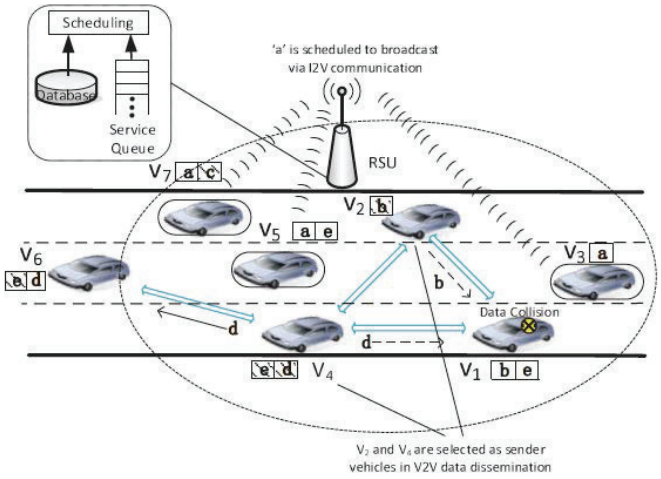


Fig. 4. Fog computing in SDN in vehicular networks [6].

react to data. The system components may then work together to lower the temperature, inject fresh air or open windows. Air conditioners can remove moisture from the air or increase the humidity. Sensors can also trace and react to movements (e.g, by turning light on or off). Fog devices could be assigned at each floor and could collaborate on higher level of actuation. With Fog computing applied in this scenario, smart buildings can maintain their fabric, external and internal environments to conserve energy, water and other resources.

IoT and Cyber-physical systems (CPSs): Fog computing based systems are becoming an important class of IoT and CPSs. Based on the traditional information carriers including Internet and telecommunication network, IoT is a network that can interconnect ordinary physical objects with identified addresses [5]. CPSs feature a tight combination of the system's computational and physical elements. CPSs also coordinate the integration of computer and information centric physical and engineered systems. IoT and CPSs promise to transform our world with new relationships between computer-based control and communication systems, engineered systems and physical reality. Fog computing in this scenario is built on the concepts of embedded systems in which software programs and computers are embedded in devices for reasons other than computation alone. Examples of the devices include toys, cars, medical devices and machinery. The goal is to integrate the abstractions and precision of software and networking with the dynamics, uncertainty and noise in the physical environment. Using the emerging knowledge, principles and methods of CPSs, we will be able to develop new generations of intelligent medical devices and systems, 'smart' highways, buildings, factories, agricultural and robotic systems.

Software Defined Networks (SDN): As shown in Figure 4, Fog computing framework can be applied to implement the SDN concept for vehicular networks. SDN is an emergent computing and networking paradigm, and became one of the most popular topics in IT industry [7]. It separates control and data communication layers. Control is done at a centralized server, and nodes follow communication path decided by the server. The centralized server may need distributed implementation. SDN concept was studied in WLAN, wireless sensor and mesh networks, but they do not involve multi-hop wireless communication, multi-hop routing. Moreover, there is no communication between peers in this scenario. SDN concept together with Fog computing will resolve the main issues in vehicular networks, intermittent connectivity, collisions and high packet loss rate, by augmenting vehicle-to-vehicle with vehicle-to-infrastructure communications and centralized control. SDN concept for vehicular networks is first proposed in [6].

IV. State-of-the-Art

A total of eight articles were identified on the concept of Fog computing [1], [2], [8], [9], [10], [11], [12], [13], [14]. There are some other concepts, not declared as Fog computing, fall under the same umbrella. We will also discuss these work in the subsection of similar work.

A. Related Work

K. Hong et al. proposed mobile Fog in [11]. This is a high level programming model for geo-spatially distributed, large-scale and latency-sensitive future Internet applications. Following the logical structure shown in Figure 1, low-latency processing occurs near the edge while latency-tolerant large-scope aggregation is performed on powerful resources in the core of the network (normally the Cloud). Mobile Fog consists of a set of event handlers and functions that an application can call. Mobile Fog model is not presented as generic model, but is built for particular application, while leaving out functions that deal with technical challenges of involved image processing primitives. Fog computing approach reduces latency and network traffic.

B. Ottenwalder et al. presented a placement and migration method for Cloud and Fog resources providers [13]. It ensures application-defined end-to-end latency restrictions and reduces the network utilization by planning the migration ahead of time. They also show how the application knowledge of the complex event processing system can be used to reduce the required bandwidth of virtual machines during their migration. Network intensive operators are placed on distributed Fog devices while computationally intensive operators are in the Cloud. Migration costs are amortized by selecting migration targets that ensure a low expected network utilization for a sufficiently long time. This work does not optimize workload mobility because Fog devices are also able to carry computationally intensive tasks. It also does not optimize the size of control information or mobility overhead, and does not describe network control policies for finding optimal paths for different applications.

In [11], K. Hong et al. proposed an opportunistic spatio-temporal event processing system that uses prediction-based continuous query handling. Their system predicts future query regions for moving consumers and starts processing events early so that the live situational information is available when

the consumer reaches the future location. Historical events for a location are processed before the mobile user arrives at that location. Live event processing begins at the moment the user arrives. To mitigate large speed of mobile user, authors propose using parallel resources to enable pipeline processing of future locations in several time steps looking ahead. Further, they proposed taking several predictions for each time step and opportunistically compute the events for all of those locations. When the user arrives at that time, the prediction among those that is closest to truth will be selected and its events returned.

J. Zhu et al. applied existing methods for web optimization in a novel manner [14]. Within Fog computing context, these methods can be combined with unique knowledge that is only available at the Fog devices. More dynamic adaptation to the user's conditions can also be accomplished with network edge specific knowledge. As a result, a user's Web page rendering performance is improved beyond that achieved by simply applying those methods at the Web server.

In the mobile Cloud concept [12], pervasive mobile devices share their heterogeneous resources and support services. Neighbouring nodes in a local network form a group called a local Cloud. Nodes share their resources with other nodes in the same local Cloud. A local resource coordinator serving as Fog device is elected from the nodes in each local Cloud. The work [12] proposed an architecture and mathematical framework for heterogeneous resource sharing based on the key idea of service-oriented utility functions. Normally heterogeneous resources are quantified in disparate scales, such as power, bandwidth and latency. However, authors in [12] present a unified framework where all these quantities are equivalently mapped to "time" resources. They formulate optimization problems for maximizing the sum and product of the utility functions, and solve them via convex optimization approaches.

The work [10] first reviews the reliability requirements of Smart Grid, Cloud, and sensors and actuators. This work then combines them towards reliable Fog computing. However, it only concludes that building Fog computing based projects is challenging and does not offer any novel concept for the reliability of the network of smart devices in the Fog computing paradigm.

B. Similar Work

BETaaS [15] proposed replacing Cloud as the resident for machine-to-machine applications by 'local Cloud' of gateways. The 'local Cloud' is composed of devices that provide smart things with connectivity to the Internet, such as smart phones, home routers and road-side units. This enables applications that are limited in time and space to require simple and repetitive interactions. It also enables the applications to respond in consistent manner.

Demand Response Management (DRM) is a key component in the smart grid to effectively reduce power generation costs and user bills. The work [16] addressed the DRM problem in a network of multiple utility companies and consumers where every entity is concerned about maximizing its own benefit. In their model, utility companies communicate with each other, while users receive price information from utility companies and transmit their demand to them. They propose a Stackelberg game [17] between utility companies and end-users to maximize the revenue of each utility company and the payoff of each user. Stackelberg equilibrium of the game has a unique solution. They develop a distributed algorithm which converges to the equilibrium with only local information available for both utility companies and end-users. Utility companies play a non-cooperative game. They inform users whenever they change price, and users then update their demand vectors and inform utility companies. This iterates until convergence. The main drawback of this algorithm is a significant communication overhead between users and utility companies. Though DRM helps to facilitate the reliability of power supply, the smart grid can be susceptible to privacy and security issues because of communication links between the utility companies and the consumers. They study the impact of an attacker who can manipulate the price information from the utility companies, and propose a scheme based on the concept of shared reserve power to improve the grid reliability and ensure its dependability.

The work [18] investigated how energy consumption may be optimized by taking into consideration the interaction between both parties. The energy price model is a function of total energy consumption. The objective function optimizes the difference between the value and cost of energy. The power supplier pulls consumers in a round-robin fashion, and provides them with energy price parameter and current consumption summary vector. Each user then optimizes his own schedule and reports it to the supplier, which in turn updates its energy price parameter before pulling the next consumers. This interaction between the power company and its consumers is modelled through a two-step centralized game, based on which the work [18] proposed the Game-Theoretic Energy Schedule (GTES) method. The objective of the GTES method is to reduce the peak to average power ratio by optimizing the users energy schedules.

The closest work for SDN in vehicular networks are several implementations in wireless sensor network and mesh networks [19], [20]. Moreover, B. Zhou et al. studied adaptive traffic light control for smoothing vehicles' travel and maximizing the traffic throughout for both single and multiple lanes [21], [22]. In addition, the work [23] proposed a three-tier structure for traffic light control. First, an electronic toll collection (ETC) system is employed for collecting road traffic flow data and calculating the recommended speed. Second, radio antennas are installed near the traffic lights. Third, road traffic flow information can be obtained by wireless communication between the antennas and ETC devices. A branch-and-bound-based real-time traffic light control algorithm is designed to smooth vehicles' travels.

V. Security and Privacy in Fog Computing

Security and privacy issues were not studied in the context of fog computing. They were studied in the context of smart grids [24] and machine-to-machine communications [25].

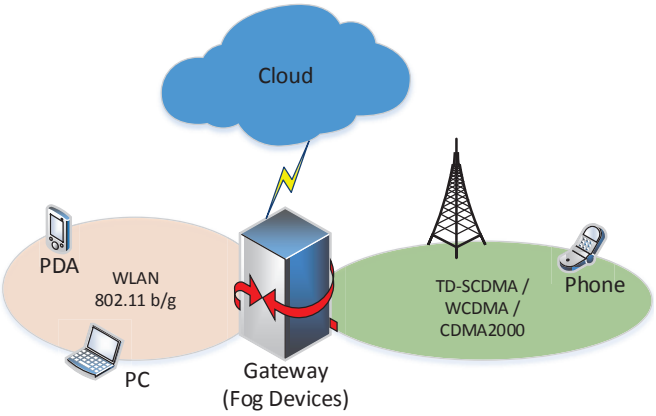


Fig. 5. A scenario for a man-in-the-middle attack towards Fog.

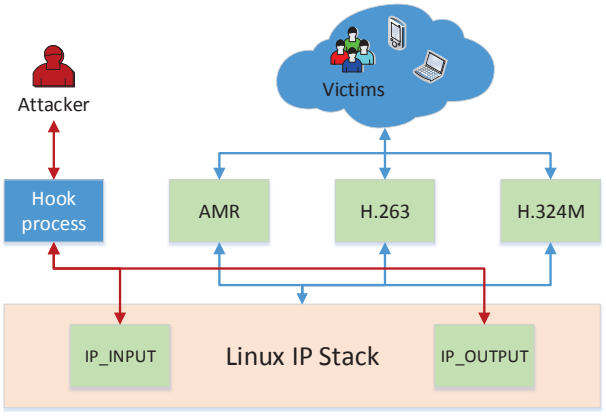


Fig. 6. A system design of man-in-the-middle-attack in Fog.

There are security solutions for Cloud computing. However, they may not suit for Fog computing because Fog devices work at the edge of networks. The working surroundings of Fog devices will face with many threats which do not exist in well managed Cloud. In this section, we discuss the security and privacy issues in Fog Computing.

A. Security Issues

The main security issues are authentication at different levels of gateways as well as (in case of smart grids) at the smart meters installed in the consumer's home. Each smart meter and smart appliance has an IP address. A malicious user can either tamper with its own smart meter, report false readings, or spoof IP addresses. There are some solutions for the authentication problem. The work [26] elaborated public key infrastructure (PKI) based solutions which involve multicast authentication. Some authentication techniques using Diffie-Hellman key exchange have been discussed in [27]. Smart meters encrypt the data and send to the Fog device, such as a home-area network (HAN) gateway. HAN then decrypts the data, aggregates the results and then passes them forward.

Intrusion detection techniques can also be applied in Fog computing [28]. Intrusion in smart grids can be detected using either a signature-based method in which the patterns of behaviour are observed and checked against an already existing database of possible misbehaviours. Intrusion can also be captured by using an anomaly-based method in which an observed behaviour is compared with expected behaviour to check if there is a deviation. The work [29] develops an algorithm that monitors power flow results and detects anomalies in the input values that could have been modified by attacks. The algorithm detects intrusion by using principal component analysis to separate power flow variability into regular and irregular subspaces.

B. An Example: Man-in-the-Middle Attack

Man-in-the-middle attack has potential to become a typical attack in Fog computing. In this subsection, we take man-in-the-middle attack as an example to expose the security problems in Fog computing. In this attack, gateways serving as Fog devices may be compromised or replaced by fake ones [30]. Examples are KFC or Star Bar customers connecting to malicious access points which provide deceptive SSID as public legitimate ones. Private communication of victims will be hijacked once the attackers take the control of gateways.

1) Environment Settings of Stealth Test: Man-in-the-middle attack can be very stealthy in Fog computing paradigm. This type of attack will consume only a small amount of resources in Fog devices, such as negligible CPU utilization and memory consumption. Therefore, traditional anomaly detection methods can hardly expose man-in-the-middle attack without noticeable features of this attack collected from the Fog. In order to examine how stealthy the man-in-the-middle attack can be, we implement an attack environment shown in Figure 5. In this scenario, a 3G user sends a video call to a WLAN user. Since the man-in-the-middle attack requires to control the communication between the 3G user and the WLAN user, the key of this attack is to compromise the gateway which serves as the Fog device.

Two steps are needed to realize the man-in-the-middle attack for the stealth test. First, we need to compromise the gateway, and second, we insert malicious code into the compromised system. For susceptible gateways, we can either refresh the ROM of a normal gateway or place a fake active point in

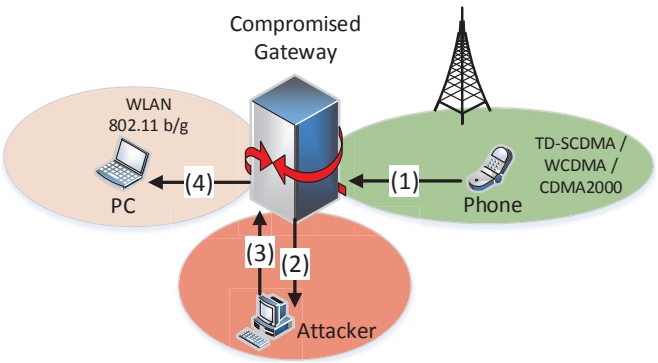


Fig. 7. The hijacked communication in Fog (e.g. from phone to PC).

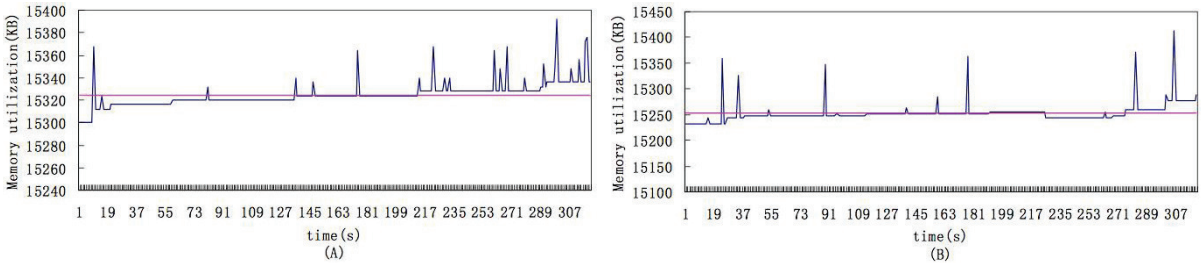


Fig. 8. Memory Consuming of man-in-the-middle-attack in Fog.

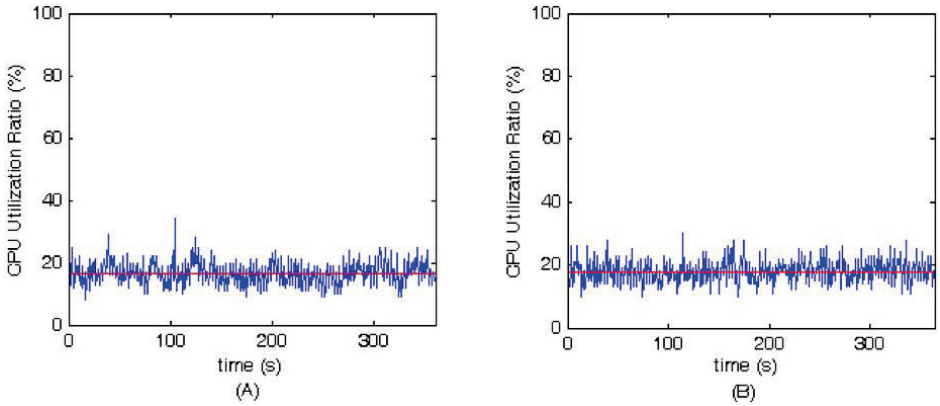


Fig. 9. CPU consuming of man-in-the-middle-attack in Fog.

the environment. Both methods can be easily implemented in the real world, such as in the KFC or Star Bar environments. In our experiment, we choose the former and use Broadcom BCM5354 as the gateway [31]. This device has a high-performance MIPS32 processor, IEEE 802.11 b/g MAC/PHY and USB2.0 controller. Video communication is set up on BCM5354 between a 3G mobile phone and a laptop which adopts Wifi for connection. We refresh the ROM of BCM4354 and update its system to the open-source Linux kernel 2.4.

In order to hijack and replay victims' video communication, we insert a hook program into the TCP/IP stack of the compromised system. Hook is a technique of inserting code into a system call in order to alter it [32]. The typical hook works by replacing the function pointer to the call with its own, then once it is done doing its processing, it will then call the original function pointer. The system structure is implemented in Figure 6. We further employ the relevant APIs and data structures in the system to control the gateway device, such as boot strap, diagnostics and initialization code. The IP packets from WLAN will be transferred to and processed in 3G related modules. We plug a 3G USB modem on BCM5354 device, on which we implement H.324M for video and audio tunnel with 3G CS. H.263 and AMR functions are also implemented as the video and audio codec modules in the system.

2) Work Flow of Man-in-the-Middle Attack: The communication between 3G and WLAN needs a gateway to translate the data of different protocols into the suitable formats. Therefore, all the communication data will firstly arrive at the gateway and then be forwarded to other receivers.

In our experiment, the man-in-the-middle attack is divided into four steps. We illustrate the hijacked communication from 3G to WLAN in Figure 7. In the first two steps, the embedded hook process of the gateway redirects the data received from the 3G user to the attacker. The attacker replays or modifies the data of the communication at his or her own computer, and then send the data back to the gateway. In the final step, the gateway forwards the data from the attacker to the WLAN user. In fact, the communication from the WLAN user will also be redirected to the attacker at first, and then be forwarded by the hook in the gateway to the 3G user. We can see clearly from Figure 7 that the attacker can monitor and modify the data sent from the 3G user to the WLAN user in the 'middle' of the communication.

3) Results of Stealth Test: Traditional anomaly detection techniques rely on the deviation of current communication from the features of normal communication. These features include memory consumption, CPU utilization, bandwidth usage, etc. Therefore, to study the stealth of man-in-the-middle attack, we examine the memory consumption and the CPU utilization of gateway during the attack. If man-in-the-middle attack does not greatly change the features of the communication, it can be proofed to be a stealthy attack. For simplicity, we assume the attacker will only replay the data at his or her own computer but will not modify the data.

Firstly, we compare the memory utilization of gateway before and after a video call tunnel is built in our experiment.

The results are shown in Figure 8, and the red line in plots indicates the average amount of memory consumption. We can see clearly that man-in-the-middle attack does not largely influence the video communication. In Figure 8(A), the average value is 15232 K Bytes, while after we build the video tunnel on gateway, the memory consumption reaches 15324.8 K Bytes in Figure 8(B). Secondly, we show the CPU consumption of gateway in Figure 9. Based on the results in Figure 9, we can also see that man-in-the-middle attack does not largely influence the video communication. In the Figure 8(A), the average value is 16.6704%, while after the video tunnel is built, the CPU consumption reaches 17.9260%. We therefore conclude that man-in-the-middle attack can be very stealthy in Fog computing because of the negligible increases in both memory consumption and CPU utilization in our experiments.

Man-in-the-middle attack is simple to launch but difficult to be addressed. In the real world, it is difficult to protect Fog devices from compromise as the places for the deployment of Fog devices are normally out of religious surveillance. Encrypted communication techniques may also not protect users from this attack since attackers can set up a legitimate terminal and replay the communication without decryption. Particularly, complex encryption and decryption techniques may not be suitable for some scenarios. For example, the encryption and decryption techniques will consume lots of battery power in 3G mobile phones. In fact, this attack is not limited to the scenario of our experiment environment. We can find many applications running in Fog computing are susceptible to man-in-the-middle attack. For example, many Internet users communicate with each other using MSN (Windows Live Massager). The communication data of MSN is normally not encrypted and can be modified in the 'middle'. Future work is needed to address the man-in-the-middle attack in Fog computing.

C. Privacy Issues

In smart grids, privacy issues deal with hiding details, such as what appliance was used at what time, while allowing correct summary information for accurate charging. R. Lu et al. described an efficient and privacy-preserving aggregation scheme for smart grid communications [33]. It uses a super-increasing sequence to structure multi-dimensional data and encrypt the structured data by the homomorphic cryptogram technique. A homomorphic function takes as input the encrypted data from the smart meters and produces an encryption of the aggregated result. The Fog device cannot decrypt the readings from the smart meter and tamper with them. This ensures the privacy of the data collected by smart meters, but does not guarantee that the Fog device transmits the correct report to the other gateways. For data communications from user to smart grid operation center, data aggregation is performed directly on cipher-text at local gateways without decryption, and the aggregation result of the original data can be obtained at the operation center [33]. Authentication cost is reduced by a batch verification technique.

VI. Conclusions and Future Work

We investigate Fog computing advantages for services in several domains, and provide the analysis of the state-of-the-art and security issues in current paradigm. Based on the work of this paper, some innovations in compute and storage may be inspired in the future to handle data intensive services based on the interplay between Fog and Cloud.

Future work will expand on the Fog computing paradigm in Smart Grid. In this scenario, two models for Fog devices can be developed. Independent Fog devices consult directly with the Cloud for periodic updates on price and demands, while interconnected Fog devices may consult each other, and create coalitions for further enhancements.

Next, Fog computing based SDN in vehicular networks will receive due attention. For instance, an optimal scheduling in one communication period, expanded toward all communication periods, has been elaborated in [6]. Traffic light control can also be assisted by the Fog computing concept. Finally, mobility between Fog nodes, and between Fog and Cloud, can be investigated. Unlike traditional data centres, Fog devices are geographically distributed over heterogeneous platforms. Service mobility across platforms needs to be optimized.

References

[1] F. Bonomi, "Connected vehicles, the internet of things, and fog computing," in *The Eighth ACM International Workshop on Vehicular Inter-Networking (VANET)*, Las Vegas, USA, 2011.

[2] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing*, ser. MCC'12. ACM, 2012, pp. 13–16.

[3] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Commun. ACM*, vol. 53, no. 4, pp. 50–58, Apr 2010.

[4] C. Wei, Z. Fadlullah, N. Kato, and I. Stojmenovic, "On optimally reducing power loss in micro-grids with power storage devices," *IEEE Journal of Selected Areas in Communications*, 2014 to appear.

[5] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Comput. Netw.*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010.

[6] K. Liu, J. Ng, V. Lee, S. Son, and I. Stojmenovic, "Cooperative data dissemination in hybrid vehicular networks: Vanet as a software defined network," *Submitted for publication*, 2014.

[7] K. Kirkpatrick, "Software-defined networking," *Commun. ACM*, vol. 56, no. 9, pp. 16–19, Sep. 2013.

[8] Cisco, "Cisco delivers vision of fog computing to accelerate value from billions of connected devices," Cisco, Tech. Rep., Jan. 2014.

[9] K. Hong, D. Lillethun, U. Ramachandran, B. Ottenwälder, and B. Koldehofe, "Opportunistic spatio-temporal event processing for mobile situation awareness," in *Proceedings of the 7th ACM International Conference on Distributed Event-based Systems*, ser. DEBS'13. ACM, 2013, pp. 195–206.

[10] H. Madsen, G. Albeanu, B. Burtschy, and F. Popentiu-Vladicescu, "Reliability in the utility computing era: Towards reliable fog computing," in *Systems, Signals and Image Processing (IWSSIP), 2013 20th International Conference on*, July 2013, pp. 43–46.

[11] K. Hong, D. Lillethun, U. Ramachandran, B. Ottenwälder, and B. Koldehofe, "Mobile fog: A programming model for large-scale applications on the internet of things," in *Proceedings of the Second ACM SIGCOMM Workshop on Mobile Cloud Computing*, ser. MCC'13. ACM, 2013, pp. 15–20.

[12] T. Nishio, R. Shinkuma, T. Takahashi, and N. B. Mandayam, "Service-oriented heterogeneous resource sharing for optimizing service latency in mobile cloud," in *Proceedings of the First International Workshop on Mobile Cloud Computing and Networking*, ser. MobileCloud'13. ACM, 2013, pp. 19–26.

[13] B. Ottenwalder, B. Koldehofe, K. Rothermel, and U. Ramachandran, "Migcep: Operator migration for mobility driven distributed complex event processing," in *Proceedings of the 7th ACM International Conference on Distributed Event-based Systems*, ser. DEBS'13. ACM, 2013, pp. 183–194.

[14] J. Zhu, D. Chan, M. Prabhu, P. Natarajan, H. Hu, and F. Bonomi, "Improving web sites performance using edge servers in fog computing architecture," in *Service Oriented System Engineering (SOSE), 2013 IEEE 7th International Symposium on*, March 2013, pp. 320–323.

[15] BETaaS, "Building the environment for the things as a service," BETaaS, Tech. Rep., Nov. 2012.

[16] S. Maharjan, Q. Zhu, Y. Zhang, S. Gjessing, and T. Basar, "Dependable demand response management in the smart grid: A stackelberg game approach," *Smart Grid, IEEE Transactions on*, vol. 4, no. 1, pp. 120–132, March 2013.

[17] D. Korzhyk, V. Conitzer, and R. Parr, "Solving stackelberg games with uncertain observability," in *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 3*, ser. AAMAS '11, 2011, pp. 1013–1020.

[18] Z. Fadlullah, D. Quan, N. Kato, and I. Stojmenovic, "Gtes: An optimized game-theoretic demand-side management scheme for smart grid," *Systems Journal, IEEE*, vol. 8, no. 2, pp. 588–597, June 2014.

[19] T. Luo, H.-P. Tan, and T. Quek, "Sensor openflow: Enabling software-defined wireless sensor networks," *Communications Letters, IEEE*, vol. 16, no. 11, pp. 1896–1899, Nov. 2012.

[20] Y. Daraghmi, C.-W. Yi, and I. Stojmenovic, "Forwarding methods in data dissemination and routing protocols for vehicular ad hoc networks," *Network, IEEE*, vol. 27, no. 6, pp. 74–79, November 2013.

[21] B. Zhou, J. Cao, X. Zeng, and H. Wu, "Adaptive traffic light control in wireless sensor network-based intelligent transportation system," in *Vehicular Technology Conference Fall (VTC 2010-Fall), 2010 IEEE 72nd*, Sept 2010, pp. 1–5.

[22] B. Zhou, J. Cao, and H. Wu, "Adaptive traffic light control of multiple intersections in wsn-based its," in *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*, May 2011, pp. 1–5.

[23] C. Li and S. Shimamoto, "An open traffic light control model for reducing vehicles co2 emissions based on etc vehicles," *Vehicular Technology, IEEE Transactions on*, vol. 61, no. 1, pp. 97–110, Jan 2012.

[24] W. Wang and Z. Lu, "Survey cyber security in the smart grid: Survey and challenges," *Comput. Netw.*, vol. 57, no. 5, pp. 1344–1371, Apr. 2013.

[25] R. Lu, X. Li, X. Liang, X. Shen, and X. Lin, "Grs: The green, reliability, and security of emerging machine to machine communications," *Communications Magazine, IEEE*, vol. 49, no. 4, pp. 28–35, April 2011.

[26] Y. W. Law, M. Palaniswami, G. Kounga, and A. Lo, "Wake: Key management scheme for wide-area measurement systems in smart grid," *Communications Magazine, IEEE*, vol. 51, no. 1, pp. 34–41, January 2013.

[27] Z. Fadlullah, M. Fouda, N. Kato, A. Takeuchi, N. Iwasaki, and Y. Nozaki, "Toward intelligent machine-to-machine communications in smart grid," *Communications Magazine, IEEE*, vol. 49, no. 4, pp. 60–65, April 2011.

[28] C. Modi, D. Patel, B. Borisaniya, H. Patel, A. Patel, and M. Rajarajan, "A survey of intrusion detection techniques in cloud," *Journal of Network and Computer Applications*, vol. 36, no. 1, pp. 42–57, 2013.

[29] J. Valenzuela, J. Wang, and N. Bissinger, "Real-time intrusion detection in power system operations," *Power Systems, IEEE Transactions on*, vol. 28, no. 2, pp. 1052–1062, May 2013.

[30] L. Zhang, W. Jia, S. Wen, and D. Yao, "A man-in-the-middle attack on 3g-wlan interworking," in *Communications and Mobile Computing (CMC), International Conference on*, vol. 1, April 2010, pp. 121–125.

[31] Broadcom bcm 5354. [Online]. Available: http://www.broadcom.com/products/Wireless-LAN/802.11-Wireless-LAN-Solutions/BCM5354

[32] Wikipedia. (2014) Hooking, what is hooking? [Online]. Available: http://en.wikipedia.org/wiki/Hooking

[33] R. Lu, X. Liang, X. Li, X. Lin, and X. Shen, "Eppa: An efficient and privacy-preserving aggregation scheme for secure smart grid communications," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 23, no. 9, pp. 1621–1631, Sept 2012.

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 9–16

DOI: 10.15439/2014F509
ACSIS, Vol. 2

The Smart Grid's Data Generating Potentials

Marco Aiello
Johann Bernoulli Institute for
Mathematics and Computer Science
University of Groningen
Groningen, The Netherlands
Email: m.aiello@rug.nl

Giuliano Andrea Pagani
Johann Bernoulli Institute for
Mathematics and Computer Science
University of Groningen
Groningen, The Netherlands
Email: g.a.pagani@rug.nl

***Abstract*—The Smart Grid is the vision underlying the evolution the power grid is currently undergoing. Its pillars are increased efficiency, self-healing, operation automation, and renewable energy integration obtained through real-time control and digitalization of the infrastructure. Thus, an important ingredient—if not the main one—is information technology support for power transmission and distribution. Given the size of the power grid, its pervasiveness, and the need for its availability, it is easy to imagine that any serious ICT infrastructure dealing with it will have to manage a great deal of rapidly forming data. Now the question is whether the amount, diversity, and uses of such data put the smart grid in the category of Big Data applications, followed by the natural question of what is the value of such data. To provide an initial answer to this question, we analyze the current state of data generation of the Dutch grid, its evolution towards a smart grid, and a future realistic scenario. The scenario considered shows that the amount of data generated is comparable to some of today's social media and "classic" Big Data examples.**

***Index Terms*—smart grid; power systems; Big Data**

I. Introduction

THE EVOLUTION of the power grid towards a smart grid is based on a massive deployment of Information and Communication Technology (ICT) in sensing, analyzing and controlling the operations of the power grid, from generation to utilization. This shift to a more information technology-based power grid will require considerable amounts of data to be produced by the sensing equipment, and by the new generation of (smart) meters. Thus, setting a challenge for the current ICT architectures of utilities and electricity distribution companies. If this trend is underway now, it is especially because the diffusion of renewable generation sources at all levels of the power grid calls for timely and precise monitoring of the infrastructure. In addition metering equipment is more affordable and reliable.

If the shift towards a smart digitalized grid is broadly accepted, little is known about the actual data generation potential of the future grid and on the management and utilization of large electricity data streams. The few works in the literature provide only a qualitative analysis of the amount of data that the smart grid is likely to generate with almost no results of quantitative experiences and field tests. To fill such knowledge gap, with the current treatment, we assess the amenability of the Big Data definition to the smart grid considering mainly the volume and velocity features of smart grid data. We perform such assessment by comparing the amount of data produced and transmitted in today's prominent Big Data examples, coming from the areas of social media and Internet based services. For the comparison, we use information coming from the Dutch power grid. The Netherlands has one of the infrastructures with one the highest availability in Europe with the electrical system being in 2012 99,99486% of the time available [1]. Moreover, The Netherlands plans to have a full rollout of smart meters by 2018 and a standard for the information to be read and exchanged by smart meters is currently being finalized.

The rest of the paper is organized as follows: Section II briefly defines the Big Data concept and offers some examples of today's sources of data. Section III provides a description of the concept of smart grid, while Section IV provides a quantitative analysis of the amount of data generated by a smart grid infrastructure with special focus on the Dutch grid as concrete example. A survey of related work is presented in Section V and the concluding remarks complete the paper, Section VI.

II. Big Data Examples

Big Data refers to information systems characterized by having to manage high volumes of data, which is rapidly created and that potentially has added value. It is common practice to refer to the *5Vs* when talking about Big Data. The *5Vs* stand for *Volume, Velocity, Variety, Value and Veracity*: big quantity of information that moves fast on a network; data that are diverse and provide relevant facts (implicit) in the information they carry, while being reliable as data sources.

Communication capacity: The telecom sector has seen an exponential growth in its infrastructure. The Internet alone in the short time span of 7 years (2000-2007) has increased network utilization by 29 times, with the necessity of transmitting 65 exabytes per year (optimally compressed) [2]. Cisco computed a monthly IP traffic of 43 exabyte for 2012 and lead to an almost 30% increase rate till 2016 reaching 110 exabytes/month and 1.3 zettabytes/year [3].

Prominent examples of Big Data today have to do with social media, where billions of users interact by exchanging data in various forms. Let us consider YouTube and Facebook as representative examples of the trend.

978-83-60810-58-3/$25.00

YouTube: YouTube is an on-line user-based broadcasting service where every user can watch and publish videos. Some key-facts numbers from YouTube's statistics[1]: more than 1 billion unique users visit YouTube monthly, over 4 billion hours of video are watched each month on YouTube, and 72 hours of video are uploaded to YouTube every minute.

Facebook: Facebook is the best-known social platform having recently passed the 1 billion users mark. The statistics concerning the amount of data that Facebook deals with are quite impressive: more than 500 terabytes of new data every day, 300 millions photo uploads, 2.7 billions 'Likes' each day[2]. It is then no wonder that with all these data Facebook requires an Hadoop Distributed File System cluster with more than 100 petabytes of physical space.[3]

Many more examples exist. E.g., computational science projects such as the data obtained from the Large Hadron Collider experiments; the astronomic images collected by (radio) telescopes; data generated on stock markets and used, for instance, for algorithmic trading; private businesses having 1 million transactions per hour, such as, Walmart, or the almost 500 transactions per second managed during peaks at Amazon.

III. The Smart Grid

The term smart grid does not yet have a unique definition [4], rather the various stakeholders and scientific disciplines involved have their own point of view on the area. From a physical and technical perspective, the system has always to satisfy the equilibrium between energy supply and demand in order to keep the correct operations and safety of the system. The information flow related to the operations of the grid is the real innovation of the Smart Grid. The grid will become more and more digital with information recorded by the sensors and the digital meters deployed at users' premises and along the grid and power stations. The benefit is to have enriched information of the performance of the system, its stability, and customer consumption. Another important motivation that drives the modernization of the grid lies in the ability to accommodate more renewable sources [5]. With more and more unpredictable (renewable) power sources, the electrical system needs more flexibility in managing the demand and supply equilibrium. One of the mechanisms to achieve such flexibility is through the use of variable electricity tariffs. These tariffs vary even several times per day and are transmitted to the users that are able to react to them by increasing or decreasing their consumption. For example, on a very windy day the wind energy production surges and in order to keep the balance of the grid, tariffs are lowered to incentivize the use of electricity. Naturally, to have such real-time flexibility, information has to be exchanged with the users both in terms of dynamic tariffs and energy used at a given price to enable the accounting.

An essential component in this scenario is the smart meter. It has the same primary duty of the traditional analog meter, but in addition it has a memory to store information, a processor, a digital clock, it can use multiple tariffs, it is able to measure bidirectional power flows (i.e., consumed and produced energy), and it is connected with a telecommunication link to the utility. All these features let us talk today of Advanced Metering Infrastructure (AMI).

[1]http://www.youtube.com/yt/press/statistics.html)
[2]http://goo.gl/2wSzw
[3]http://goo.gl/Ic0x9

TABLE I: 2012 Dutch smart grid data.

Metering	
Metered customers	7,827,350
Installed smart meters	450,000
Smart meter sampling period (min)	86,400 (by law 2 months)
Smart Devices	
Electric vehicles	6,275
Battery packs	N/A
Intelligent appliances per household	0
Grid Infrastructure	
Nodes HV (380/220kV)	36
Nodes MV/LV	155,000

IV. The Smart Grid and Big Data

Taking The Netherlands as our case study, we consider the current and a possible future situation for the power grid, specifically considering the amount of generated data. We have chosen The Netherlands because the electrical system is very reliable, with just an average of 27 minutes downtime per year per customer [1], because it has a modern infrastructure with an unbundled energy market, and because it is going fast towards a digitalization of the electrical system.

A. Volume and Velocity

Volume and velocity are the most two prominent aspects of the transition towards a smart grid.

Metering and Smart Buildings: In the last years in The Netherlands, a compulsory smart meter roll-out had been planned and then canceled due to privacy concerns [6]. Thus, currently smart meters are only partially installed in the country and the approach is first to test the AMI in controlled settings and in pilots before proceeding to a massive roll out, expected to cover the country by 2018.

In our quest for Big Data, we consider the current situation of advancement of the AMI and smart grid in The Netherlands and we compare it to two realistic scenarios of the near future. The current situation is synthesized in Table I. To date, in The Netherlands only about 5% of the meter installations are smart. Furthermore, the data measuring capabilities of the meter is only partially used, since current law mandates that they can only be read once every two months. A number of *smart devices* might also interact and exchange information with the grid about tariffs or the power use. Most notably, electric vehicles, batteries in residential premises and intelligent home appliances contribute to the picture [7]. Weather information represents another essential source in the future smart gird. Knowing the weather in advance allows an home energy management system to forecast how much energy will be produced by user's solar panels or small wind turbines, and the internal needs for heating and cooling.

TABLE II: Near future scenario (e.g., one decade).

Metering	
Metered customers	8,000,000
Installed smart meter	8,000,000
Smart meter sampling period (min)	15
Smart Devices	
Electric vehicles	790,000
Battery packs	45,000
Intelligent appliances per household	10
Grid Infrastructure	
Nodes HV (380/220kV)	45
Nodes MV/LV	158,419

TABLE III: Far future scenario (e.g., four/five decades).

Metering	
Metered customers	9,000,000
Installed smart meter	9,000,000
Smart meter sampling period (min)	5
Smart Devices	
Electric vehicles	3,950,000
Battery packs	135,000
Intelligent appliances per household	20
Grid Infrastructure	
Nodes HV (380/220kV)	60
Nodes MV/LV	178,221

Table II shows a short-medium term scenario for the same objects, when the roll-out of the smart meters will be completed. The 100% penetration is based on the commitment of the European Union to deploy smart meters to at least 80% of all customers (c.f. EU directives 2009/72/EC and 2009/73/EC) and the commitment of the Dutch government. As a sampling period we consider 15 minutes. This is the time interval used in many of the AMI installations and tests (e.g., CenterPoint AMI in Texas.[4] We envision most of the advancements in the adoption of smart devices with almost 800,000 electric vehicles (equal to 10% of the passenger vehicles in 2012), 45,000 battery packs for local energy storage (based on the figures of solar capacity installed till 2011) and the penetration of home intelligent appliances of 10 per metering point (as about half of the average number of appliance per family).

Table III provides a long term vision, say few decades. We consider that the fully digital smart meter infrastructure reaches 9 millions customers throughout the whole country. As sampling period we consider an infrastructure that is closer to real-time measurement with 5 minutes interval. We suppose that smart appliances become the norm, therefore with 20 smart devices on average in each home or office leading to a total of 180 millions devices. The number of electric passenger vehicles reaches 3.95 millions (equal to 50% of the passenger vehicles in 2012), while the battery packs increase threefold. This last assumption is based on the estimate of the increase in distributed generation capacity by three times compared to the 2011 figures.

The Power Grid: Complementary to the users and metering of the electricity, there is the transmission and distribution infrastructure. In the bottom of Tables I, II, and III we report

[4]http://goo.gl/MzGPX

TABLE IV: Data size for various parameters/devices/services in bytes per sample.

Data Use in Metering	**Size**
Consumption only	193
Consumption/production	245
Consumption/production, instantaneous power (3-phase), and current	530
Consumption/production, instantaneous power (3-phase), current, failures, and gas metering	1,100
Data Use in Smart Devices	**Size**
Electric vehicle consumption only	193
Electric vehicle consumption/feed-in	245
Electric vehicle consumption/feed-in, instantaneous power (3-phase), and current	530
Electric vehicle consumption/feed-in, instantaneous power (3-phase), current, failures, and gas metering	1,100
Battery	200
Intelligent appliances	200
Data Use in Weather Forecast	**Size**
Essential weather parameter	13,000
Improved weather parameter	15,000
Advanced weather parameter	20,000

on the current size of the Dutch power grid, size that we do not expect to increase dramatically in the medium-long period. We consider the nodes (e.g., power and transformation station) since these contain the equipment and the sensors that monitor the status of the power assets and the lines. Considering the high voltage, the current number of nodes is provided in the work of Rosas-Casals and Corominas-Murtra [8], whereas the information concerning the medium and low voltage nodes is provided by the various distribution utilities of the Netherlands.[5] For the high voltage grid, we consider a moderate evolution based on the public plans of the transmission operator (TenneT[6]). In considering the evolution, we make the conservative assumption that the ratio between the medium-low voltage nodes and the metered customers is constant and equal to the current ratio. That means that the total number of medium-low voltage nodes is 158,419 and 178,221 in the short term and long term future scenario, respectively.

Data Size: The Dutch Smart Meter requirement document [9] defines the size and function of the information sent. Based on this information we consider the user generated data: data from self generation, failure statistics, additional energy metering (e.g., 3-phase installation, gas metering) as shown on the top of Table IV. In the current situation, the amount of data produced by the Dutch smart grid is extremely limited as shown in Figure 1. The most parsimonious case entails just a yearly amount of about 500 megabytes, while measuring the whole set of parameters requires almost 3 gigabytes.

Things change when considering future scenarios where the diffusion of smart meters, electric vehicles and appliances will be pervasive. As shown in the central part of Table IV, we consider four categories of data generation related to the

[5]http://www.energieleveranciers.nl/netbeheerders/overzicht-netbeheerders
[6]http://www.tennet.eu

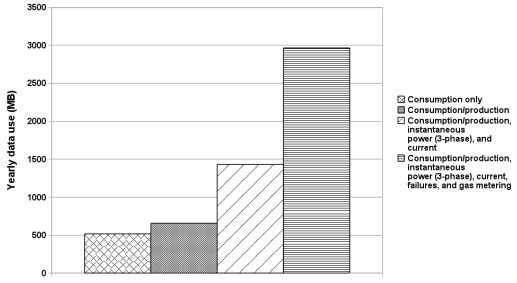


Fig. 1: Current Dutch AMI yearly data generation.

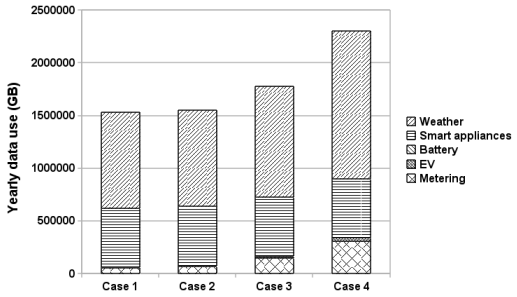


Fig. 2: Near future Dutch smart grid yearly data generation.

electric vehicles. In essence, an electric vehicle is not much different from a house metering point. Another similarity is in the possibility of feeding energy to the grid when required, thus acting as a production unit. In this future scenario, we also consider the interaction with meteorological services that provide weather information and forecast which influences users energy behavior, their energy production units, and the schedule of their appliances. The last three lines of Table IV show the amount of data required by a weather service. The National Oceanic and Atmospheric Administration (NOAA) provides eXtensible Markup Language based information regarding temperature, wind speed, and cloud coverage conditions up to 1 week forecast. The 'improved' and 'advanced' weather service data are considered for improved weather services with meteorological information with fine grain timescale. This granularity of information is essential for reliable forecasting of local production of energy, and therefore needed in the smart grid.

For the big data estimation, we identify four cases. Each case has increasing data quantities as more units are involved. The generated data can come from:

- *Case 1:* Consumption data from smart metering and electric vehicles; data from batteries and intelligent appliances; essential weather data.
- *Case 2:* Data for consumption/feed-in from smart metering and electric vehicles; data from batteries and intelligent appliances; essential weather data.
- *Case 3:* Data for consumption/feed-in, instantaneous power, and current from smart metering and electric vehicles; data from batteries and intelligent appliances; rich weather data.
- *Case 4:* Data for consumption/feed-in, instantaneous power, failures from smart metering and electric vehicles; data from gas metering; data from batteries and intelligent appliances; rich weather data.

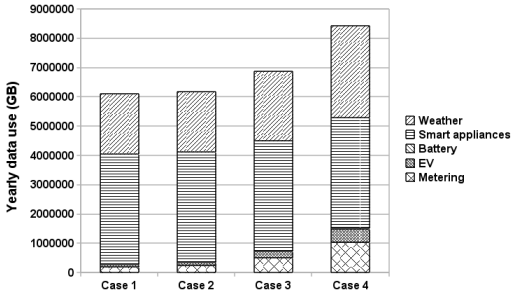


Fig. 3: Far future Dutch smart grid yearly data generation.

Considering the near future scenario for full smart grid implementation the amount of data produced yearly for the whole of The Netherlands starts to assume the connotation of "Big Data" (Figure 2). We remark that the sampling frequency is 15 minutes for metering whereas the weather information are provided every hour. In the most conservative scenario (Case 1) the data reaches the amount of 1.5 petabytes. Case 2 is similar in data generation having a total amount again close to 1.5 petabytes. Case 3 is close to 1.8 petabytes due to the increased information recorded and more comprehensive weather information. In Case 4 we note that more than half of the data generated comes from the meteorological information. The actual metering accounts for about 300 terabytes, while the electric vehicles are about 10% of this figure. The high number of smart appliances considered in a full fledged implementation of the smart grid, makes them responsible of the generation of more than half a petabyte of data.

Considering a future scenario with even higher penetration of smart devices and higher sampling rates, the amount of data naturally grows (Figure 3). We assume a sampling period of 5 minutes for metering-related data, and 30 minutes for weather. In the first case (most conservative), the data reaches the 6 petabytes value. Case 2 is almost as data rich as Case 1, whereas Case 3 reaches almost 7 petabytes. In all the cases the source that causes the most of data are smart appliances responsible of a little less than 4 petabytes of data. In Case 4 one reaches the level of about 8.5 petabytes, distributed between smart appliances (almost 3.8 petabytes), weather information (about 3 petabytes), and smart meters (1 petabyte).

Before concluding the analysis of the volume and, to some extent, of the velocity of the data that can come from the future smart grid users, we analyze the data generating potentials of the power grid infrastructure itself. Table V provides the values for the high voltage and the low voltage grid for the current situation and for the near and far future scenarios. For each of the three temporal variants, we consider the percentage of

TABLE V: Data for grid monitoring.

Type of Power Station	% of Station Monitored	Sampling period (min)	Data Size (byte)
Current Scenario			
HV	100	15	12,950
MV/LV	10	1,440	40
Near Future Scenario			
HV	100	5	12,950
MV/LV	50	60	40
Far Future Scenario			
HV	100	1	1,295,000
MV/LV	100	5	4,000

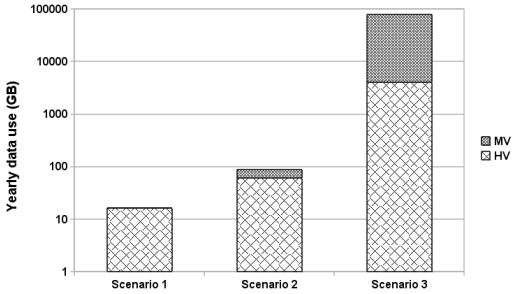


Fig. 4: Dutch power grid data generation.

monitored infrastructure, the sampling period and the size of the data.

Today, with very few exceptions, the medium/low voltage stations have manual switching device operations and therefore no data or remote monitoring is in place. The sampling period will shorten in the future and we assume to go from days to 5 minutes. Regarding the data size, we consider the data involved in the query and response of a distribution relay which is 40 bytes following the Modbus standard[7] that is a typical standard for electrical equipment monitor. We assume that in the far future scenario the data will be increased to 4,000 bytes per sample. The assumption is based on the presence of more sensing equipment and richer protocols. Considering the high voltage stations, they are already fully monitored; we consider only a decrease in the period of monitoring from 15 to 5 and 1 minute in the three scenarios. The amount of data required is based on the study of Sanchez *et al.* on the implementation of telecontrol functions for electrical stations using IP technologies, by considering just one picture frame per monitoring function [10]. Even here, we keep the current and the near future value constant and increase 100 folds the situation for the far future since we assume more equipment and richer protocols.

Based on these projections, we report the expectations for how much data is generated by the Dutch power grid in Figure 4. One notices an important growth for the high voltage related data, though what is remarkable is the important appearance of the distribution grid (medium and low voltage) data. In fact, this is where the smart grid concept will mostly change the current status quo.

B. Variety

Several standardization bodies and international organizations are working to provide new standardization documents to guarantee the interoperability between electric and electronic equipment in the context of the (smart) grid. To date IEEE alone has more than 100 standardization initiatives. IEC is going in the same direction with more than 100 relevant standards involving the smart grid in its various domains. These standards are not all related to data transmission and information-oriented aspects, since there are also aspects inherited by the electric aspects. In general, the application domains can be divided into different portions where standards and information exchanges can be considered. In the IEC representation are defined: the IEC-61850 that describes the communication networks and systems inside a substation; the IEC-61968 describes the information exchange in the electricity distribution system; the IEC-61970 describes the common information model (CIM) used in the control center to interoperate with the components and equipments outside the control center. Security aspects of the various communications are addressed in the IEC-62351.

In addition to technical aspects, it is important to consider the data exchange factor related to the market. The efforts of two standardization bodies (IEC and NIST) and State Grid Corporation of China are compared in [11] where similarities and differences in the approach are compared and a probabilistic model on the amount of development is provided. IEC-62325 deals with the market aspects putting the accent on deregulation. The market efforts are also addressed by the OASIS technical committee for the smart grid. Both the OASIS Energy Inter-operation specifications and the OASIS Market Information Exchange provide a framework for standardization where energy interactions are considered as transactions, i.e., energy is an operation comprising a price, a quantity and a date of delivery that can be negotiated on the market.

The home automation setting [12] is also a rich one in terms of data variety. There are many technologies and protocols for home energy saving and many initiatives are underway in relation to the smart grid. Zigbee Alliance Smart Energy Profile provides the protocol stack (based on the IP protocol) to have smart grid application at home such as meter reading, and appliance automation in response to changing energy prices. In order to enable a true interoperability between so many different and heterogeneous services, the tendency is that of using solutions that are based on services available through the Internet [13]. In summary, many standards are appearing with different data specifications and formats.

C. Veracity

In the broad panorama of the smart grid, the problem of having accurate and meaningful data is essential. Having so many different sources and levels of aggregation requires a fine grained filtering to ensure that data are consistent and

[7] www.modbus.org/

that no compromised or mistaken data are present. In addition to the traditional mechanism of error correction coding [14], one of the issues that are highly debated about the smart grid is the security of a digital power grid infrastructure. There are several possible attacks, ranging from the acquisition of private user data, to tampering the energy bills, to modifying the data recordings of the equipment of the grid, to malware injection into general purpose and embedded computers, modifications of the demand-response mechanism creating imbalances in the production and consumption and therefore compromising the stability of the grid [15]. Therefore accuracy, integrity, confidentiality, and authenticity have to be all satisfied to guarantee the safe operations of the smart grid. The cyber-security aspects of the future power grid are a source of concern. Solutions to the security issues are under investigation for instance by NIST that has established a Cyber Security Working Group on the smart grid topic. In [16] the proposed solution is based on technologies already applied to ICT domains such as public key infrastructure for providing a secure way of authentication and authorization. Encryption based on AES or 3DES algorithms can provide a solution to guarantee privacy. In order to guarantee the trust in a system made of many different components (several thousands to millions) with different technologies and several manufacturers, the solution proposed by Metke and Ekl [16] uses trust anchor security i.e., a series of chains of certification authorities and at the end of the chain a digitally signed certificate for a piece of equipment or device in the smart grid is provided. Among the different security concerns of the smart grid analyzed in [17] some are definitely related to the data: security of devices and their communication interactions, and privacy of the users whose energy consumption is digitally metered.

D. Value

It is difficult to give a value to information regarding an infrastructure not yet in place and which is managed as a monopoly in most countries, or at most as a slightly unbundled market. However, we can estimate the value of the smart grid data considering some of the actors involved and the benefits they can achieve with an accessible and rich digital infrastructure.

The end user can obtain value by energy saving and appropriate planning of its energy use. This can be done only with an automation system in place and would be even more profitable with a dynamic tariff system, as shown in [7]. The value can be even greater if a totally open energy market existed where end-users could trade their excess and stored energy freely with any other partner, a scenario we depicted in [18].

The value for the energy utilities will be manyfold. First, the AMI provides accurate measure of power consumed thus a more accurate billing for the end user. Second, the AMI and the data flow of a smart meter helps to discover and fight electricity thefts and have better revenues. Worldwide estimations of electricity thefts account for 25 billion dollars [19]. Third, energy utilities with smart grid data are able to early diagnose problems (especially at distribution level) in the network and provide a fast solution, which results in less customer interruptions and therefore less penalties for service interruption. Fourth, and most importantly, precise knowledge of energy flows allows utilities to better plan the generation of energy and save money, since energy production costs are not a linear function of the amount, but rather a convex one. Fifth, the inclusion of renewables requires a better prediction of the distributed generation to keep the demand-response balance.

With the smart grid concept, there is also the possibility that new actors can find value in the power grid (Big) data. New companies are emerging (e.g., Opower) that work for the utilities in smart meters data analytics, to motivate the customer in energy conservation, dynamic tariffs usage, and energy efficiency through automation. Furthermore, several benefits are achievable through smart appliances. In fact, they enable totally new scenarios of use of the home environment. The home appliances of the next generation will receive a configuration by the user who sets the preferences of the usage time and tariffs limits to be used [20]. These appliances will interact with the home energy management system, with the utilities that might directly control them, and with the manufacturer that receives logs about the appliance life conditions. Electric vehicles will also interact with the utilities, to provide a balancing capacity for the network, therefore they will participate in the demand response functionality receiving energy tariff information, remotely provide the battery status to the user and the log of anomalies to the vehicle manufacturer. Utilities and energy services companies can take advantage of energy analytics to develop new tailor made products. The approach will be similar to the personalized offers that Amazon or other on-line retailers offer to the customers having a more detailed set of information concerning the products that the customer likes or visits frequently. Utilities will provide energy offers based on the energy consumption patterns of customers, on the appliance usage and charging time of the electric vehicle.

V. Related Work

The smart grid is a new industrial and research theme open to exploration. The impact of data production and management is still under investigation. The few related works on the topic consider the amount of data generated to be huge, but very few provide actual numbers not to mention field experiences.

The direction towards an electricity sector that is more complex is synthesized by Rusitschka *et al.* [21]. The idea is to exploit cloud and a service-oriented approaches for processing power grid data. The paper does not provide an estimation of the amount of data that the smart grid is going to generate. The authors take for granted that the amount will be big: “The smart grid will be the largest increase in data any energy company has ever seen.” [21].

Parikh *et al.* [22] investigate the appropriate wireless technology to be used for the smart grid. The various solution proposed range between WiMAX and cellular for long distance communication to the wireless LAN, ZigBee and Bluetooth to more short distance applications. The review is interesting,

though the paper only considers the theoretical bandwidth of the technologies without considering the data size and sampling requirement of the smart grid.

The important problem of privacy in metering is addressed in [23] where the authors propose an anonymization mechanism for smart metering data. The approach considers two types of metering services a *high frequency* and a *low frequency*. The solution proposed appears effective and able to actually solve the main privacy concerns. The authors only refer to high frequency of sampling without giving any quantitative aspects on the amount of data. Another interesting aspect to be investigated is the amount of overhead in data size necessary to guarantee privacy.

The authors of [24] propose a new communication infrastructure to deal with the enhanced amount of data that will be generated by the smart grid. The infrastructure is a distributed one. The authors explicitly state that the new smart grid infrastructure will generate a significant amount of data given the increased sampling frequency of the grid sensing equipment (e.g., phasor measurement units) and that the current infrastructure is not ready to handle them. The authors do not provide any quantitative value concerning the data generated by the current or the future grid. Only a table shows the main characteristic of the sampling rate of part of the today's sensing and actuating infrastructure.

The concept of a distributed control for the power grid in the future evolution of the grid is illustrated also in [25]. The authors consider that the system will evolve from a mechanical-electrical control system to a fully electronics based one. They also state that more computation will take place locally on the grid with agent-based technologies operating within substations. Again, no sizing concerning the communication infrastructure to enable this additional information exchange between the substations is provided, thus difficult to have a quantitative picture of the Big Data of the smart grid.

In order to test the new scenarios of the data generation of the smart grid, EDF has estimated in its French network 35 millions smart meters and a sampling frequency of 10 minutes to have a total amount of data about 120 terabytes/year [26]. In that report, the authors show how a solution based on Hadoop and an accurate optimization in the data modeling, partitioning, and compression can improve the performance of the system that manages these data.

The IBM white paper on Big Data and smart grid provides some quantitative insights on the amount of data that utilities are going to deal in a smart grid future [27]. With a smart meter infrastructure and a 15 minutes sampling period, IBM forecasts a 3000 fold increase in the amount of data compared to the current monthly metering situation. Another quantitative example that is provided in the report to show the goodness of the IBM data management solution shows that in a 31-day period the total amount of storage required in a 100 millions meter scenario was less than 4 terabytes. The report explicitly emphasizes that the volume of data will not reach the same amount as for traditional data intensive industries, though in the power systems utilities world the projected amount could be overwhelming. The report also emphasizes that in addition to the *velocity* characteristic of the smart grid data in collection, processing, and use, the utility will have to deal with the *variety* of the data to handle from power control system data, to surveillance videos, to geographical and meteorological data, to social media mining.

The literature provides good examples of how the power grid will evolve. For instance, in [28] we look at the topological aspects of the distribution grid through the lenses of complex network analysis [29]. In general, all the works agree that additional amount of computation will be required to monitor the supplementary information coming from more sensors deployed in the grid. However, the studies never provide quantitative information concerning the amount of data that AMI or data gathered at substation level will produce. Only few studies try to provide indication of the communication infrastructure required, but in a coarse way.

VI. Concluding Remarks

Given the current state of affairs, the smart grid, or better said AMI, is not amenable to be referred to as an example of Big Data, not in the Netherlands, not in other countries. However, when more and more houses and businesses will be equipped with smart meters and the sampling period will be reduced, then the Dutch smart grid will become closer to be a Big Data system. If we compare the absolute numbers of the Big Data examples that we have analyzed in Section II, then there is no competition with the annual amount of data generated by social media or video repositories. However, to make the comparison fair considering Facebook, for example, if we account the the traffic generated by the more than 8 millions Dutch users of Facebook,[8] they are responsible for about 1.6 petabytes a year. We indeed note that the amount of data produced by the future Dutch smart grid is similar to the amount of data that the Dutch users produce each year on Facebook.

The numbers provided in this paper should be taken with a grain of salt. For instance, the Dutch smart meter standard is not yet finalized, therefore the amount of data and information metered could change, thus increasing or decreasing the data required. One aspect that we have not considered in our investigation is the overhead required by the communication infrastructures (i.e., the extra data in addition to the application layer that need to be considered) that could be a substantial burden. Further, all the data containing user and privacy related information will need to be protected by an additional layer of security (e.g., encryption) that require even more bits to be transmitted and stored. On the other hand, we have not considered any compression level possible for the smart grid data, but the figures here presented are raw data. Compression both in the communication and storage of data has achieved remarkable results [30]. For the protocols to control the grid we have considered one for the high

[8] http://goo.gl/HyKzxQ

voltage (IEC 60870-5 [10]) and one for medium/low voltage (Modbus) among many others available and in use.

Utilities are likely to have to face a number of data related issues in the short and medium term. In the implementation of big-scale AMIs utilities will deal with ICT challenges that are not the core of their business. In addition to a secure telecommunication infrastructure to transfer the metered data and other information, the utilities will have to deal with the storage and management of that data. Of course this is not a big problem from a technical perspective, similar solutions exist in the financial sector. However, the smart grid will require a modernization of the utilities that in addition to the energy orientation will have to become also information oriented. The power industry from a fully analog business where consumer data was recorded annually with pen and paper is going to become the most data intensive industry with enormous quantity of data generated every day.

In this paper, we have made projections and calculations on how the relatively small grid of The Netherlands could become a Big Data generator similar to Facebook. Certainly, sensing more data along the grid to monitor its performance is essential for the utilities and provides an improvement in their operation, it will also ease the billing process and it will help the inclusion of renewable sources. Interesting challenges are not only in the management of the future smart grid data, but also on how to extract value from this data and on novel business models based on the availability of Big Data.

ACKNOWLEDGEMENTS

We thank Frank Blauw for useful comments on a previous version of the article. The work is supported by the Dutch National Research Council under the NWO Smart Energy Systems programme, contract no. 647.000.004. Pagani is supported by University of Groningen with the Ubbo Emmius Fellowship 2009 and IBM PhD Fellowship 2013-14.

REFERENCES

[1] Netbeheer Nederland, "Betrouwbaarheid van elektriciteitsnetten in nederland - resulta ten 2012," Netbeheer Nederland, Tech. Rep. RM-ME-13L10440006, 2013.

[2] M. Hilbert and P. López, "The world's technological capacity to store, communicate, and compute information," *Science*, vol. 332, no. 6025, pp. 60–65, 2011.

[3] Cisco Systems, "Cisco Visual Networking Index: Forecast and Methodology, 2011-2016," Cisco Systems, Tech. Rep., 2012.

[4] M. G. Morgan, J. Apt, L. B. Lave, M. D. Ilic, M. Sirbu, and J. M. Peha, "The many meanings of "smart grid"," Carniege Mellon University, Tech. Rep., 2009.

[5] National Energy Technology Laboratory, "A system view of the modern grid," U.S. Department of Energy - Office of Electricity Delivery and Energy Reliability, Tech. Rep., 2007.

[6] C. Cuijpers and B.-J. Koops, "Smart metering and privacy in europe: Lessons from the dutch case," in *European Data Protection: Coming of Age*, ser. S. Gutwirth et al. (eds). Springer, 2013, pp. 269–293.

[7] I. Georgievski, V. Degeler, G. Pagani, T. A. Nguyen, A. Lazovik, and M. Aiello, "Optimizing energy costs for offices connected to the smart grid," *Smart Grid, IEEE Transactions on*, vol. 3, no. 4, pp. 2273–2285, 2012.

[8] M. Rosas-Casals and B. Corominas-Murtra, "Assessing European power grid reliability by means of topological measures," *Trans. of Ecology and the Environment*, no. 121, pp. 515–525, 2009.

[9] Netbeheer Nederland, "Dutch smart meter requirements v. 4.0," Netbeheer Nederland, Tech. Rep., 2011.

[10] G. Sanchez, I. Gomez, J. Luque, J. Benjumea, and O. Rivera, "Using internet protocols to implement iec 60870-5 telecontrol functions," *Power Delivery, IEEE Transactions on*, vol. 25, no. 1, pp. 407–416, 2010.

[11] X. Miao, X. Chen, X. ming Ma, G. Liu, H. Feng, and X. Song, "Comparing smart grid technology standards roadmap of the iec, nist and sgcc," in *Electricity Distribution (CICED), 2012 China International Conference on*, 2012, pp. 1–4.

[12] E. Kaldeli, E. U. Warriach, A. Lazovik, and M. Aiello, "Coordinating the web of services for a smart home," *ACM TWEB*, vol. 7, no. 2, p. 10, 2013.

[13] G. A. Pagani and M. Aiello, "Service orientation and the smart grid state and trends," *Service Oriented Computing and Applications*, vol. 6, no. 3, pp. 267–282, 2012.

[14] W. W. Peterson and E. J. Weldon, *Error-correcting codes*. The MIT Press, 1972.

[15] H. Khurana, M. Hadley, N. Lu, and D. Frincke, "Smart-grid security issues," *Security Privacy, IEEE*, vol. 8, no. 1, pp. 81–85, 2010.

[16] A. Metke and R. Ekl, "Security technology for smart grid networks," *Smart Grid, IEEE Transactions on*, vol. 1, no. 1, pp. 99–107, 2010.

[17] M. HADLEY, N. Lu, and A. DEBORAH, "Smart-grid security issues," *IEEE Security and Privacy*, vol. 8, no. 1, pp. 81–85, 2010.

[18] N. Capodieci, G. Cabri, G. A. Pagani, and M. Aiello, "An agent-based application to enable deregulated energy markets," in *36th Annual IEEE Computer Software and Applications Conference, COMPSAC 2012*, 2012, pp. 638–647.

[19] S. S. ahnd Sreenadh Reddy Depuru, L. Wang, and V. Devabhaktuni, "Electricity theft: Overview, issues, prevention and a smart meter based approach to control theft," *Energy Policy*, vol. 39, no. 2, pp. 1007 – 1015, 2011, special Section on Offshore wind power planning, economics and environment. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S030142151000861X

[20] C. Warmer, K. Kok, S. Karnouskos, A. Weidlich, D. Nestle, P. Selzam, J. Ringelstein, A. Dimeas, and S. Drenkard, "Web services for integration of smart houses in the smart grid," *Grid-Interop-The road to an interoperable grid, Denver, Colorado, USA*, pp. 17–19, 2009.

[21] S. Rusitschka, K. Eger, and C. Gerdes, "Smart grid data cloud: A model for utilizing cloud computing in the smart grid domain," in *Smart Grid Communications (SmartGridComm), 2010 First IEEE Int. Conf. on*, 2010, pp. 483–488.

[22] P. Parikh, M. Kanabar, and T. Sidhu, "Opportunities and challenges of wireless communication technologies for smart grid applications," in *Power and Energy Society General Meeting, 2010 IEEE*, 2010, pp. 1–7.

[23] C. Efthymiou and G. Kalogridis, "Smart grid privacy via anonymization of smart metering data," in *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, 2010, pp. 238–243.

[24] Y.-J. Kim, M. Thottan, V. Kolesnikov, and W. Lee, "A secure decentralized data-centric information infrastructure for smart grid," *Comm. Mag., IEEE*, vol. 48, no. 11, pp. 58–65, 2010.

[25] S. Amin and B. Wollenberg, "Toward a smart grid: power delivery for the 21st century," *Power and Energy Magazine, IEEE*, vol. 3, no. 5, pp. 34–41, 2005.

[26] L. dos Santos, A. G. da Silva, B. Jacquin, M. Picard, D. Worms, and C. Bernard, "Massive smart meter data storage and processing on top of hadoop," in *BigData Workshop, VLDB 2012*, 2012.

[27] IBM, "Managing big data for smart grids and smart meters," IBM Corporation, Tech. Rep., 2012.

[28] G. A. Pagani and M. Aiello, "Towards decentralization: A topological investigation of the medium and low voltage grids," *IEEE Trans. Smart Grid*, vol. 2, no. 3, pp. 538 –547, 2011.

[29] ——, "The power grid as a complex network: a survey," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 1, pp. 2688–2700, 2013.

[30] K. Sayood, *Introduction to data compression*. Morgan Kaufmann, 2012.

9th International Symposium
Advances in Artificial Intelligence and Applications

THE AAIA'14 will bring researchers, developers, practitioners, and users to present their latest research, results, and ideas in all areas of artificial intelligence. We hope that theory and successful applications presented at the AAIA'14 will be of interest to researchers and practitioners who want to know about both theoretical advances and latest applied developments in Artificial Intelligence. As such AAIA'14 will provide a forum for the exchange of ideas between theoreticians and practitioners to address the important issues.

Topics

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

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

Professor Zdzislaw Pawlak Best Paper Awards

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

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

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

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

IFSA Award for Young Scientist

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

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

Founding Chairs

Kwaśnicka, Halina, Wroclaw University of Technology, Poland

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

Event Chairs

Krawczyk, Bartosz, Wroclaw University of Technology, Poland

Slezak, Dominik, University of Warsaw & Infobright Inc., Poland

Program Committee

Bartkowiak, Anna, Wroclaw University, Poland

Bazan, Jan, University of Rzeszów, Poland

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

Budnik, Mateusz, University of Grenoble, France

Błaszczyński, Jerzy, Poznan University of Technology, Poland

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

Czarnowski, Ireneusz, Gdynia Maritime University, Poland

Herrera, Francisco, University of Granada, Spain

Hippe, Zdzislaw, University of Information Technology and Management in Rzeszow, Poland

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

Korbicz, Józef, University of Zielona Gora, Poland

Kwaśnicka, Halina, Wroclaw University of Technology

Marek, Victor, University of Kentucky, United States

Markowska-Kaczmar, Urszula, Wroclaw University of Technology

Mercier-Laurent, Eunika, IAE Lyon3, France

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

Myszkowski, Pawel, Wroclaw University of Technology, Poland

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

Nguyen, Hung Son, University of Warsaw, Poland

Porta, Marco, University of Pavia, Italy

Ramanna, Sheela, University of Winnipeg, Canada

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 19–26
DOI: 10.15439/2014F368
ACSIS, Vol. 2

Neural network approach to ECT inverse problem solving for estimation of gravitational solids flow

Hela Garbaa
Email: helagarbaa@gmail.com

Lidia Jackowska-Strumiłło
lidia_js@kis.p.lodz.pl

Krzysztof Grudzień
kgrudzi@kis.p.lodz.pl

Andrzej Romanowski
androm@kis.p.lodz.pl

Lodz University of Technology, Institute of Applied Computer Science, Poland.

***Abstract*—A new method to solve the inverse problem of electrical capacitance tomography is proposed. Our method is based on artificial neural network to estimate the radius of an object present inside a pipeline. This information is useful to predict the distribution of material inside the pipe. The capacitance data used to train and test the neural network is simulated on Matlab using the electrical capacitance tomography toolkit ECTsim. The provided accuracy is promising and shows efficiency to solve the inverse problem in a simple manner and on reduced computational time about 120 times when compared to the existing Landweber iterative algorithm for tomographic image reconstruction that can be encouraging for dynamic industrial applications.**

***Index terms*—Electrical Capacitance Tomography, Inverse Problem, Artificial Neural Networks, Gravitational Flow of Solids.**

I. Introduction

ELECTRICAL Capacitance Tomography (ECT) is a non-invasive technique used to image the spatial distribution of merged materials with different dielectric properties inside a pipe [1]. The spatial distribution is determined by measuring the capacitances between all pairs of electrodes placed around the vessel containing the process to be examined [2]. The provided measurements depend on the electrical permittivity value of the combined materials and their spread inside the isolated pipe. Studying the relationship between capacitance records and permittivity distribution and converting them to an ECT image has been an attractive research era since 1980's [13, 15, 16]. Researchers are still investigating to improve the performances of this technique and agree on the complexity of the task due to the difficulties with inverse problem solution, nonlinearity of the system and the limited number of obtained capacitances. Typical used sensors are with N electrodes (e.g. N=8, 12, 16) lead to M capacitance measurements (M= 28, 66, 120) respectively and typical generated 2D ECT image with resolution of 32x32 pixels, which in turn makes the inverse problem ill-posed.

Forward and inverse problems are in fact the two main tasks in tomography visualization. First one is looking for the measurement records given permittivity distribution (in case of ECT tomography) while the later one is nothing else but looking for a relationship of the measurement (result) with the source data (cause) that is looking for the cause, which gives the measurement. In most cases, a forward problem is solved by numerical methods with the finite elements method (FEM) being the most popular one. FEM allows calculating the inter-capacitance values relying on the known permittivity distribution.

Typical results of inverse problem solution are the reconstructed images of ECT sensor space. Depending on type of the applied reconstruction algorithm the image quality can vary. Choice of appropriate reconstruction method for industrial application is limited by time of calculation in real time. In the case of iterative reconstruction methods the reconstructed image is updated iteratively until reaching a satisfactory error based on the difference between the calculated and real capacitances.

Authors in [4] reviewed various Iterative ECT methods: Newton–Raphson, Landweber iteration and Algebraic Reconstruction Techniques. Evaluation and simulations results highlighted the superiority of the performances of Landweber iterative algorithm in terms of lowest capacitance calculation errors comparing to the other cited algorithms. Artificial Neural Networks (ANN) have been used for solving both ECT problems since they represent a powerful and effective tool to dealing with complex and non-linear computations. The type of the applied neural network differs depending on the purpose of investigations. A multi-layer Feed-Forward Network (FFN) was applied to solve the forward problem. The network was trained to predict capacitance data from different permittivity distributions and then when integrated with Landweber iteration method provided a satisfactory quality of reconstructed image [3]. 313 parallel multi-layer perceptron with 2 hidden layers each

of which were applied in [5] to reconstruct different ECT image pixels and visualize the oil distribution inside a pipe.

Large radial basis function (RBF) neural network is used in [7], with 66 nodes, representing the measured capacitances, in the input layer and 804 neurons in the output layer to flow patterns image reconstruction. The 804 outputs of the network correspond to the mesh grid elements considered to model the ECT phantoms in an earlier step.

In our present work we propose to use a Multi-Layer Perceptron with one hidden layer to estimate the radius of an object present inside the sensor and thus reconstruct the tomographic image. Our approach has the advantage to have a simple neural network with simplified structure: 66 capacitances form the input layer and one neuron at the output which corresponds to the radius of the phantom. The performance rate is set at 0.004 while it was equal to 0.3 in [7].

II. ECT INVERSE PROBLEM

Electrical capacitance tomography inverse problem is to, by means of appropriate algorithm; determine permittivity distribution based on measured capacitances into a form of a tomographic image [10, 14]. Fig. 1 shows a schematic of an image reconstruction process in ECT

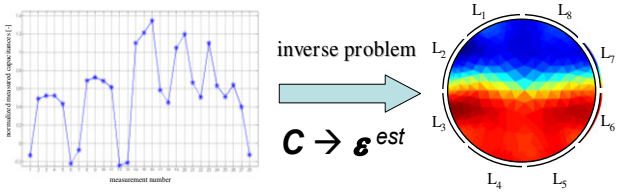


Fig. 1.Example of a graph of normalized measured capacitances and corresponding reconstructed tomography image following the solving of permittivity matrix using the inverse problem method.

Two major characteristic obstacles of inverse problem are: its ill-posed and non-linear properties. The first refers to the limited number of measured capacitances compared to the number of pixels to be reconstructed while the relationship between capacitance and permittivity distribution nonlinearity is modeled by the Gauss Law, [11, 13, 14] (Eq.1):

$$C = \frac{Q}{V} = -\frac{1}{V}\iint_{\Gamma} \varepsilon(x,y)\nabla\phi(x,y)\,d\Gamma, \quad (1)$$

where Q is the electric charge; V the potential difference between two electrodes, ε(x,y) denotes the permittivity and φ(x,y) represents the electrical potential distributions. Γ stands for the electrode surface and $d\Gamma$ an element orthogonal to this surface.

A discrete linear approximation of the previous equation is formulated as following (Eq. 2):

$$C = \mathbf{S}\,\varepsilon, \quad (2)$$

where C is a vector of measured capacitances, **S** linearized sensitivity matrix and ε is a vector of permittivity distribution. The problem of image reconstruction is then reduced to solving the linear discrete form (2) which is still challenging and attracting researchers 'efforts. The problem of image reconstruction is then reduced to solving equation (3) represented:

$$\varepsilon = S^{-1}\,C \quad (3)$$

The inverse of **S** does not exist; because **S** is not a square matrix (number of measurements is not equal of number of pixels in image). Instead of having its inverse S^{-1}, the pseudo-inverse **S*** must be calculated, for instance, using an approximated solution. Other important problem in image reconstruction procedure is the dependence of the sensitivity matrix on the permittivity distribution, what causes that the inverse problem is non-linear. A lot of publications where different kind of methods to solve the inverse problem (linear, non-linear, directly methods and iterative methods) are presented in literature [3]. Depending on computational method, the algorithm generates images with different levels of quality, in the case of LBP algorithm or much higher quality, using iterative or non-linear methods. In the case of iterative and non-linear methods, high image quality is occupied by long computational time.

The Landweber iterative algorithm is one of the most popular methods in the field of ECT image reconstruction. The iteration process, in Landweber algorithm is governed by the following formula [8, 14]

$$\varepsilon_{k+1} = \varepsilon_k - l\,S^T(S\varepsilon_k - C) \quad (4)$$

where ε_{k+1} and ε_k are the estimated permittivity distributions at the k^{th} and $(k+1)^{th}$ iterations respectively, S is the calculated sensitivity matrix and l is a relaxation parameter of Landweber algorithm.

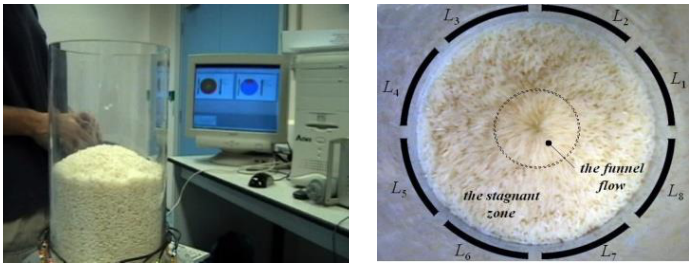
The method cited above owns the advantages of easy implementation and low computational complexity but suffers from the numerical optimization point of view as it possess a relatively low convergence rate and hardly provides a global optimization solution, [12].

Artificial Neural Network constitutes a competitive optimization based- method applied in the same research era [3, 17].

III. PROBLEM STATEMENT

The extraction of industrial process characteristic parameters gives possibility of predicting unwanted incidents and feasibility of in depth exploration of dynamic spatial temporal phenomena occurring during industrial process such as flows [13, 15, 16, 18-20]. Visualization of flow processes by means of tomography image reconstruction gives a non-invasive tool for extraction of these parameters and given sufficient reconstruction time combined with fast image processing algorithms can promise on-the-fly flow characterization.

One of the examples of successful ECT application for dynamic flow is hopper gravitational discharging process [19, 20]. Model of such process is depicted on Fig. 2. where the laboratory setup photograph is presented. The mentioned type of flow is widely present in a range of branches such as pharmaceutical, chemical, food processing, construction and others. Processing of 2D reconstructed images allowed examination of hopper discharging funnel type of flow parameters. The set of parameters such as solids concentration in funnel and funnel area size characterize the dynamics of hopper flow [18]. This knowledge gives the information about correct/incorrect hopper flow.



a) b)

Fig. 2. Photos of the hopper flow model. Photos of the hopper model: a) side view of the hopper and the image reconstruction visualisation; b) top view of the container with 8 sensor electrodes depicted around the silo.

The so-called funnel flow occurring during the silo discharging process not obvious to be analysed in its full volume since the non-transparent nature of the process. Hence, ECT is the ultimate tool to examine this flow on the base of reconstructed image where funnel area, in the center of silo, with the smaller material concentration value than rest of the sensor space can be observed (Fig. 3)

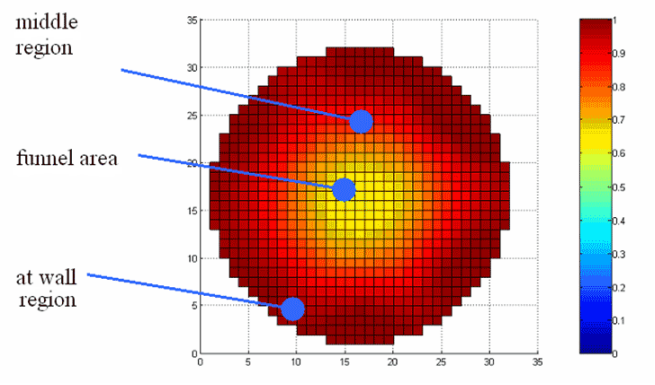


Fig. 3. Basic ECT tomographic image reconstructed for funnel hopper flow with characteristic flow areas indicated.

In order to develop the method for calculating these characteristic parameters a scheme of data processing was proposed. It allows estimating the radius of the object of a different permittivity inside a sensor cross-section space. As a proof-of-concept study we present a simulation with a phantom as useful information to predict the distribution inside the vessel knowing the capacitance measurements. A uniform distribution inside a circular sensor and a circular object is situated at the center of the vessel was assumed for calculations as shown in Fig. 4.

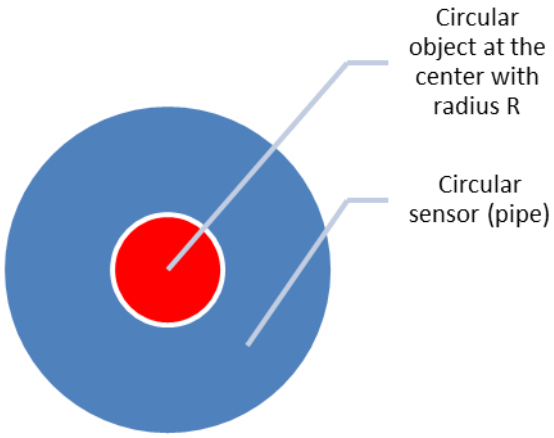


Fig. 4. Phantom considered with circular sensor and circular object in the center.

We present, in the following sections, the ECT inverse problem solving by use of Artificial Neural Networks, an evaluation of our method by comparing it to Landweber iterative approach and mention further works.

IV. ARTIFICIAL NEURAL NETWORKS

An artificial neural network is a powerful information processing tool recognized for its abilities to model complex input/output relationships and to learn these relationships directly from the data being modeled. They are principally designed to mimic the human brain functions in the following two ways: they acquire knowledge through a learning process, and the knowledge is stored within inter-neuron connection strengths known as synaptic weights, [6].

Two learning approaches are available: the first one is supervised and the second is unsupervised. A supervised learning, mostly applied, requires a desired output in order to learn. To perform the task of learning a training algorithm is applied in order to adjust the synaptic weights of the network in an orderly fashion so as to generate a model that maps the input to the output using historical data. The provided model can then be generalized to produce the output when the desired output is unknown. In addition to a powerful training algorithm, a neural network needs to have an appropriate structure to deal with the complexity of the problem to be solved. The most commonly used network structure, which has been also used in the presented approach, is the Multi-Layer Perceptron (MLP). An example of a MLP network is shown in figure 5.

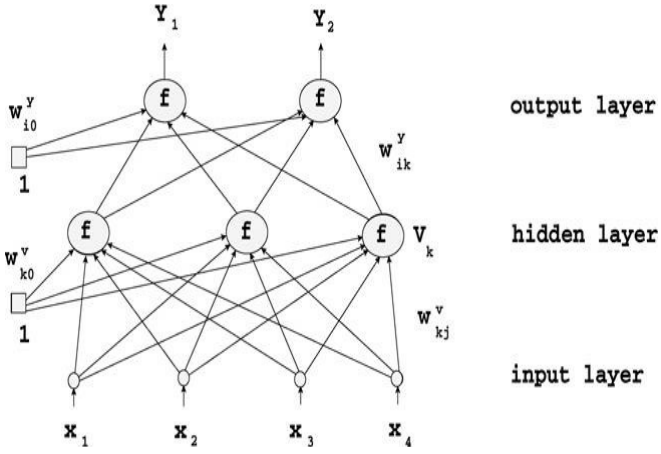


Fig.5. An example of multilayer perceptron network with one hidden layer.

The inputs are fed into the input layer and get multiplied by interconnection weights as they are passed from the input layer to the hidden layer. Within the hidden layer, they get summed then processed by a nonlinear function called activation function. The data is multiplied by interconnection weights then processed one last time within the output layer to produce the neural network outputs. A mathematical representation of the signal of i−th network output Yi is given by the Eq. (5):

$$Y_i = f\left(\sum_{k=0}^{M} \omega_{ik}^{y}\, y_k\right) = f\left(\sum_{k=0}^{M} \omega_{ik}^{y}\, f\left(\sum_{j=0}^{N} \omega_{kj}^{v}\, x_j\right)\right) \quad (5)$$

Where: N—number of inputs, M—number of neurons in the hidden layer, w_{kj}—weight of kth neuron in the hidden layer for the j th input, w_{yik}—weight of ith neuron in the output layer for the signal y_k, which is the output of the kth neuron

The activation function f (·) may be a simple linear or a non-linear function. The most commonly used activation functions are: threshold function, sigmoidal function, hyperbolic tangent function and radial basis function. In our case we used the sigmoidal function in both hidden and output layers. Sigmoidal function is mathematically expressed as:

$$f(z) = \frac{1}{1+e^{-z}} \quad (6)$$

ANN-based inverse model is built on the basis of relations between the network input and output vectors. The knowledge about the inverse mapping is stored within the network structure and network connection weights [6,22]. Sixty six values of capacitances **C**= [$C_{1,\ldots}$, C_{66}] constitute the network input vector. The approximated values of the corresponding radius $\hat{R}$ are calculated at the network output. An unknown mapping of the input vector to the output vector is approximated in an iterative procedure known as neural network training [5]. The objective of the learning algorithm is to adjust network weights on the basis of a given set of input-output pairs for a given cost function to be minimized. Back propagation algorithm, a supervised learning network algorithm, uses the gradient of the performance function to determine how to adjust the weights to minimize performance. In back propagation, the error data is propagated from the output layer backwards through the network. The effected computations allow the update of the incoming weights at each layer. In our present approach, during the network learning phase the error is propagated until a set value of the training error is reached.

V. EXPERIMENTAL PART

A. Description

We treat on our present work the inverse problem which aims to determine the material distribution relying on measured capacitances. The simulation is done using ECTSIM Matlab's toolbox, [8, 21]. ECTSIM was designed to evaluate existing image reconstruction algorithms applied on the field of ECT like Landweber algorithm and LBP method. Our work can be described on 2 main steps: (1) Sensor Modeling and capacitances measurement (2) Image Reconstruction using Artificial Neural Network: A multi-layer Perceptron (MLP) is applied to determine the phantom radius). We enclose the experimental part by comparing the obtained results with our method with others obtained with Landweber algorithm. Computations were performed on PC computer with Intel(R) Xeon(R) CPU E5630 @ 2.53GHz processor and 24,0 GB RAM. The different algorithms were implemented in MATLAB.

We detail on the following paragraphs the steps mentioned above.

B. Sensor Modeling and capacitances measurement

The current work is done under the 2D version of the ECTSIM toolbox using the circular sensor model. The provided sensor model is composed of four layers: pipe layer, electrode layer, insulation layer and the screen layer as shown on Fig.3. As the user of the toolbox is able to set the different parameters of each layer like thickness, electrical permittivity of insulation material and number of electrodes, we chose a sensor with 12 electrodes (N= 12) placed inside the pipe (inner insulator thickness=0), background permittivity =1 and elements' permittivity =3. The field of view diameter was set to 84 mm. The sensor field of the view was then divided into 96x96 square meshes. For each phantom the simulated capacitances is calculated using the discrete linear approximation of Gauss Law (Eq.2).

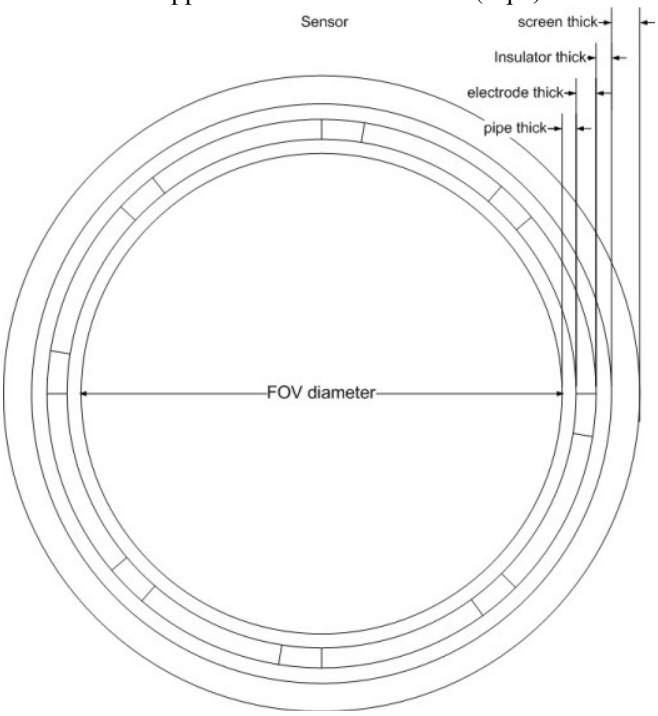


Fig.6. Sensor's diagram used for simulation [21].

We varied the diameter of the object present inside pipe in order to generate the phantoms to train and test the ANN in a farther step.

C. Image Reconstruction using Artificial Neural Network

For the given parameters we generated 320 phantoms divided into 241 training examples and 80 test examples. A Multi-Layer Perceptron with a single hidden layer is applied to estimate the radius of the object inside the pipe. The number of nodes in the input layer is given by the number of measured capacitances L= N (N-1)/2 = 66 and one neuron at the output layer. The network was trained using the back propagation algorithm with a training error set to E= 1/240=0.004. We made several experiences to determine the number of nodes in the hidden layer. Table1 summarizes the obtained testing errors with different numbers of hidden neurons.

We consider on the selection of the appropriate structure of MLP :

- Number of iterations at the learning phase
- Mean Square Error (MSE) during the test process:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(R - \widehat{R})^2 \quad (7)$$

- Mean testing error:

$$\text{Mean} = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

where $e_i(R) = (R - \hat{R})$ (8)

We designate by R the desired MLP output / desired radius, $\hat{R}$ the estimated radius at the testing phase and n the number of testing examples. Table1 summarizes the obtained testing errors with different numbers of hidden neurons.

TABLE 1.
TESTING ERRORS AND NUMBER OF ITERATIONS WITH DIFFERENT MLP STRUCTURES

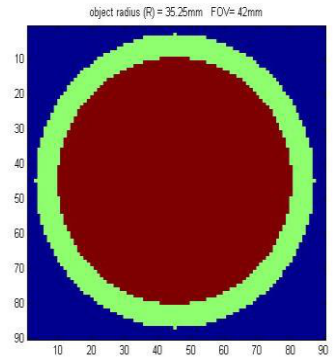
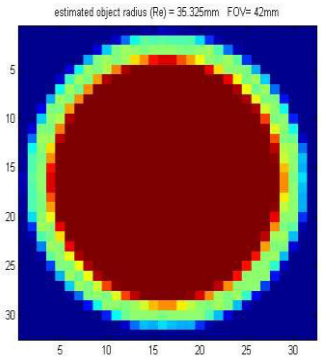
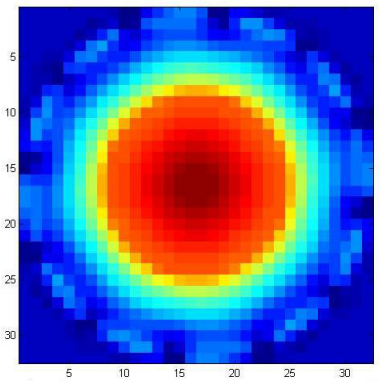
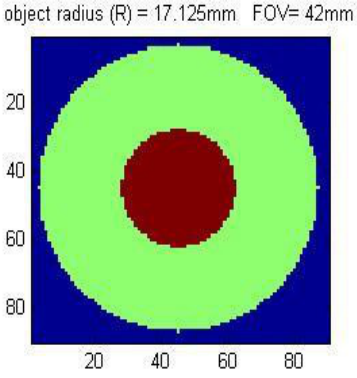
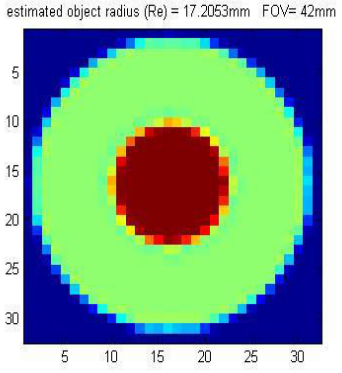
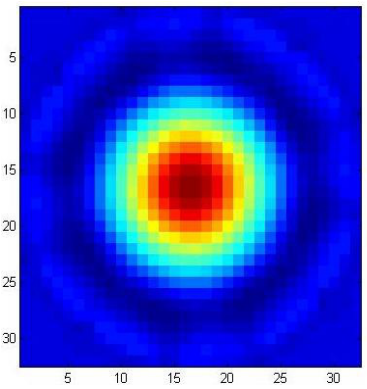
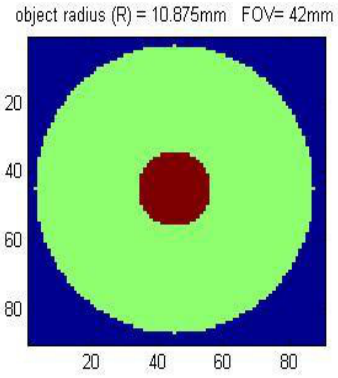
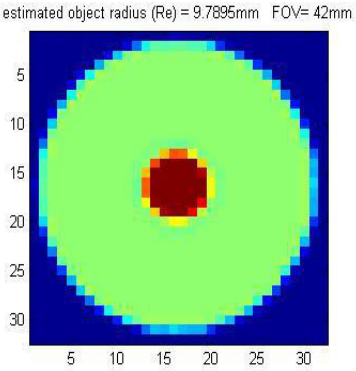
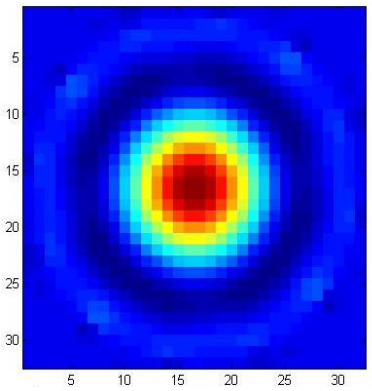
Network structure	Testing error		Number of iterations (learning)
	MSE	Mean	
(66-10-1)	**0.1670**	**0.2839**	**1080885**
(66-20-1)	0.1688	0.3148	1419490
(66-30-1)	0.1721	0.2846	1815935
(66-66-1)	0.1832	0.4086	1225485

The provided results show that minimum testing errors MSE=0.1670 and Mean=0.2839 are obtained for a network structure with 10 hidden neurons (66-10-1). The maximum errors were obtained with 66 neurons in the hidden layer. The use of both Mean Square Error and the Mean Error was beneficial to decide the best structure to retain for the reconstruction phase since the obtained MSE errors were close for the two first structures (66-10-1) and (66-20-1). The learning process was the fastest for the given structure and was the slowest for the structure (66-30-1) as the smaller the network structure is the less numerical complexity we have.

To consider the MLP structure (66-10-1) the appropriate structure to solve our handled problem, the process of learning and testing the selected was repeated several times and results uphold that 10 neurons in the hidden layer are sufficient to estimate the radius of the object inside the pipe.

A second attempt to evaluate the performances of the proposed method is to compare the reconstructed images from ANN with images reconstructed with Landweber algorithm under the ECTsim toolbox.

TABLE. 2.
RECONSTRUCTED IMAGES FROM MLP AND LANDWEBER ALGORITHM

Desired Phantom / Distribution	**Image reconstructed from MLP**	**Image reconstructed from Landweber/ Time elapsed for reconstruction (in s)**
		
	time elapsed for reconstruction (in s) 0.070273	time elapsed for reconstruction (in s) 10.22
		
	time elapsed for reconstruction (s) 0.075846	time elapsed for reconstruction (s) 9.910792
		
	time elapsed for reconstruction (s) 0.082285	time elapsed for reconstruction (s) 10.112285

D.Evaluation of the proposed method

We present in table2. the reconstructed images from MLP network and Landweber algorithm for different radius taken from the neural network test base. The reconstruction from MLP is performed by drawing the circular object with the estimated radius. The elapsed time included one iteration to estimate the radius from the capacitances fed to the input of the MLP with the retained structure (66-10-1) and the time to draw the outer circular sensor and the object inside. The number of iterations to reconstruct the same object, using Landweber algorithm, with the same radius is set equal to 100.

The obtained images from the neural network method present satisfactory shape with small relative radius estimation error. The circular shape is better maintained with MLP based method especially in boundaries. Other advantage of the proposed methods is better estimation of permittivity value for object. In the case of MLP the relative permittivity value is much closer to the prepared simulated phantom than for Landweber algorithms. Next aspect is the time calculation. The phantom reconstruction with MLP method is about 120 times faster than with Landweber iteration algorithm.

The relative testing error for different values of radius was calculated based on formula:

$$E_R = \frac{R-\hat{R}}{R} \quad (9)$$

where R is desired radius and $\hat{R}$ is the estimated radius.

The results are shown in Fig.7. The error is significant, for small values of radiuses - about 1 mm. For R<5 mm the relative error value is still negative which means that the estimated value is bigger than the real value. For 5mm<R< 15mm, the relative error value becomes positive and the gap between the desired and estimated values is smaller. For radius values higher than 15 mm the relative testing error is near to zero. The high pic of relative error could be referred either to the capacitances measurement process (i.e. the difficulty to sense the permittivity distribution in a very small objects) or to the lack of training examples for this interval of radii. Providing more training examples for small objects (R<15mm) would improve the performances of our neural network.

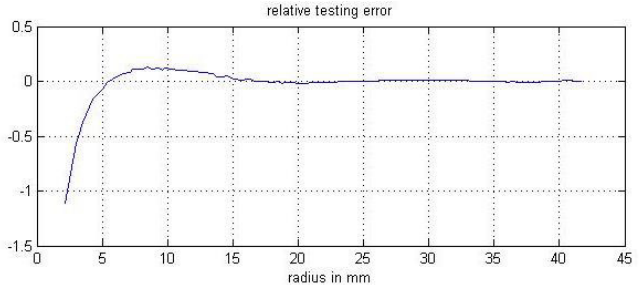


Fig. 7. Relative testing error in term of the object radius.

VI. CONCLUSION

The aim of the work was to estimate the radius of the object present inside a circular cross section ECT sensor based on neural network technique. The calculated radius was then used to image the distribution of content inside the sensor. The different steps for solving the inverse ECT problem were depicted. The obtained results are promising especially under a simple MLP structure (66-10-1) and back-propagation training algorithm. The provided accuracy is satisfactory and ANN based approach allowed to solve the inverse problem in a simple manner and on reduced computational time about 120 times when compared to the existing Landweber iterative algorithm. Results revealed potential to estimate flow patterns such as funnel-type hopper flow with reconstruction speed sufficient for on-the-fly industrial applications. Possibility to estimate more than one useful parameter (such as radius and permittivity) and generalization of the pattern of phantoms will be the subject to further work.

ACKNOWLEDGMENTS

Work partially funded by the European Commission under the Erasmus Mundus GreenIT project (GreenIT for the benefit of civil society. 3772227-1-2012-ES-ERA MUNDUS-EMA21; Grant Agreement nº 2012-2625/001-001-EMA2)

REFERENCES

[1] Chathurika M.G.P. Mediwaththe, Kasun E. Wijethilake, Damith B.W. Abeywardana, Sachini E. Wijethilake, Janaka V. Wijayakulasooriya, Non Invasive Cross Sectional Imaging Using Electric Capacitance Tomography, 6th International Conference on Industrial and Information Systems, ICIIS 2011, pp234-238, 10.1109/ICIINFS.2011.6038072.

[2] Norberto Flores, J Carlos Gamio, Carlos Ortiz-Alemán and Enrique Damián ,Sensor modeling for an electrical capacitance tomography system applied to oil industry, Excerpt from the Proceedings of the COMSOL Multiphysics User's Conference 2005 Boston.

[3] Qussai Marashdeh, Warsito Warsito, Liang-Shih Fan, and Fernando L. Teixeira, Nonlinear Forward Problem Solution for Electrical Capacitance Tomography Using Feed-Forward Neural Network, IEEE SENSORS JOURNAL, VOL. 6, NO. 2, APRIL 2006, pp441-449, 10.1109/JSEN.2005.860316.

[4] W Q Yang and Lihui Peng, Image reconstruction algorithms for electrical capacitance tomography, Measurement Science and Technology, Vol.44,No.1, January2003, 10.1088/0957-0233/14/1/201.

[5] Norberto Flores, Ángel Kuri-Morales, Carlos Gamio , An Application of Neural Networks for Image Reconstruction in Electrical Capacitance Tomography Applied to Oil Industry, Progress in Pattern Recognition, Image Analysis and Applications,Lecture Notes in Computer Science, Vol 4225, 2006, pp 371-380, 10.1007/11892755_38.

[6] L. Jackowska-Strumillo, J. Sokolowski, A. Żochowski and A. Henrot, On Numerical Solution of Shape Inverse Problems, Computational Optimization and Applications , Vo. 23, Issue 2, ,pp 231-255, 10.1023/A:1020528902875.

[7] Jianwei Li, Xiaoguang Yang, Youhua Wang and Ruzheng Pan , An Image Reconstruction Algorithm Based on RBF Neural Network for Electrical Capacitance Tomography, Sixth International Conference on Electromagnetic Field Problems and Applications (ICEF), 2012,pp1-4, 10.1109/ICEF.2012.6310416.

[8] Smolik W and Radomski D, The matlab's toolbox for iterative image reconstruction in electrical capacitance tomography, 5th Int.Symp.on Process tomography (Poland), pp98-103.

[9] Ziqiang Cui, Chengyi Yang, Benyuan Sun, Huaxiang Wang, Liquid film thichness estimation using electrical capacitance tomography, MEASUREMENT SCIENCE REVIEW, Volume 14, No. 1, 2014, pp8-15, 10.2478/msr-2014-0002.

[10] Yunjie Yang and Lihui Peng, Data Pattern With ECT Sensor and Its Impact on Image Reconstruction, IEEE SENSORS JOURNAL, VOL. 13, NO. 5, MAY 2013, pp1582-1593, 10.1109/JSEN.2013.2237763.

[11] Jing Lei , Shi Liu , Xueyao Wang and Qibin Liu ,An Image Reconstruction Algorithm for Electrical Capacitance Tomography Based on Robust Principle Component Analysis, Sensors2013, VOL.13, pp 2076-2092, 10.3390/s130202076.

[12] Jing Lei and Shi Liu, Dynamic Inversion Approach for ElectricalCapacitance Tomography, IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, VOL. 62, NO. 11, November2013, pp3035-3049, 10.1109/TIM.2013.2270039 .

[13] Scott DM., McCann H.: Process Imaging for automatic control, Taylor and Francis Group, p.439, 2005.

[14] Lionheart WRB.: Review: Developments in EIT reconstruction algorithms: pitfalls, challenges and recent development. Physiol. Meas., Vol. 25, 125-142, 2004, 10.1088/0967-3334/25/1/021.

[15] Yang WQ., Liu S.: Role of tomography in gas/solids flow measurement, Flow Meas. and Instrum., Vol. 11, pp. 237-244, 2000, 10.1016/S09555986(00)000236.

[16] Dyakowski T., Edwards R.B., Xie C.G., Williams R.A: Application of capacitance tomography to gas-solid flows, Chemical Engineering Science, Vol.52, pp. 2099-2110, 1997, 10.1016/S0009-2509(97)00037-7.

[17] Warsito W. and Fan L-S (2003) Development of 3-Dimensional Electrical Capacitance Tomography Based on Neural Network Multi-criterion Optimization Image Reconstruction, Proc. of 3rd World Congress on Industrial Process Tomography (Banff), 2003, 942-947.

[18] Romanowski A., K. Grudzien, R.A. Williams Analysis and interpretation of hopper flow behavior using electrical capacitance tomography Part. Part. Syst. Charact., 23 (2006), pp. 297–305, 10.1002/ppsc.200601060.

[19] Niedostatkiewicz M., Tejchman J., Chaniecki Z., Grudzień K.: Determination of bulk solid concentration changes during granular flow in a silo with ECT sensors. Chemical Engineering Science. 64, 2008, pp. 20-30.

[20] Chaniecki Z., Dyakowski T., Niedostatkiewicz M., Sankowski D.: Application of electrical capacitance tomography for bulk solids flow analysis in silos. Particle & Particle Systems Characterization. 23, 3-4, 2006, pp. 306-312, 10.1002/ppsc.200601061.

[21] ECTSim 3D MATLAB's Toolbox,http://ectsim.ire.pw.edu.pl/, retrieved on April 2014.

[22] Jackowska-Strumiłło L., Sokołowski J., Żochowski A.: Topological Derivative and Training Neural Networks for Inverse Problems. In: W. Duch, J. Kacprzyk, E. Oja (Editors), et al. Artificial Neural Networks: Biological Inspirations – ICANN 2005: 15th International Conference, Warsaw, Poland, September 11-15, 2005. Lecture Notes in Computer Science, Vol. 3697/2005, Springer-Verlag GmbH, pp. 391-396, 10.1007/11550907_62.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 27–34
DOI: 10.15439/2014F492
ACSIS, Vol. 2

Adaptive Learning for Improving Semantic Tagging of Scientific Articles

Andrzej Janusz
Institute of Mathematics,
The University of Warsaw,
Banacha 2, 02-097 Warszawa, Poland
janusza@mimuw.edu.pl

Sebastian Stawicki
Institute of Mathematics,
The University of Warsaw,
Banacha 2, 02-097 Warszawa, Poland
stawicki@mimuw.edu.pl

Hung Son Nguyen
Institute of Mathematics,
The University of Warsaw,
Banacha 2, 02-097 Warszawa, Poland
son@mimuw.edu.pl

***Abstract*—In this paper we consider a problem of automatic labeling of textual data with concepts explicitly defined in an external knowledge base. We describe our tagging system and we also present a framework for adaptive learning of associations between terms or phrases from the texts and the concepts. Those associations are then utilized by our semantic interpreter, which is based on the Explicit Semantic Analysis (ESA) method, in order to label scientific articles indexed by our SONCA platform. Apart from the description of the learning algorithm, we show a few practical application examples of our system, in which it was used for tagging scientific articles with headings from the MeSH ontology, categories from ACM Computing Classification System and from OECD Fields of Science and Technology Classification.**

***Keywords*—semantic indexing; Explicit Semantic Analysis; Adaptive Semantic Analysis; multi-label tagging; adaptive learning;**

I. Introduction

THE MAIN idea of a keyword search is to look for texts (documents) that contain one or more words specified by a user. Then, using a dedicated ranking algorithm, relevance of the matching documents to the user query is predicted and the results are served as an ordered list [1]. In contrast, semantic search engines try to improve the search accuracy by understanding both, the user's information need and the contextual meaning of texts, which are then intelligently associated [2].

From the data processing point of view, the semantic search engine may be divided into three main components: semantic text representation module, interpretation and representation of a user query, and an intelligent matching algorithm [3]. The scope of the first two modules may be categorized as a semantic data representation [4]. In opposite to the keyword search, the semantic data representation, and thus the semantic indexes, cannot be calculated once and then utilized by intelligent matching algorithms. The text representation, as well as a query interpretation should be assessed with respect to the type of the users' group, a context of the words in the query and many others factors [5].

This work is partially supported by the National Centre for Research and Development (NCBiR) under Grant No. SP/I/1/77065/10 by the Strategic scientific research and experimental development program: "Interdisciplinary System for Interactive Scientific and Scientific-Technical Information" and by Polish National Science Centre (NCN) grants DEC-2011/01/B/ST6/03867 and DEC-2012/05/B/ST6/03215.

The better part of current search engines is based on a combination of a keyword search and sophisticated document ranking methods [1]. Only some of them process search queries, analyzing both, a query and documents' content with respect to their meaning, and return the semantically relevant search results [2]. However, even this approach becomes insufficient. The process of information retrieval needs to be made intelligently in order to help users in finding relevant information. The key role in this process is the recognition of the users' information needs and collecting feedback about the search effectiveness. The gathered information should be utilized to improve search algorithms and forge better responses to user requirements. Those challenges are in the scope of studies on adaptive search engines which interact with experts (users) and operate in a semantic representation space [6].

The SONCA (Search based on ONtologies and Compound Analytics) platform [7] is developed at the Faculty of Mathematics, Informatics and Mechanics of the University of Warsaw. It is a part of SYNAT project focusing on development of Interdisciplinary System for Interactive Scientific and Scientific-Technical Information (www.synat.pl). SONCA aims at extending the functionality of search engines by more efficient search of relevant documents, intelligent extraction and synthesis of information, as well as a more advanced interaction between users and knowledge sources.

Within the SYNAT project, some successful methods for the semantic text representation and indexing have already been developed [4], [8]. In this paper we discuss an adaptive learning model of terms-to-concepts associations which can be treated as an extension of the Explicit Semantic Analysis (ESA) method [9], [10]. By an analogy, we call it Adaptive Semantic Analysis (ASA). The main purpose of this model is to adjust the links between words and well-defined concepts. Those links are automatically derived from natural language definitions of the concepts which are stored in an external knowledge base. The associations are then used for labeling and indexing scientific articles. The definitions of the concepts can be extracted from different knowledge sources such as domain ontologies. In our experiments we show how the model can be constructed using descriptions of the concepts in a natural language and how it can be improved by using feedback from domain experts. We also show how to deal with a lack of concept descriptions in a case when there is available a sufficient number of training examples of labeled articles. Finally,

978-83-60810-58-3/$25.00

we present our most recent developments and improvements to the model, which are related to a problem of deciding how many concepts should be associated with a given document. As case studies we use a task of tagging biomedical articles from the PubMed repository with concepts from the MeSH ontology [11], a task of labeling abstracts of computer-science-related documents with terms from ACM Computing Classification System [12] and a problem of assigning the OECD Fields of Science and Technology Classification categories to articles from the Infona system (www.infona.pl).

II. Explicit Semantic Analysis

Explicit Semantic Analysis (ESA) proposed in [9] is a method for automatic tagging of textual data with Wikipedia concepts. It utilizes natural language texts of Wiki articles as textual representations of the corresponding concepts. It is assumed that the Wiki articles contain definitions of the concepts and describe their semantic. Those representations are regarded as a regular collection of texts and are matched against documents to find the best associations [10].

In ESA, the semantic relatedness between concepts and documents is computed two-fold. First, after the initial processing (tokenization, stemming, stop words removal), the corpus and the concept definitions are converted to the *bag-of-words* representation. Each of the distinct terms in the documents is given a weight expressing a strength of its association to the text. Assume that after the initial processing of a corpus consisting of M documents, $D = \{D_1, \ldots, D_M\}$, there have been identified N distinct terms (e.g. words, stems, n-grams) $t_1, \ldots, t_N$. Any text D_i in the corpus D can be represented by a vector $W_i = \langle w_{1,i}, \ldots, w_{N,i} \rangle \in \mathbb{R}_+^N$, where each coordinate $w_{j,i}$ expresses a value of some relatedness measure for j-th term in vocabulary (t_j), relative to this document. The most common measure used to calculate $w_{j,i}$ is the *tf-idf* (term frequency-inverse document frequency) index [1], defined as:

$$w_{j,i} = tf_{i,j} * idf_j = \frac{n_{i,j}}{\sum_{k=1}^{N} n_{i,k}} \log\left(\frac{M}{|\{i : n_{i,j} \neq 0\}|}\right), \quad (1)$$

where $n_{i,j}$ is the number of occurrences of the term t_j in the document D_i.

In the second step, the bag-of-words representation of the concept definitions is transformed into an inverted index that maps the terms $t_1, \ldots, t_N$ into lists of K concepts $C_1, \ldots, C_K$, described in an external knowledge source. The inverted index can be used as a semantic interpreter. Given a document from the corpus D, we may iterate over terms from the text, retrieve the corresponding entries from the inverted index and merge them into a vector of concept weights that represents the analyzed document.

Let $W_i = \langle w_{1,i}, \ldots, w_{j,i}, \ldots, w_{N,i} \rangle$ be a bag-of-words representation of an input document D_i, where $w_{j,i}$ is the tf-idf index of t_j defined by the formula (1). We can analogically quantify the association between the term t_j and the k-th concept C_k by computing the *bag-of-words* representations of the concept descriptions. Those associations constitutes the inverted index. Let $inv_{j,k}$ be the inverted index entry for the term t_j and the concept C_k. For convenience, all the weights $inv_{j,k}$ can be arranged in a sparse matrix structure with N rows and K columns, denoted by INV, such that $INV[j,k] = inv_{j,k}$ for any pair (j,k), such that $j = 1, \ldots, N$ and $k = 1, \ldots, K$. The new vector representation of D_i will be denoted by $V_i = \langle v_{1,i}, \ldots, v_{K,i} \rangle$ where:

$$v_{k,i} = \sum_{j : t_j \in D_i} w_{j,i} * inv_{j,k}. \quad (2)$$

In other words, the above equation expresses a standard dot product of the k-th column of the matrix INV and the vector W_i. This new representation will be called a *bag-of-concepts* of a document D_i.

For practical reasons it may also be useful to represent documents only by the most relevant concepts. In such a case, the association weights can be used to rank the concepts and to select only the top concepts from the ranked list. One can also apply some more sophisticated methods that involve utilization of internal relations in the knowledge base (e.g. for semantic clustering of concepts and assigning only the most representative ones to the documents).

The original purpose of Explicit Semantic Analysis was to provide means for computing semantic relatedness between texts. However, an intermediate result – weighted assignments of concepts to documents (induced by the term-concept weight matrix) may be naturally utilized in document retrieval as a semantic index [3], [5]. Although originally ESA was meant to utilize the Wikipedia articles as the external knowledge source, it seems reasonable that for specialized tasks, such as indexing articles from a specific branch of science, it is better to use concepts described in dedicated knowledge bases or ontologies. A user (an expert) may query a document retrieval engine for documents matching a given ontology concept. If the concepts are already assigned to documents, this problem is conceptually trivial. However such a situation is relatively rare, since the employment of experts who could manually labeled documents from a huge repository is expensive. On the other hand, the utilization of an automatic tagging method, such as ESA, allows to infer a labeling of previously untagged documents or at least it can support the experts in that task.

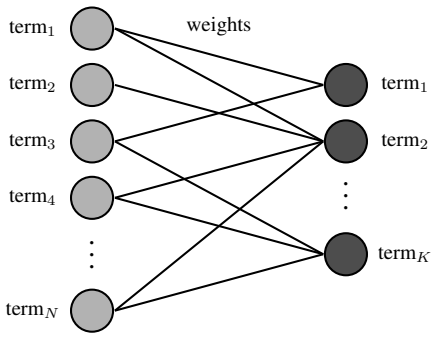


Fig. 1. A schema of the inverted index utilized by ESA.

III. The Adaptive Learning Algorithm

During our research on the automatic tagging methods we noticed that the structure of the inverted index used by ESA resembles a structure of an artificial neural network [13]. This network consists of a single layer of perceptrons (neurons)

that correspond to the different concepts and their inputs correspond to the distinct terms from the concept definitions. However, unlike in a classical neural network, in the case of ESA the inputs are not connected with every neuron. In fact, the net of the connections is rather sparse, since only a small fraction of terms appears in a single description of a concept. Each connection has a non-negative weight from the inverted index, which quantifies the association between the term and the concept. Such a schema of the inverted index structure is depicted in Figure 1.

In the classical neural networks, the activation of each neuron is determined by computing a value of, so called, an activation function. This function takes as its argument a weighted sum of input values and returns a real-valued output. In ESA the activation function corresponds to the identity function. Nevertheless, it is possible to use other types of functions, including those which are typically used in perceptrons (e.g. sigmoid, hyperbolic tangent [13]), in order to scale the association values into a desired range. We can also easily modify the network structure from Figure 1 by adding an additional input, connected to each of the neurons. This input will be treated as an activation threshold. We will assign a concept C_k to a document D_i only if its association $v_{k,i}$ exceeds the corresponding activation threshold a_k, $k = 1, \ldots, K$.

Since the model resembles a neural network, in our previous research [8] we proposed to use a simple learning algorithm for the adaptation of weights from the inverted index to a feedback regarding the tagging quality, obtained from domain experts. The algorithm was based on a typical perceptron learning schema, namely the error backpropagation approach [14]. It is shown in Figure 2. Here we present an improved version of this algorithm that does not require a prior information regarding a number of concepts that should be assigned to each document. For this purpose, we first need to discuss the types of errors that can be made in predicting a set of labels that should be assigned to a given document.

Let us denote by $esa(D_i, INV, A)$ a set of concepts assigned by ESA to a document D_i, using the inverted index INV and the vector of activation thresholds A. This set consists of those concepts whose associations to D_i exceeded the activation threshold values, i.e., $esa(D_i, INV, A) = \{C_k : v_{k,i} > a_k, k = 1, \ldots, K\}$. We assume that there is available a corpus D of training documents, for which we can get the sets of truly related concepts. Since those sets of reference labels usually have to be obtained from domain experts, we will denote them by $exp(D_i)$.

Knowing the sets $esa(D_i, INV, A)$ and $exp(D_i)$ we can divide their union into three mutually disjoint subsets: $TP_i = esa(D_i, INV, A) \cap exp(D_i)$, $FP_i = esa(D_i, INV, A) \setminus exp(D_i)$ and $FN_i = exp(D_i) \setminus esa(D_i, INV, A)$. They can be interpreted as the sets of *Truly Positive*, *Falsely Positive* and *Falsely Negative* cases in the classical machine learning theory [13]. The set TP_i contains truly relevant concepts which were assigned by ESA. Since we want to maximized its cardinality, in every iteration of the learning algorithm we will increase the weights of the connections between the terms t_j from D_i and the concepts from TP_i. The update will be proportional to the association strength between the terms and D_i, which is quantified by the values of $w_{j,i}$. Analogically, we will increase the weights of the concepts from FN_i and decrease those of the concepts from FP_i. At the same time we will be updating the activation thresholds in order to move the concepts from FN_i into TP_i and to remove the FP_i concepts from the set $esa(D_i, INV, A)$. Details of this procedure are explained by Algorithm 1. We call it Adaptive Semantic Analysis (ASA) by an analogy to the ESA algorithm.

We impose one constraint on the weight refinement procedure. Only the available concept descriptions determine the network structure of the inverted index. During the learning procedure we do not construct any new connections in the network, i.e. we restrict the weights $inv_{j,k}$ equal zero to remain zero for a whole learning process. Moreover, the updates in our algorithm are multiplicative, which guarantees that $inv_{j,k} \geq 0$ for every j and k. This restriction is motivated by an intuition that the original concept descriptions, which are usually provided by domain experts, contain sufficient vocabulary to characterize the concepts, thus they define a good model of the terms-to-concepts relations. Additionally, by tuning a large number of weights it is possible to fall into a trap of over-fitting the inverted index to the reference data. Moreover, the reduced number of connections in the inverted index makes the learning more efficient, since there are needed considerably less updates at every iteration of the ASA algorithm.

In the algorithm, the activation thresholds are tuned along the concept weights. In practice, however, they do not need to be updated in every iteration. In order to speed up the learning process, the line number **(31)** of Algorithm 1 can be executed periodically, with the length of the period controlled by an additional parameter.

IV. Experiments

We tested our multi-label tagging system on three different problems, namely automatic labeling of biomedical articles from the PubMed Central repository with headings from the MeSH ontology, assigning categories from ACM Computing Classification System (ACM CCS) to articles from ACM Digital Library and labeling research papers from the Infona repository [6] with the OECD Fields of Science and Technology Classification (OECD FOS) categories. In all those experiments we followed the same testing methodology. We split the available corpus into a training and a test set. We use the training data for adaptive learning of the associations between terms and concepts with the proposed ASA algorithm (see Section III), and then we verify the performance of our tagging system on the test data. We repeat the whole procedure several times with different divisions of the data and report the average results. As the quality measures we use average values of the F_1-$score$, $Precision$ and $Recall$, obtained for all the test documents by comparing the predicted tags to those which were assigned by experts or authors. This type of evaluation of a tagging quality is popular for the multi-label classification problems [15].

A. Experiment on Biomedical Articles

In our first series of experiments we performed the tests on a corpus from the PubMed Central repository [16], consisting of roughly $38,000$ publicly available articles. As the external knowledge base we used the MeSH ontology [11],

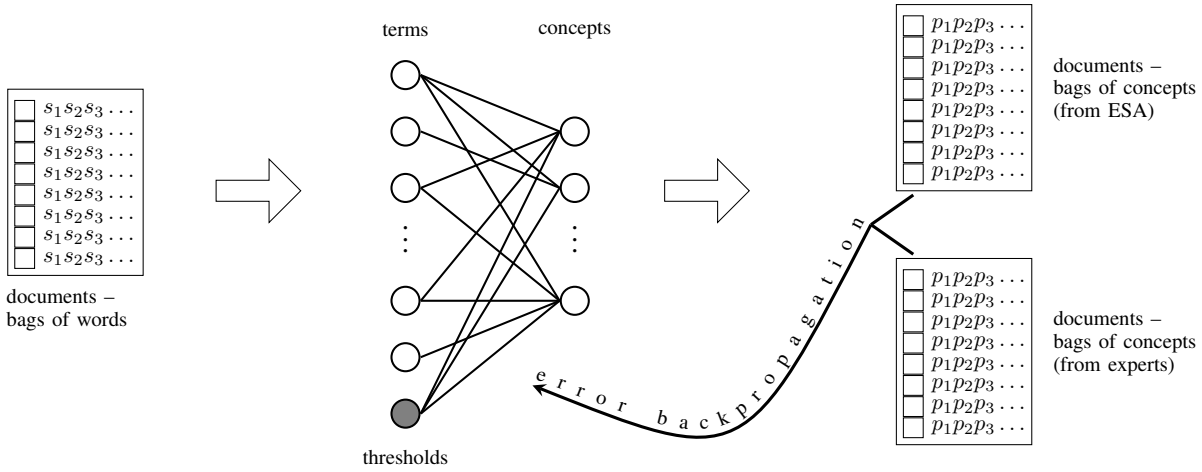


Fig. 2. The learning schema for Adaptive Semantic Analysis.

which is also employed by PubMed to index articles and to facilitate search through its resources. We adapted the ESA method to enable tagging documents from the corpus with the MeSH concepts (also known in MeSH terminology as the headings). In the MeSH ontology, each heading is accompanied by a descriptor data record prepared by domain experts. We composed the final natural language description of the MeSH headings by concatenating the following elements of the corresponding MeSH record: *MESH HEADING* (the name), *MESH SCOPE NOTE* (a short textual description), *ENTRY* (synonyms), *PREVIOUS INDEXING* (previous naming) and *PHARMACOLOGICAL ACTION* (known pharmacological activity). We processed those descriptions using text mining tools in order to determine the initial inverted index structure of our model (i.e. relations between the terms and concepts) and the initial values of the weights. In the experiments we used the edition of MeSH from the year 2012, which contained records on 26.142 main concepts (the headings).

Additionally, for each document in the corpus we obtained sets of MeSH headings assigned by experts from the U.S. National Library of Medicine. The average number of tags assigned to a single article by the NLM experts was ≈ 13.5. We treat those tags as a reference and we utilize them for improving the terms-to-concepts associations with the adaptive learning algorithm described in Section III. We also used those tags for the evaluation purpose. In the tests, we run the adaptation of the inverted index on randomly selected $20,000$ documents and then we use it to tag the remaining part of the corpus. As the starting activation thresholds we use a vector with all coordinates equal 5. This value was chosen using a common sense, based on an observation of a distribution of the concept associations for several exemplary documents. We assess the quality of the tagging by computing the average values of F_1-$score$, $Precision$ and $Recall$ measures. Results of those tests are shown in Figure 3.

The results of the tests turned out to be very promising. On the test data we observed a significant improvement of performance over the regular ESA (the iteration number 0 in the plots) in terms of the computed statistics. For instance, F_1-$score$ value improved by approximately 160% (from ≈ 0.15 to ≈ 0.39). Even a greater improvement was noticed with regards to the values of $Recall$. After the last iteration, its average value exceeded 0.43 while for the regular ESA the average $Recall$ was ≈ 0.16. This however, can be partially explained by the fact that in the initial learning iterations the resulting tagging model usually returned a lower number of labels than the experts.

B. Experiment on Papers from ACM Digital Library

This experiment was conducted on a corpus consisting of publicly available meta-data entries for articles from ACM Digital Library. The corpus contain information on approximately $400,000$ research papers from the field of computer science. The available meta-data included a title, an abstract and in some cases a list of key phrases assigned by authors. We concatenated those information for each document into a single text and we used it to compute its bag-of-words representation. Additionally, the data contained a list of associated ACM CCS categories which also were inputted by the authors. On average, every article was associated with only three out of 1571 possible categories.

The task in this experiment was to label the articles with the ACM CCS categories based on the remaining meta-data. Unlike in the previous experiment, however, this time we did not possess any additional knowledge base with natural language descriptions of the concepts. To deal with this problem we had to slightly modify the procedure of our experiment. After the initial division of the data into the training and test sets (in proportion of 1:1), we divided the training data into two separate sets. For each of the ACM CCS categories we concatenated into a single text the meta-data of all articles that were labeled with this category by the authors. In this way we obtained the textual representation of the categories that could be used for the computation of the initial term-to-concepts associations for our tagging system.

In the second step, we used those associations as a starting point for the ASA algorithm. We performed the adaptive learning of the associations on the remaining part of the training data. We initiated the learning process with the activation thresholds set to 0.30 for all the categories. The starting value of this parameter was much lower than in the experiments

Algorithm 1: Computation of a new inverted index matrix INV^{l+1} and activation thresholds A^{l+1} in the l-th iteration of the adaptive learning algorithm (ASA).

Input: A corpus $D = \{D_i : i \in 1, \ldots, M\}$; INV^l; activation thresholds $A^l = \langle a_1, \ldots, a_K \rangle$;
Output: An updated matrix INV^{l+1}; a vector A^{l+1};

begin
  Initiate ΔINV and CU as empty $N \times K$ matrices;
  Initiate ΔA as a zero vector of length K;
  for $i = 1$ **to** M **do**
    $TP_i = esa(D_i, INV^l, A^l) \cap exp(D_i)$;
    $FP_i = esa(D_i, INV^l, A^l) \setminus exp(D_i)$;
    $FN_i = exp(D_i) \setminus esa(D_i, INV^l, A^l)$;
    foreach $C_k \in esa(D_i, INV^l, A^l) \cup exp(D_i)$ **do**
      $tIds = \{j : t_j \in D_i \wedge INV^l[j,k] > 0\}$;
      $wSum = \sum_{j \in tIds} w_{j,i}$;
      if $C_k \in FP_i$ **then**
        foreach $j \in tIds$ **do**
          $\Delta INV[j,k] = \Delta INV[j,k] - INV^l[j,k] * w_{j,i} / wSum$;
          $CU[j,k] = CU[j,k] + 1$;
        $\Delta A[k] = \Delta A[k] + A^l[k] * \left(1 - \frac{|TP_i|}{|TP_i \cup FP_i|}\right)$;
      else
        foreach $j \in tIds$ **do**
          $\Delta INV[j,k] = \Delta INV[j,k] + INV^l[j,k] * w_{j,i} / wSum$;
          $CU[j,k] = CU[j,k] + 1$;
        $\Delta A[k] = \Delta A[k] - A^l[k] * \left(1 - \frac{|TP_i|}{|TP_i \cup FN_i|}\right)$;
  foreach (j,k) *such that* $CU[j,k] > 0$ **do**
    $\Delta INV[j,k] = \Delta INV[j,k] / CU[j,k]$;
  $INV^{l+1} = INV^l + \Delta INV$;
  $A^{l+1} = A^l + \Delta A / M$;
  return INV^{l+1} *and* A^{l+1}

on biomedical articles due to a fact that this time we had to operate on significantly shorter texts. We measured the quality of our tagging system by comparing the labels assigned to the test articles with the labels which were given by the authors. The results of those comparisons in the consecutive iterations of the learning algorithm are depicted on Figure 4.

Similarly to the previous experiments, the results clearly show that the learning algorithm significantly improves the quality of the tagging system in comparison to the standard ESA (the iteration number zero on the plots). In the test, the highest F_1-$score$ on the test set was ≈ 0.20. It was obtained in the last iteration (50-th) of the algorithm, which suggests that it would be possible to slightly improve the results by giving the algorithm some more time for learning. There is also a noticeable difference between the results on the training and test sets. The highest training F_1-$score$ exceeded 0.33, which is over 50% higher than the corresponding test score. On one hand this difference may be partially explained by the fact that authors do not follow any strict rules or guidelines when they assign the categories to their papers. It makes the assigned labels very subjective. As a consequence, the prediction of the ACM CCS categories becomes a very difficult task. On the other hand, the differences in the tagging quality for the training and test data may be caused by the way we generated the textual descriptions of the ACM CCS categories. The concatenation of many article abstracts had to result in the inclusion of many highly specialized terms into the descriptions. Such terms often allow to identify individual papers, thus their presence may lead to the over-fitting of the learning algorithm to the training data.

C. Experiment on Data from the Infona System

The last series of experiments was conducted on a corpus obtained from the Infona repository [6]. Infona contains meta-data of over 1.8 million articles from a wide range of science fields. However, in our experiments we were restricted to only a small sample of all the data from this repository, i.e. our corpus contained information from 1000 meta-data entries. Each entry consisted of an article title in English, author names and an English abstract. Additionally, for many entries there were available key phrases assigned by the authors. Similarly as in the previous experiments, for each article we concatenated the available meta-data (we skipped the information regarding the authors) to create their textual representation.

The task in those experiments was to learn how to tag the documents from Infona with the categories from the OECD FOS classification. This classification system consists of 42 main categories grouped into six different upper-level categories, namely *Natural sciences*, *Engineering and technology*, *Medical and Health sciences*, *Agricultural sciences*, *Social sciences* and *Humanities* [17]. In order to construct

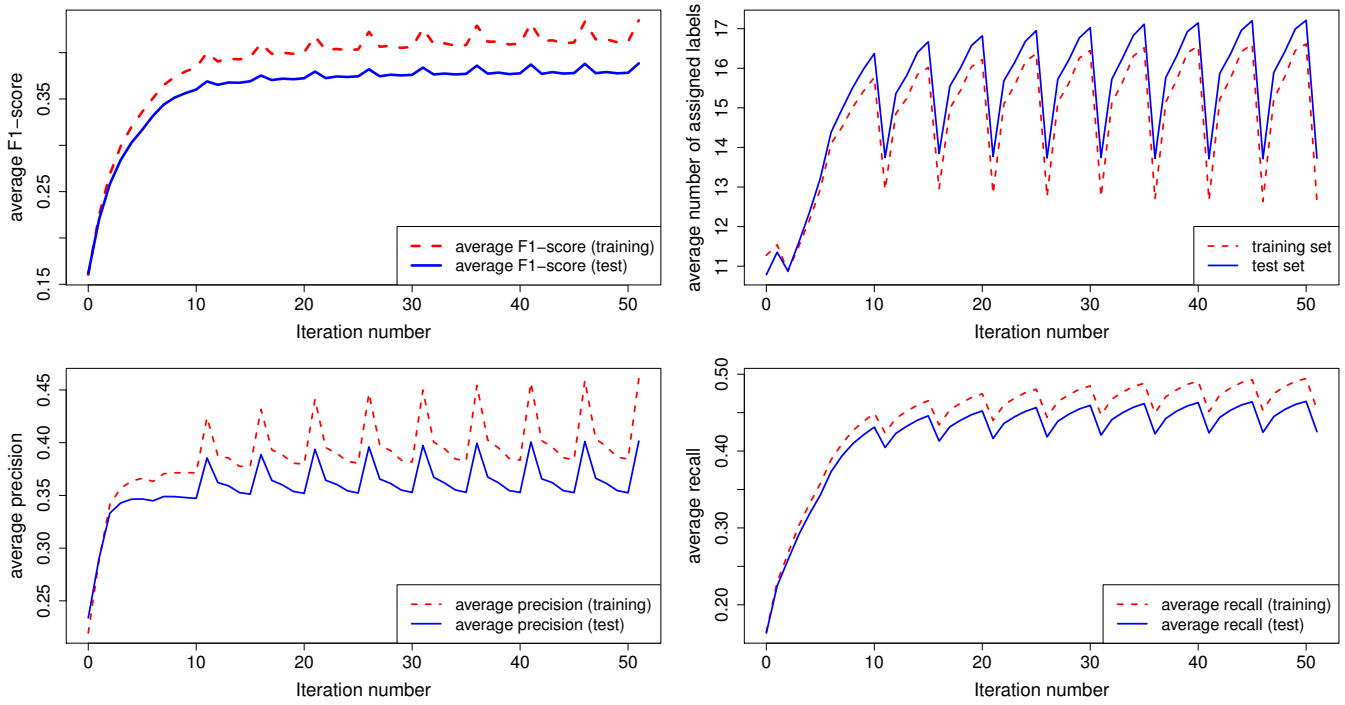


Fig. 3. Results of the adaptive learning algorithm on the PubMed Central corpus (50 iterations). The documents were labeled with the concepts from the MeSH ontology. For each learning iteration the average F_1-$score$, $Precision$, $Recall$ and a number of assigned categories is shown. The results in the iteration number 0 correspond to the standard ESA algorithm. The activation thresholds in those experiments were updated in every fifth iteration of the algorithm.

the textual descriptions for all the 42 categories we manually selected a number of related English Wikipedia articles and we concatenated their content.

The size of the corpus for those experiments was limited by the availability of the expert knowledge. Infona did not provide us any information about the OECD FOS categories of the documents. In order to create a reference set of labeled documents we had to ask volunteers to manually tag the data. In this way we obtained a total of 1000 labeled meta-data entries which we could use in the experiments. A single document on average was assigned to 1.7 categories. Approximately 30% of the documents were labeled by more than one person. In this way we could check how difficult this task is and get a good estimation of a reference quality assessment. It turned our that the average cross-expert F_1-$score$ merely exceeds 0.51, while the average $Precision$ and $Recall$ values are about 0.55. It means that on average, two different experts agree only on about a half of the assigned categories, thus we should not expect a better result from an automatic tagging method.

In the experiments we used 800 documents as a training set and the remaining 200 served as a test set. Due to the small size of the test sets, we repeated the testing procedure 20 times on different divisions of the data in order to get reliable estimations of the tagging quality. Similarly as in the case of the ACM Digital Library corpus, we set the initial values of the activation thresholds to a low value, i.e. they were all equal 0.25. The average results of those tests for the consecutive iterations of the ASA algorithm are presented in Figure 5. In those plots, the values of the cross-expert quality estimations are marked by the thick black lines.

The experimental results once again clearly demonstrate usefulness of our learning algorithm. The average F_1-$score$ value obtained using the adapted inverted index was greater by over 40% than the score of the standard ESA algorithm. For ASA it was approximately 0.47. It is worth noting that this improvement was possible, even though the number of available training documents was very limited. The best F_1-$score$ on the test set was usually achieved around thirtieth iteration of the algorithm and after that point we noted a slight decrease in the results. Since the scores obtained on the training set systematically grew and often exceeded 0.8, this can be most likely explained by the over-fitting problem. Nevertheless, the score achieved using ASA was very close to the cross-expert F_1-$score$ which confirms the effectiveness of the proposed adaptive learning algorithm.

V. Conclusions

In the paper we discussed an adaptive learning framework, called Adaptive Semantic Analysis, which can be utilized for improving the terms-to-concepts associations from the inverted index of the ESA algorithm. We described in details the learning procedure and we showed its effectiveness in dealing with real-life problems. In particular, we presented results of experiments on three document corpora, in which the ASA algorithm was used to facilitate the automatic tagging of documents with concepts from different knowledge bases.

We hope that in a future our automatic tagging module can become a part of a larger scientific article repository platform. We are currently trying to integrate our SONCA platform with the Infona repository. This may enable an

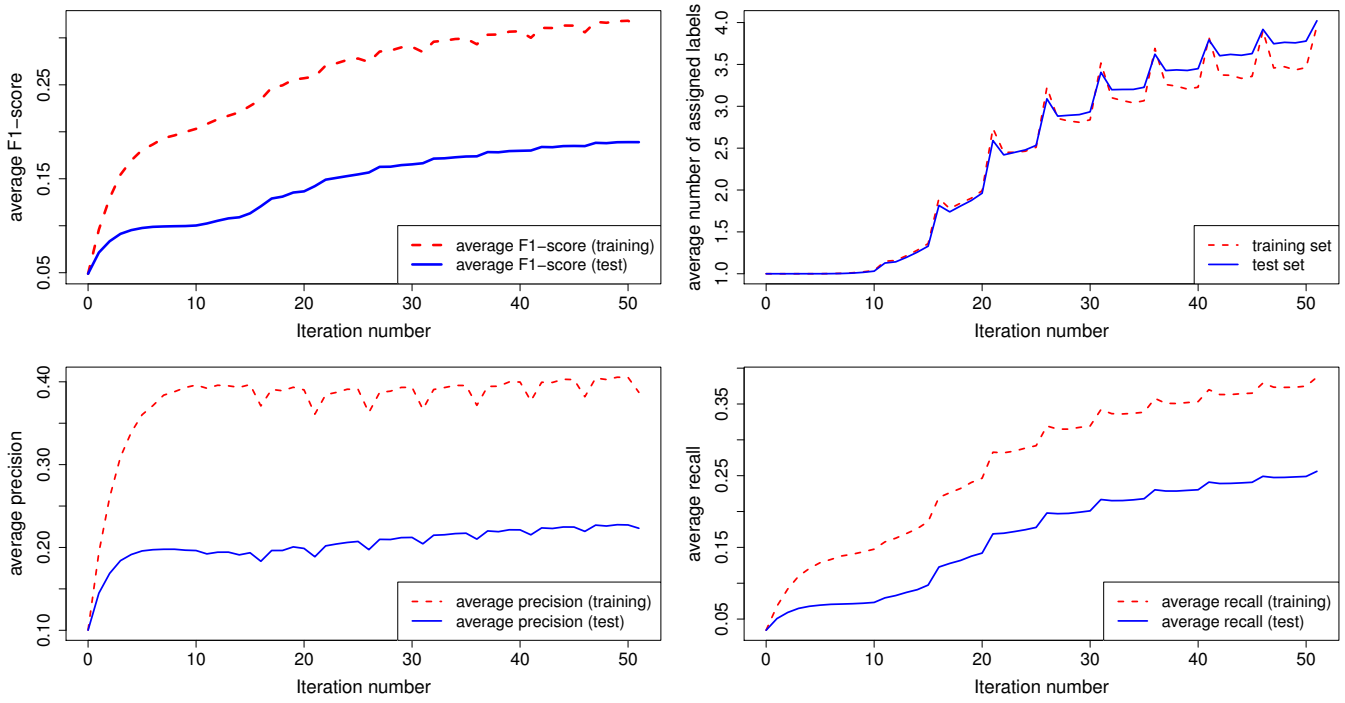


Fig. 4. Results of the adaptive learning algorithm on the ACM Digital Library corpus (50 iterations). The documents were labeled with the ACM Computing Classification System categories. For each learning iteration the average F_1-$score$, $Precision$, $Recall$ and a number of assigned categories is shown. The results in the iteration number 0 correspond to the standard ESA algorithm. The activation thresholds in those experiments were updated in every fifth iteration of the algorithm.

efficient and automatic semantic indexing of Infona's resources which would allow to better fulfill the information needs of Infona's users. Apart from the direct use as an indexing module of the search engine, the tags returned by our system could be utilized for, e.g., improving the clustering of search results or assigning comprehensible names to document clusters.

REFERENCES

[1] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[2] B. Fazzinga, G. Gianforme, G. Gottlob, and T. Lukasiewicz, "Semantic web search based on ontological conjunctive queries," *Web Semantics: Science, Services and Agents on the World Wide Web*, 2011.

[3] A. Hliaoutakis, G. Varelas, E. Voutsakis, E. G. M. Petrakis, and E. Milios, "Information retrieval by semantic similarity," *Int. Journal on Semantic Web and Information Systems (IJSWIS). Special Issue of Multimedia Semantics*, vol. 3, no. 3, pp. 55–73, 2006.

[4] D. Ślęzak, A. Janusz, W. Świeboda, H. S. Nguyen, J. G. Bazan, and A. Skowron, "Semantic analytics of PubMed content," in *Information Quality in e-Health - 7th Conference of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2011, Graz, Austria, November 25-26, 2011. Proceedings*, ser. LNCS, A. Holzinger and K.-M. Simonic, Eds., vol. 7058. Springer, 2011, pp. 63–74.

[5] A. M. Rinaldi, "An ontology-driven approach for semantic information retrieval on the web," *ACM Trans. Internet Technol.*, vol. 9, pp. 1–24, 2009. [Online]. Available: http://doi.acm.org/10.1145/1552291.1552293

[6] R. Bembenik, L. Skonieczny, H. Rybinski, M. Kryszkiewicz, and M. Niezgodka, Eds., *Intelligent Tools for Building a Scientific Information Platform - Advanced Architectures and Solutions*, ser. Studies in Computational Intelligence. Springer, 2013, vol. 467.

[7] L. A. Nguyen and H. S. Nguyen, "On designing the sonca system," in *Intelligent Tools for Building a Scientific Information Platform*, R. Bembenik, L. Skonieczny, H. Rybiński, and M. Niezgódka, Eds. Springer-Verlag New York, 2012, pp. 9–36.

[8] A. Janusz, W. Świeboda, A. Krasuski, and H. S. Nguyen, "Interactive document indexing method based on Explicit Semantic Analysis," in *Proceedings of the 8th International Conference on Rough Sets and Current Trends in Computing (RSCTC 2012), Chengdu, China, August 17-20, 2012*, ser. LNAI, J.T. Yao et al., Ed., vol. 7413. Springer, Heidelberg, 2012, pp. 156–165.

[9] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *Proc. of The 20th Int. Joint Conf. on Artificial Intelligence*, Hyderabad, India, 2007, pp. 1606–1611. [Online]. Available: http://www.cs.technion.ac.il/~shaulm/papers/pdf/Gabrilovich-Markovitch-ijcai2007.pdf

[10] O. Egozi, S. Markovitch, and E. Gabrilovich, "Concept-based information retrieval using explicit semantic analysis," *ACM Trans. Inf. Syst.*, vol. 29, no. 2, pp. 8:1–8:34, Apr. 2011. [Online]. Available: http://doi.acm.org/10.1145/1961209.1961211

[11] United States National Library of Medicine, "Introduction to MeSH - 2011," Online: http://www.nlm.nih.gov/mesh/introduction.html, 2011. [Online]. Available: http://www.nlm.nih.gov/mesh/introduction.html

[12] Association for Computing Machinery, "The 2012 acm computing classification system," Online: http://www.acm.org/about/class/2012, 2012. [Online]. Available: http://www.acm.org/about/class/2012

[13] T. M. Mitchell, *Machine Learning*, ser. McGraw Hill series in computer science. McGraw-Hill, 1997.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[15] A. Janusz, H. S. Nguyen, D. Ślęzak, S. Stawicki, and A. Krasuski, "JRS'2012 Data Mining Competition: Topical Classification of Biomedical Research Papers," in *Proceedings of the 8th International Conference on Rough Sets and Current Trends in Computing (RSCTC 2012), Chengdu, China, August 17-20, 2012*, ser. LNAI, J.T. Yao et al., Ed., vol. 7413. Springer, Heidelberg, 2012, pp. 417–426.

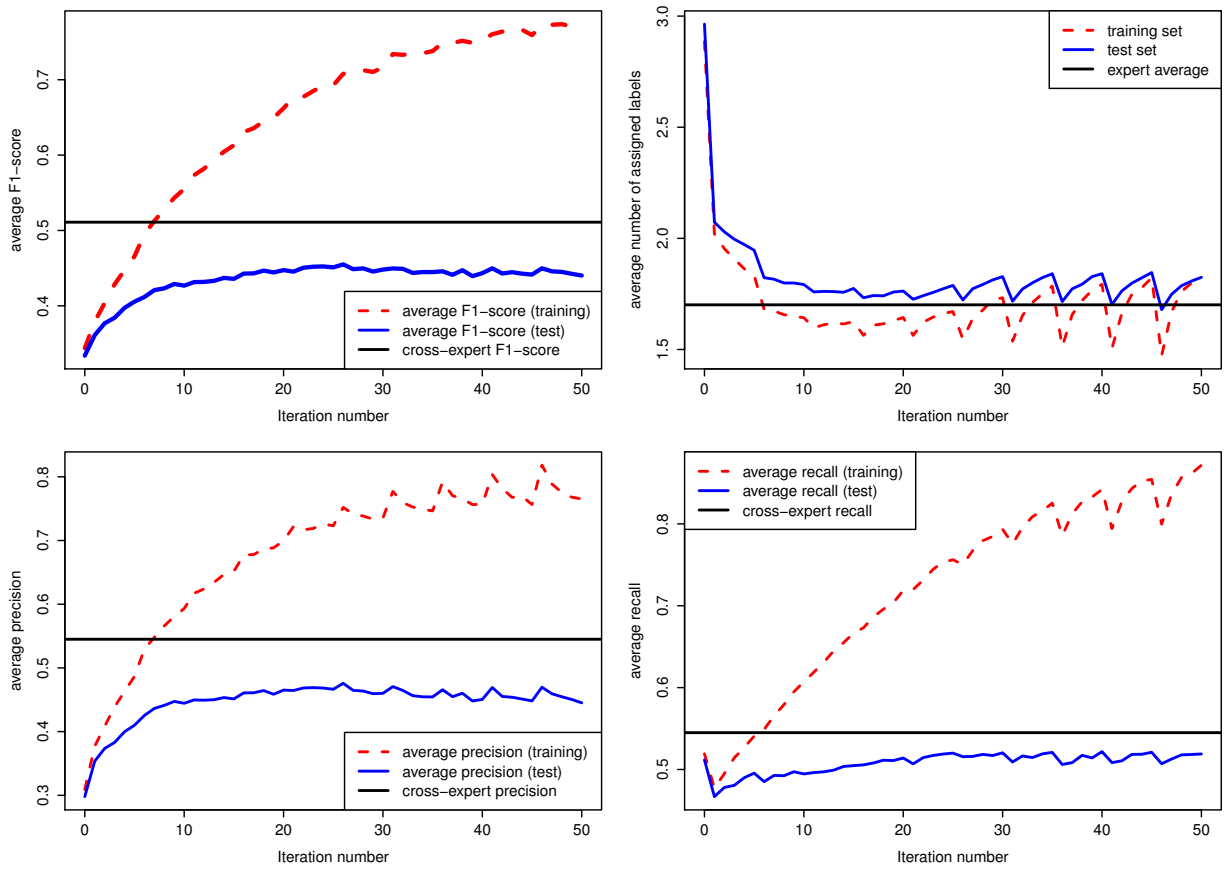


Fig. 5. Results of the adaptive learning algorithm on the corpus from the Infona repository (50 iterations). The documents were labeled with the OECD Fields of Science and Technology Classification categories. For each learning iteration the average F_1-$score$, $Precision$, $Recall$ and a number of assigned categories is shown. Moreover, the thick black line in the plots denotes the cross-expert statistic values which can be regarded as an additional reference. The results in the iteration number 0 correspond to the standard ESA algorithm. The activation thresholds in those experiments were updated in every fifth iteration of the algorithm.

[16] R. J. Roberts, "PubMed Central: The GenBank of the published literature," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 2, pp. 381–382, 2001. [Online]. Available: http://www.pnas.org/content/98/2/381.abstract

[17] "Revised Field of Science and Technology (FoS) Classification in the Frascati Manual," Committee for Scientific and Technological Policy, Directorate for Science, Technology and Industry, OECD, Feb. 2007.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 35–42
DOI: 10.15439/2014F231
ACSIS, Vol. 2

A Brain Emotional Learning-based Prediction Model for the Prediction of Geomagnetic Storms

Mahboobeh Parsapoor[1, 2], Urban Bilstrup[1], Bertil Svensson[1]
[1] School of Information Science, Computer and Electrical Engineering (IDE),Halmstad University, Halmstad, Sweden
[2]School of Computer Science, Faculty of Engineering & Physical Science, The University of Manchester, Manchester, UK
[1]{Mahboobeh.Parsapoor, Urban.Bilstrup, Bertil.Svensson }@hh.se
[2]mahboobeh.parsapoor@postgrad.manchester.ac.uk

***Abstract*—This paper introduces a new type of brain emotional learning inspired models (BELIMs). The suggested model is utilized as a suitable model for predicting geomagnetic storms. The model is known as BELPM which is an acronym for Brain Emotional Learning-based Prediction Model. The structure of the suggested model consists of four main parts and mimics the corresponding regions of the neural structure underlying fear conditioning. The functions of these parts are implemented by assigning adaptive networks to the different parts. The learning algorithm of BELPM is based on the steepest descent (SD) and the least square estimator (LSE). In this paper, BELPM is employed to predict geomagnetic storms using the Disturbance Storm Time (Dst) index. To evaluate the performance of BELPM, the obtained results have been compared with the results of the adaptive neuro-fuzzy inference system (ANFIS).**

I. Introduction

THE geomagnetic storm, which originates from the solar wind, disturbs the Earth's magnetosphere and has caused harmful damage to the ground based communication, electricity power network, etc. Therefore, developing alert systems for geomagnetic storms is essential in order to prevent these harmful effects [1]-[6].

The disturbance storm time, Dst, is one of the main indices of a geomagnetic storm and was defined by Bruce Tsurutani [1], [5]-[6]. It is a measurement to count 'the number of solar charged particles that enter the Earth's magnetic field' [6]. The Dst index has been proposed to characterize the phases of geomagnetic storms i,e., the initial phase, main phase and recovery phase and has been recorded by several space centers such as the World Data Center for Geomagnetism, Kyoto, Japan.

Different machine learning methods e.g., linear input-output techniques or linear prediction filtering neural networks [8][9], neurofuzzy methods [2],[4], have been investigated for predicting geomagnetic storms using the Dst index [1]-[4], [6]-[12]. Amongst them, neural networks and neuro-fuzzy models have shown high generalization [13],[14] capabilities to model nonlinear behavior of the Dst index .

Recently, inspiration from the mammalian emotional systems to develop emotion-based models has received fairly good attention [15]-[25]. The emotion-based models that were proposed in [15]-[17] have been developed by a limited modification of a computational model of emotional learning that is referred to as 'amygdala-orbitofrontal system'; this computational model simulates the emotional learning in the amygdala (i.e.,one region of the mammalian brain) [26]. The obtained results from [15]-[17] verify that there are not able to accurately predict chaotic behavior of nonlinear systems.

This paper suggests a new instance of Brain Emotional Learning-Inspired Models (BELIMs) that are emotion-based models. The suggested model is applied to predict the Dst index of geomagnetic storms. So far different variations of BELIMs have been [18]-[25] examined for forecasting solar activity and geomagnetic storms.

The main contribution of this paper is to present a new version of BELIMs to be used as an accurate prediction method for the long horizon prediction of the Dst index. Another contribution of this paper is to provide comparative results when predicting the Dst index.

The rest of the paper is organized as follows: Section II reviews related works to emotions. Section III describes the structure, function and learning algorithm of the BELPM (Brain Emotional Learning Based Prediction Model). Section IV reviews the related studies in predicting the Dst index and the results of BELPM to predict the Dst index are described. Finally, conclusions about the performance of BELPM and the further improvements to the suggested model are discussed in Section V.

II. Related Works To 'Emotions'

Neuroscientists and psychologists have tried to describe emotions on the basis of different hypotheses, e.g., psychological, neurobiological, philosophy, and learning hypotheses [27]. Cognitive neuroscientists have also tried to describe the neural system underlying the emotional process. One of the earlier works is the 'Papez circuit' (See Fig. 1) that was proposed by Jamez Papez. As Fig. 1 shows, this circuit includes the 'hypothalamus, anterior thalamus, cingulate gyrus and hippocampus' [27].

978-83-60810-58-3/$25.00

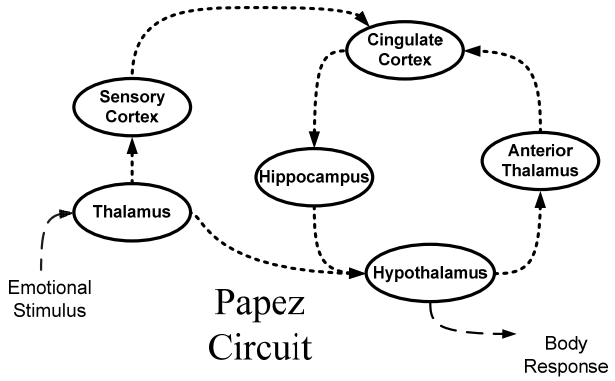


Fig.1. The Papez Circuit and its component.

MacLean modified the Papez circuit and proposed the limbic system theory to describe the regions of the brain that are responsible for emotional processing. The limbic system includes the hippocampus, amygdala, thalamus and sensory cortex [27]. Later, neuroscientists rejected the limbic system theory and stated that different parts of the brain are responsible for different emotional behavior [28]. Fear is a common emotional behavior that exists as well for humans as animals. Fear conditioning has been defined as a 'behavioral paradigm' which means learning fearful stimuli to predict aversive events [28].

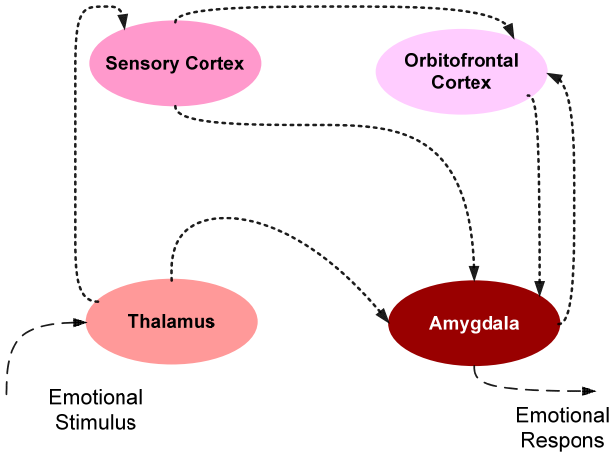


Fig. 2. A schematic of the brain's parts and their interconnections in fear conditioning.

Figure 2 displays how the amygdala and other parts of the brain, thalamus, sensory cortex and orbitofrontal cortex connected to process a fearful stimulus in the mammalian brain. As the diagram indicates, the amygdala is the central part to process the emotional stimulus. The neural structures of emotional behavior have been the foundation of the computational model of emotional learning.

Computational models of emotion [23],[25],[26] are computer-based models that have been developed to simulate different aspects of the emotional system. A good example of computational models is a model that is referred to as amygdala-orbitofrontal system and has been proposed to simulate emotional learning in the amygdala [26].

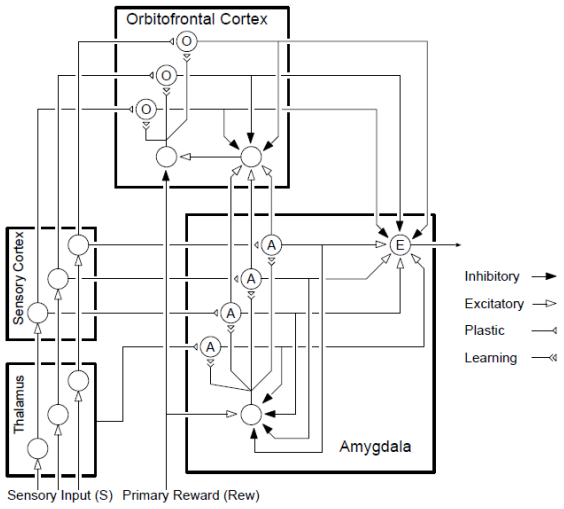


Fig.3. The Amygdala-orbitofrontal system [26].

Figure 3 depicts the internal structure of the amygdala-orbitofrontal system and describes that orbitofrontal cortex and amygdala consists of several nodes; the output of each node of the amygdala is represented as A_i while the output of each node of the orbitofrontal cortex is represented as O_i. The overall output of the model is represented as **E** and it is formulated as equation (1) [26].

$$\mathbf{E} = \sum_i A_i - \sum_i O_i \tag{1}$$

Here A_i and O_i are the output of the i^{th} node of the amygdala and the orbitofrontal cortex, respectively. The updating rules of the model are based on the reinforcement signal **REW**. The updating rules are formalized as equations (2) and (3) and are utilized to adjust the weights **V** and **W** that are associated with the nodes of the amygdala and the orbitofrontal part, respectively [26]. Here s_i is the input stimulus for the i^{th} node of the amygdala and the orbitofrontal cortex [26].

$$\Delta V_i = \alpha(S_i \times \max(0, \mathbf{REW} - \sum_i A_i)) \tag{2}$$

$$\Delta W_i = \beta(S_i \times (\sum_i O_i - \mathbf{REW})) \tag{3}$$

The amygdala-orbitofrontal model [26] has a simple structure and has been used as a foundation for numerous l emotion-based models [15]-[25],[29]-[33]. As was discussed earlier, the emotion-based models in [15]-[16] were proposed as chaotic time series prediction models. The foundation of these models is amygdala-orbitofrontal system; however they were developed by changing the updating rules of amygdala-orbitofrontal system. These models have not shown good results to accurately predict chaotic time series [15]-[16]. In [17], another modification of amygdala-orbitofrontal system was proposed by changing the input vector of the thalamus and the amygdala part; in addition, the updating rules of amygdala and orbitofrontal cortex were modified. The model that is named 'ADBEL'[17]

(adaptive brain emotional decayed learning) was applied to predict the hourly Dst index; however, the obtained results verify the ADBEL could not accurately predict chaotic behavior of the Dst index.

The Emotion-based controllers [29]-[33] have also been developed by imitating the structure of the amygdala-orbitofrontal.

This paper presents a new type of emotion-based prediction models. Although the general structure of this model is similar to the amygdala-orbitofrontal system, the internal structure of this model is different from the amygdala-orbitofrontal system.

III. Brain Emotional Learning-Based prediction Model

The Brain Emotional Learning-Based Prediction Model (BELPM) is a type of Brain Emotional Learning-Inspired Model (BELIM) which is a new category of computational intelligence models. The general structure of a BELIM is an extension of the amygdala-orbitofrontal system by adding new internal parts that have been inspired by the neural structure of one of the emotional systems in the brain. Different types of BELIM: Brain Emotional Learning-based Fuzzy Inference System (BELFIS), Brain Emotional Learning-based Recurrent Fuzzy System (BELRFS) and Emotional Learning Inspired Ensample Classifier (ELiEC) [18]-[25] have been proposed as prediction models and classification models.

A. Structural Aspect of BELPM

Figure 4 depicts the structure of BELPM and shows that it consists of four main parts: TH, CX, AMYG and ORBI which refer to the THalamous, sensory CorteX, AMYGdala, and ORBItofrontal cortex, respectively. The structure of BELPM copies the interconnection of those parts (THalamous, sensory CorteX, AMYGdala, and ORBItofrontal cortex) that are responsible to process the emotional stimuli. It should be noted that these regions of the brain are very complex and there is no intention to mimic their functionality and all their connections in detail.

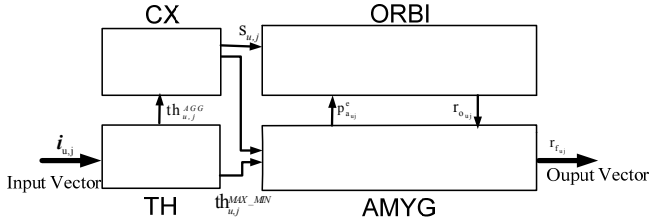


Fig. 4. The structure of BELPM.

The following steps describes the input and output of each part of BELPM; when it receives an input as $i_{u,j}$ from the training data set, $\mathbf{I}_u = \{\mathbf{i}_{u,j}\}_{j=1}^{N_u}$.

1. First, $\mathbf{i}_{u,j}$ the j^{th} input vector from $\mathbf{I}_u = \{\mathbf{i}_{u,j}\}_{j=1}^{N_u}$ (taking the assumption that the number of training samples is equal to N_u ; the subscript u has been used to determine the input data is chosen from the training data set) enters the TH, which provides two outputs, $\mathbf{th}_{u,j}^{Max\_Min}$ and $\mathbf{th}_{u,j}^{AGG}$ which are sent to the AMYG and the CX, respectively.

2. The CX provides $\mathbf{s}_{u,j}$ and sends it to both the AMYG and the ORBI.

3. The AMYG receives two inputs: $\mathbf{th}_{u,j}^{Max\_Min}$ and $\mathbf{s}_{u,j}$. It provides the primary output, $r_{a_{u,j}}$, and expected punishment, $p^e_{a_{u,j}}$, that is sent to the ORBI (the subscript a has been used to show the outputs of AMYG).

4. The ORBI receives $\mathbf{s}_{u,j}$ and $p^e_{a_{u,j}}$. It provides the secondary output, $r_{o_{u,j}}$, and sends it to the AMYG. 5. The AMYG receives $r_{o_{u,j}}$ and provides the final output, $r_{f_{u,j}}$ (the subscript f has been used to show the final outputs).

B. Functional Aspect of BELPM

The function of BELPM is implemented by assigning adaptive networks to different parts. Figure 5 describes how the adaptive networks can be assigned to each part to implement the functionality of that part.

The adaptive network (see Fig. 6) consists of a number of nodes that are connected by directional links. The nodes of the adaptive network can be classified into circle and square nodes. A circle node has a function without adjustable parameters; in contrast, the square nodes have been defined by a function with the adjustable parameters. The learning parameters of an adaptive network are a combination of linear and nonlinear parameters and can be adjusted by using a learning algorithm.

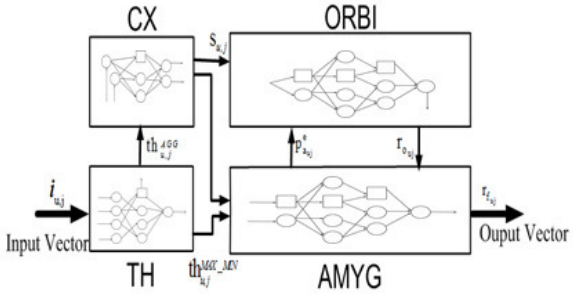


Fig. 5. Assigning different adaptive networks to different parts of BELPM.

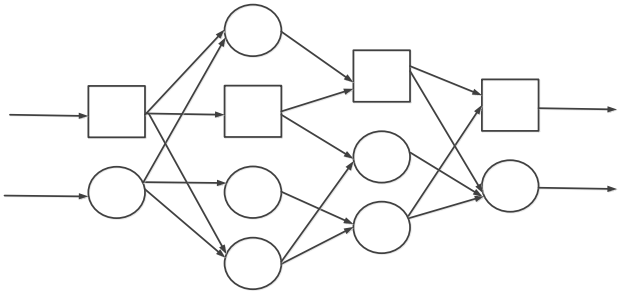


Fig. 6. A structure of an adaptive network.

C. A Weighted k-Nearest Neighbor based Adaptive Network

In BELPM, weighted k-nearest neighbor based adaptive networks are assigned to the AMYG and the ORBI. This adaptive network has been developed on the basis of weighted k-Nearest Neighbor (Wk-NN). Figure 7 describes a simple Wk-NN based adaptive network and shows that it is divided into four layers. In the following, the function, the input and the output of each layer has been explained.

The first layer consists of k square nodes with K(.) function (kernel function). Note that in Fig.7, k is equal to three. This layer has an input vector as $\mathbf{d}_{\min} = \{d_{\min,m}\}_{m=1}^{k}$. The $\mathbf{d}_{\min}$ is a set of k minimum distances of $\mathbf{d} = \{d_j\}_{j=1}^{N_u}$. The distances vector $\mathbf{d} = \{d_j\}_{j=1}^{N_u}$ can be calculated as Euclidean distances between a new input as $\mathbf{i}_{c,j}$ and the training data as $\{\mathbf{i}_{u,j}\}_{j=1}^{N_u}$. The output vector of the m^{th} node of the first layer is calculated using equation (4). Here, the input to the m^{th} node is $d_{\min,m}$.

$$n_m^1 = K(d_{\min,m}) \tag{4}$$

In general, the kernel function for the m^{th} node can be one of the functions that have been defined as equations (5), (6), and (7). The input and the parameter of K(.) of m^{th} node can be determined using d_m and b_m .

$$K(d_m) = \exp(-d_m b_m) \tag{5}$$

$$K(d_m) = \frac{1}{(1+(d_m b_m)^2)} \tag{6}$$

$$K(d_m) = \frac{\max(\mathbf{d}) - (d_m - \min(\mathbf{d}))}{\max(\mathbf{d})} \tag{7}$$

The second layer is a normalization layer and has k nodes (fixed or circle), which are labeled N to calculate the normalized values of $\mathbf{n}^1$ using (8). The input vector of this layer is $\mathbf{n}^1$ and the output of m^{th} node in this layer can be calculated as (8)

$$\bar{n}_m^1 = \frac{(n_m^1)}{\sum_{m=1}^{k} n_m^1} \tag{8}$$

The third layer has k circle nodes; the function of m^{th} node of this layer is given in (9). This layer has two input vectors, $\bar{\mathbf{n}}^1$ and $\mathbf{r}_{\min,u}$; the latter is a vector that is extracted from $\mathbf{r}_u = \left[r_{u,1}, r_{u,2}, ..., r_{u,N_u}\right]$ and is related to the target outputs of the k samples of $\{\mathbf{i}_{u,j}\}_{j=1}^{N_u}$ that have minimum distances with the new input $\mathbf{i}_{c,j}$. The output of this layer is $\mathbf{n}^3$.

$$n_m^3 = \frac{(n_m^1)}{\sum_{m=1}^{k} n_m^1} \times r_{u,m} \tag{9}$$

The fourth layer has a single node (circle) that performs the summation of the input vector, $\mathbf{n}^3$ to produce $\mathbf{r}$.

The above explanation has illustrated the function of a simple Wk-NN based adaptive network. It should be noted that in BELPM, the AMYG and the ORBI are assigned this type of adaptive network.

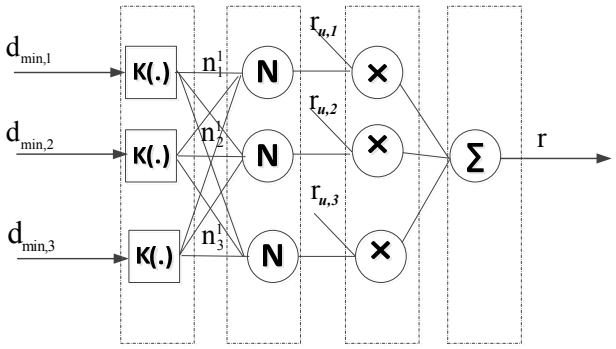


Fig.7. A weighted k-Nearest Neighbor adaptive network.

D. Learning Aspect of BELPM

To adjust the linear and nonlinear learning parameters, a hybrid learning algorithm that consists of the steepest descent (SD) and the least-squares estimator (LSE) is used. The SD updates the nonlinear parameters in a gradient related direction to minimize the loss functions, which are defined based on the outputs of the adaptive networks. The LSE is applied to update the linear parameters. The learning algorithm has been explained in detail in [25].

IV. Related work to Predict Geomagnetic Storms

As was previously discussed, developing an alert system for geomagnetic storms is essential. The Dst index that has been defined to measure the intensity of geomagnetic storms has been utilized by many data driven models. The following subsection reviews several studies that have been used to predict geomagnetic storms.

A. Related works to Predict Geomagnetic Storm

A good review of earlier studies related to use the Dst index to predict geomagnetic storms have been done in [6]. The authors of [6] provided a survey of using the Dst index to predict geomagnetic storms; they also proposed a neural network-based prediction model to predict the minimum

values of the Dst index. The model was successfully evaluated to predict geomagnetic storms of 1980 and 1989. In [9], a recurrent neural network was introduced to predict one hour step of Dst from 2001. The authors of [9] also showed that combining principal component analysis (PCA) and NN could significantly increase the accuracy of prediction of a geomagnetic storm. The damage and harmful effects of geomagnetic storms were reviewed in [10]. The authors of [10] studied the effect of the embedding dimension on the chaotic behavior of the Dst time series; the proposed model in [10] was is study tested for two super storms: 13 March 1989 and 11 January 1997. In [2] a combination of Singular Spectrum Analysis (SSA) and locally linear neuro-fuzzy model was proposed as a useful methodology to increase the accuracy of long term prediction of Dst time series. Specifically, this method was examined to predict ten steps ahead of extracted Dst time series between 1988 and 1990, Within this time window the geomagnetic storm damaged Quebec's power grid and caused a blackout in Quebec [10] is included. A nice review of the Dst index prediction models and the benefits of prediction of the Dst index was presented in [12]. The authors of [12] also proposed a long term prediction model that is known as 'Anemomils'[12] and tested it for three geomagnetic storms of 2001, 2005 and 2012.

B. Evaluating the BELPM's Performance on the Dst index

This subsection evaluates BELPM's performance by examining it on two sets of the Dst index. The code of BELPM has been written in MATLAB and the Dst index can be downloaded from the World Data Center for Geomagnetism, Kyoto, Japan.

The first data set is related to the Dst index of the super storm that occurred in March 1989. The value of the Dst index on the 13th March of 1989 reached to '-589 nanotesla (nT)'. Figure 8 depicts the hourly DST index during January1988 to January 1990.

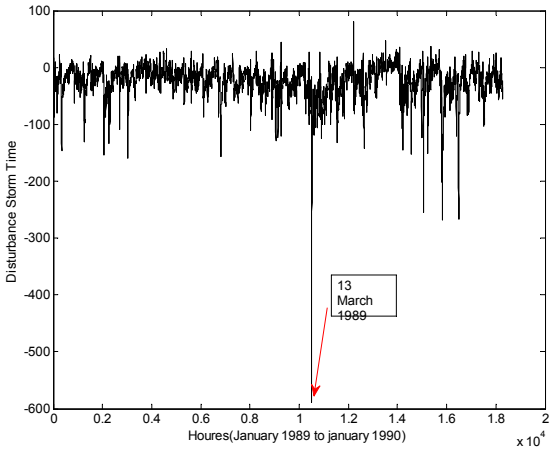


Fig. 8. The hourly Dst index from January 1988 to January 1990.

The second data set is related to the geomagnetic super storm which occurred on July 15th to 17th ; the minimum of the Dst index was – 301 nanotesla (nT). Figure 9 depicts the hourly DST index during the days of July 2000.

To provide a careful comparison with other methods, this paper utilizes two error measures: normalized mean square error (NMSE) and the correlation coefficient $\rho_{y,\hat{y}}$ that are given as equations (10) and (11).

$$\mathrm{NMSE} = \frac{\sum_{j=1}^{N}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{N}(y_j - \bar{y}_j)^2} \tag{10}$$

$$\rho_{y,\hat{y}} = \frac{\mathrm{Cov}(y,\hat{y})}{\sigma_y \sigma_{\hat{y}}} \tag{11}$$

The parameters $\hat{\mathbf{y}}$ and $\mathbf{y}$ refer to the predicted values and desired targets, respectively. The parameter $\bar{\mathbf{y}}$ is the average of the desired targets.

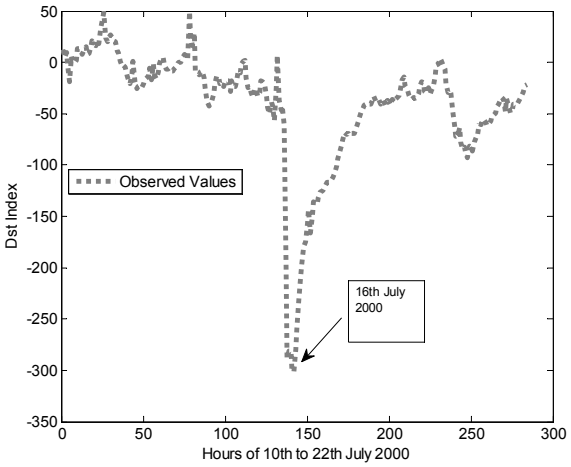


Fig.9. The hourly Dst index for July 2000.

As was mentioned earlier, in the first experiment, BELPM is tested to model the Dst time series between january1988 to January 1990. The Dst time series is related to one of the harmful geomagnetic storms which occurred during solar cycle 22 and caused severe damage to Quebec's electricity power system. In this case, the embedded dimension is selected as three. Figure 10 shows the inputs and output of BELPM as a black box to predict the Dst time series.

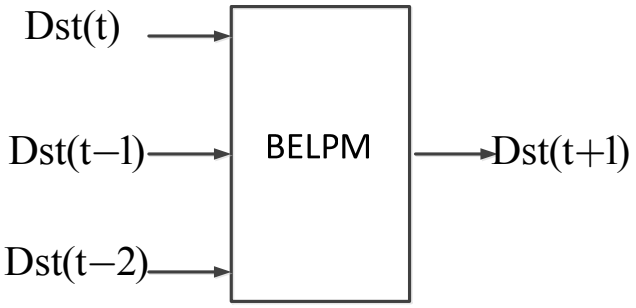
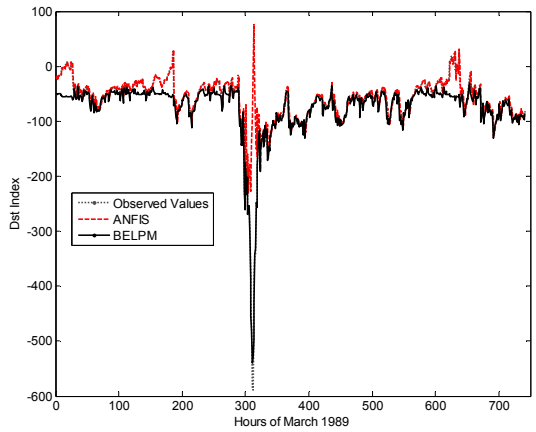
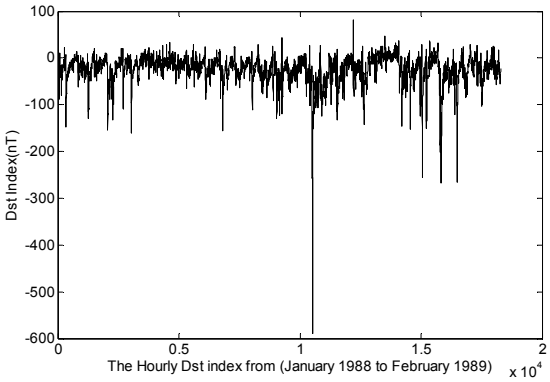


Fig.10. A black box of the BELPM to receive the Dst index and predict one step ahead of the Dst index.

The predicted values of the Dst index versus the observed values of the first data set includes the Dst index of 13th March 1989 which has been described in Fig.11(a). The training data samples of Dst index has been depicted in Fig. 11(b). This figure shows that BELPM can outperform ANFIS in modeling the Dst index.



(a)



(b)

Fig.11. (a). The observed values and predicted values by ANFIS and BELPM. (b). The training data samples.

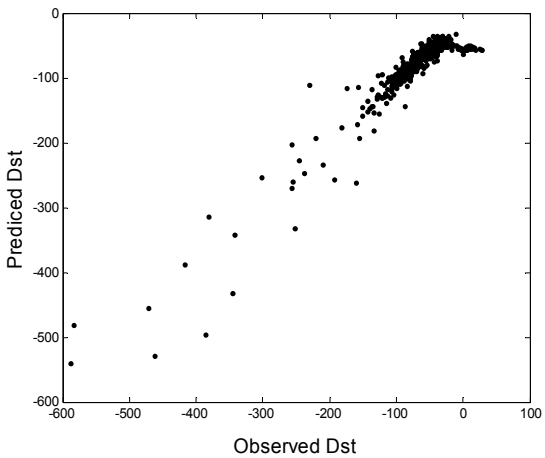


Fig.12. The correlation between the predicted values and the observed values of March 1989.

The correlation between the predicted values and the observed values are shown in Fig.12. The obtained NMSE of BELPM is 0.1041; while the NMSE index of ANFIS is more than 0.15.

For the second data set, BELPM is tested for multi-step ahead prediction of the Dst values. The main goal of this case is to evaluate the performance of BELPM for long term prediction of the Dst index. Figure 13. shows the predicted values by BELPM versus the observed values. For one step ahead prediction of the Dst index from 10th to 22nd of July 2000, the NMSE index of BELPM equals to 0.0593; while the NMSE index of ANFIS is equal 0.112. As this figure describes BELPM could predict the Dst index better than ANFIS. Figure 14 shows how increasing the prediction horizon causes increases in the prediction errors of BELPM and ANFIS. It is notable that the values of the NMSE index of BELPM are lower than the values of the NMSE index of ANFIS.

Tabel I compares the performance of three methods on six -steps ahead prediction of the Dst index. It is notable that the BELPM outperforms ANFIS and a neural network method in terms of NMSE and correlation coefficent.

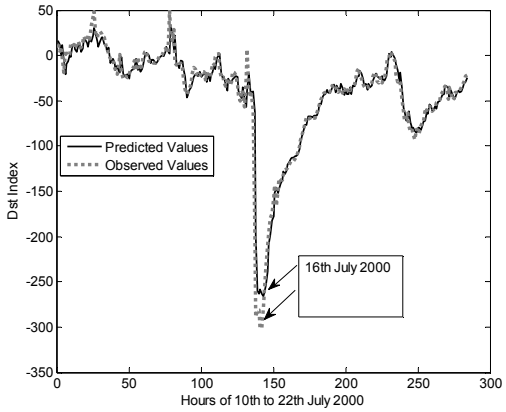


Fig.13.The predicted values of Dst index of 2000

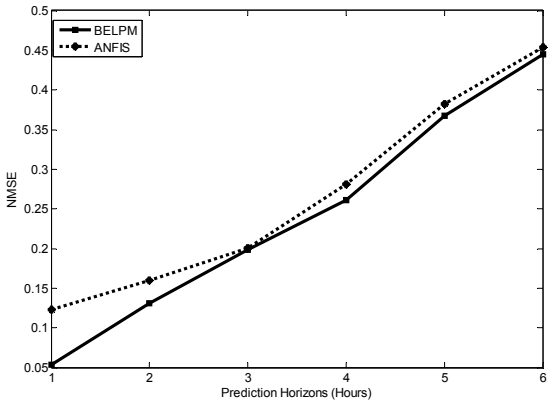


Fig.14. The values of NMSE index versus the prediction horizon of the Dst time series.

TABLE I
A COMPARISON BETWEEN BELPM, ANFIS AND ANN

Method	NMSE	Correlation
ANN[13]	-----	0.75
BELPM	0.44	0.75
ANFIS	0.45	0.74

V. CONCLUSION

This study has suggested another model of BELIMs that is referred to as Brain Emotional Learning-based Prediction Model (BELPM). This paper examined this model (BELPM) to model the Dst index and has been tested for prediction geomagnetic storms using two data sets of the Dst index. The results have verified that the proposed model can be used for long-term prediction of geomagnetic storms.

As future work, the authors consider adding an evolutionary-based optimization method e.g., genetic algorithm to find optimal values of the fiddle parameters, e.g., the number of neighbors and the initial values of nonlinear parameters. Other improvements in the model would be made on the basis of kd-Tree data structure to address "the curse of dimensionality" problem and decrease the computational time complexity of BELPM. To adjust the nonlinear parameters, different types of optimization methods (e.g., Quasi-Newton or Conjugate Directions) may be utilized. The good results obtained when employing the BELPM for predicting geomagnetic storms are a motivation for applying this model to classify patterns as well as to identify complex systems.

ACKNOWLEDGEMENT

The authors are grateful for accessing the Dst index provided by NOAA and World Data Center for Geomagnetism and Space Magnetism, Kyoto University

REFERENCES

[1] L. Xinlin Li, B.Luo and M.Temerin, "Prediction of the Dst, AL, AU Indices Using Solar Wind Parameters,'' Geophysical Research Abstracts, Vol. 15, EGU2013-3645, 2013. doi:10.1029/2006JA011918

[2] J. Sharifi, B.N. Araabi and C. Lucas, Multi_step prediction of Dst index using singular spectrum analysis and locally linear neurofuzzy modeling, Earth Planets Space, 2006, vol. 58, pp. 331–341. doi: 10.1186/BF03351929.

[3] M. Mirmomeni, and C. Lucas, Analyzing the variation of embedding dimension of solar and geomagnetic activity indices during geomagnetic storm time, Earth Planets Space, 61, 237–247, 2009., doi:10.1186/BF03352904.

[4] M. Mirmomeni, M. Shafiee, C. Lucas, and B. N. Araabi, Introducing a new learning method for fuzzy descriptor systems with the aid of spectral analysis to forecast solar activity, J. Atmos. Sol.-Terr. Phys., 68,2061–2074, 2006. DOI:10.1016/j.jastp.2006.07.001.

[5] http://en.wikipedia.org/wiki/March_1989_geomagnetic_storm

[6] S. Kugblenu, S. Taguchi, and T. Okuzawa, Prediction of the geomagnetic storm associated Dst index using an artificial NN algorithm, Earth PlanetSci., 51, 307–313, 1999.

[7] D. Jankovičová, P. Dolinský, F. Valach, and Z. Vörös, Neural network-based nonlinear prediction of magnetic storms, J. Atmos. Sol. Terr. Phys., 64, 651–656,2002. DOI:10.1016/s1364-6826(02)00025-1

[8] R. Bala, and P. Reiff, Improvements in short-term forecasting of geomagnetic activity, Space Weather, 10, S06001,2012, doi:10.1029/2012SW000779.

[9] H. Gleisner, H. Lundstedt, and P. Wintoft,, Predicting geomagnetic storms from solar wind data using time delay neural networks, Ann. Geophys., 14, 679, 1996.

[10] Z. Voros, D. Jankovicova, "Neural network prediction of geomagnetic activity: a method using local Holder exponents," Nonlinear Processes in Geophysics, no. 9, pp. 425 - 433, 2002.

[11] M. Mirmomeni and C. Lucas, "Analyzing the variation of lyapunov exponents of solar and geomagnetic activity indices during coronal mass ejections," Space Weather, vol. 7, p. S07002, July 2009. DOI: 10.1029/2008SW000454

[12] W.K. Tobiska, D. Knipp, W. J. Burke, D. Bouwer, J. Bailey, D. Odstrcil, M. P. Hagan, J. Gannon, and B. R. Bowman (2013), The Anemomilos prediction methodology for Dst, Space Weather, 11, 490–508, doi:10.1002/swe.20094.

[13] S. Haykin, Neural Networks: *A Comperhensive Foundation*.Upper Saddle River, NJ:Prentice Hall, 2nd ed., 1999.

[14] O. Nelles, *Nonlinear System Identification:From classicical Approches to Neural Networks and Fuzzy Models*. Berlin, Germany: Springer-Verlag, 2001.

[15] T. Babaie, R. Karimizandi, C. Lucas, "Learning based brain emotional intelligence as a new aspect for development of an alarm system,'' *J. Soft Computing*., vol. 9, issue 9, pp.857-873, 2008. DOI:10.1007/s00500-007-0258-8

[16] A. Golipour,C.Lucas, D. Shamirzadi, Purposeful prediction Of Space Weather Phenomena by Simulated Emotional Learning. *Modeling Journal,* 24,2004. 65-72.

[17] E. Lotfi and M.R. Akbarzadeh-Totonchi, "Adaptive brain emotional decayed learning for online prediction of geomagnetic activity indices", ;presented at Neurocomputing, 2014, pp.188-196.DOI: 10.1016/j.neucom.2013.02.040

[18] M. Parsapoor, M, U. Bilstrup, "Neuro-fuzzy models, BELRFS and LoLiMoT, for prediction of chaotic time series," in Proc. IEEE Int. Conf. INISTA., pp.1-5, 2012.doi: 10.1109/INISTA.2012.6247025

[19] M. Parsapoor, U. Bilstrup, "Brain Emotional Learning Based Fuzzy Inference System (BELFIS) for Solar Activity Forecasting," in Proc. IEEE Int. Conf. ICTAI 2012, 2012. DOI:10.1109/ICTAI.2012.78

[20] M. Parsapoor, U. Bilstrup, "Brain Emotional Learning Based Fuzzy Inference System (Modified using Radial Basis Function)," 8th IEEE International Joint Conference for Digital InformationManagement, 2013. DOI: 10.1109/ICDIM.2013.6693994.

[21] M. Parsapoor, C. Lucas and S. Setayeshi, "Reinforcement _recurrent fuzzy rule based system based on brain emotional learning structure to predict the complexity dynamic system," *in Proc. IEEE Int. Conf. ICDIM,* pp.25-32, *2008*. Doi: 10.1109/ICDIM.2008.4746712

[22] M. Pasrapoor, U. Bilstrup, "An emotional learning-inspired ensemble classifier (ELiEC)," *Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on* , vol., no., pp.137,141, 8-11 Sept. 2013.

[23] M. Parsapoor and U. Bilstrup, "Chaotic Time Series Prediction Using Brain Emotional Learning Based Recurrent Fuzzy System (BELRFS)," in *International Journal of Reasoning-based Intelligent Systems*, 2013. DOI: 10.1504/IJRIS.2013.057273.

[24] M. Parsapoor, *Prediction the price of Virtual Supply Chain Management with using emotional methods. M.S. thesis,* Dept. Computer. Eng., Science and research Branch, IAU.,

[25] M. Parsapoor, *Brain Emotional Learning-Inspired Models. Licentiate dissertation*, Halmstad: Halmstad University Press, 2014.

[26] J.Moren, C.Balkenius,"A computational model of emotional learning in the amygdala,''in From Animals to Animats, MIT, Cambridge, 2000.

[27] M. S. Gazzaniga, R. B, Ivry, G.R.Mangun, and Megan.S.Steven, *Gognative Nerosc in The Biology of the Mind.* W.W.Norton&Company, New York, 3rd ed., 2009.

[28] J.E.Ledoux, *The emotional brain: the mysterious underpinnings of emotional life,* Simon & Schuster,NY ,1998.

[29] C. Lucas, D. Shahmirzadi, N. Sheikholeslami, "Introducing BELBIC: brain emotional learning based intelligent controller,'' *J. INTELL. AUTOM. SOFT. COMPUT.,* vol. 10, no. 1, pp. 11-22, 2004. DOI: 10.1080/10798587.2004

[30] N. Sheikholeslami, D. Shahmirzadi, E. Semsar, C. Lucas., "Applying Brain Emotional Learning Algorithm for Multivariable Control of HVAC Systems,", *J. Intell. Fuzzy. Syst.* Vol. 16, pp. 1–12, 2005.

[31] A. R. Mehrabian, C. Lucas, J. Roshanian,"Aerospace Launch Vehicle Control: An Intelligent Adaptive Approach", *J. Aerosp. Sci. Technol.,* vol. 10, pp. 149–155, 2006. DOI: 10.1016/j.ast.2005.11.002

[32] R. M. Milasi, C. Lucas, B. N. Araabi, "Intelligent Modeling and Control of Washing Machines Using LLNF Modeling and Modified BELBIC," in *Proc. Int. Conf. Control and Automation.,* pp.812-817, 2005. DOI: 10.1109/ICCA.2005.1528234

[33] A. M. Yazdani1, S. Buyamin1, S. Mahmoudzadeh2, Z. Ibrahim1 and M. F. Rahmat1., "Brain emotional learning based intelligent controller for stepper motor trajectory tracking," *J. IJPS.,* vol. 7, no. 15, pp. 2364-2386, 2012. DOI: 10.5897/IJPS11.1590

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 43–50
DOI: 10.15439/2014F389
ACSIS, Vol. 2

Parallel Feature Selection Algorithm based on Rough Sets and Particle Swarm Optimization

Mateusz Adamczyk
Faculty of Mathematics, Informatics, and Mechanics
University of Warsaw
Banacha 2
02-097 Warszawa
Poland
Email: adamczyk@mimuw.edu.pl

***Abstract*—The aim of this paper is to propose a new method of solving feature selection problem. Foundations of presented algorithm lie in the theory of rough sets. Feature selection methods based on rough sets have been used with success in many data mining problems, but their weakness is their computational complexity. In order to overcome the above-mentioned problem, researches used diverse approximation techniques. This paper presents a new approach to approximation of reducts.**

Particle swarm optimization (PSO) is a stochastic meta-heuristic similar to genetic algorithms. The idea is to see each potential solution as a particle with certain velocity flying through the problem space. The PSO finds optimal solutions by interactions of individuals in population. The main advantage of the PSO over genetic algorithms, is that PSO does not require complex operators such as crossover or mutation. It only uses simple mathematical operators to update position and velocity of each particle, which makes PSO computationally inexpensive in terms of both memory and runtime.

The presented feature selection algorithm treats each feature subset as separate particle. Optimal subset, in terms of selected measure, is discovered as particles fly within the problem space. In order to speed up calculations and balance usage of hardware resources (processors, memory), parallel asynchronous version of PSO is applied. It is based on scheduling calculations of complex fitness function on slave processors, while the main one is responsible for updating particles data and checking algorithm's convergence. Applied approach scales well and provides balanced usage of given resources even if it is not feasible to use the same computational power of every processor, for instance when used resources are not homogeneous.

Proposed method was tested on selected set of data sets from the UCI repository and results were compared to some of the classical algorithms.

***Index Terms*—Feature Selection, Rough Sets, Particle Swarm Optimization, Parallel Asynchronous Particle Swarm Optimization**

I. Introduction

FEATURE selection is one of phases in data mining. The idea is to select subset of attributes, which preserves knowledge for given information or decision system. There are two main reasons for doing it. Firstly, it is some kind of data compression which eases comprehension of analysed data. Secondly, most classifiers are better trained on non-redundant data (compare with [6]). Moreover, removal of superfluous attributes leads to smaller data, which will quicken classifier training.

There are many algorithms for feature selection. Some of them are simple filters concerning every attribute separately. The more useful ones try to rank sets of attributes. They have also high computational complexity (see [1]). One of methods for finding „good" subset of attributes is calculating reducts. It is based on rough sets theory. The reduct is a set of attributes which: preserve information contained in full set of attributes and is minimal, i.e. removing one attribute from reduct leads to losing some knowledge about data. Unfortunately, calculating reducts is quite complex [7].

Feature selection algorithms can be divided into two groups: *filters* and *wrappers* ([2] and [3]). The most standard filters select attributes based on some measure and are classifier agnostic. On the other hand, wrappers use selected classifier accuracy as a measure of quality of subset of attributes. By doing that, they are somewhat „tied" to used classifier and their results will probably not work well with other learning algorithms.

The aim of this paper is to present a new feature selection algorithm. It is wrapper algorithm, that uses classical, exhaustive method for finding reducts and inducing decision rules from them. Because reducts are induced on a subset of attributes, time of their calculations should be smaller than time for obtaining reducts from all attributes. Finding new candidates for reducts is done by using particle swarm optimization. The idea was inspired by work presented in [4]. Particle swarm optimization is meta-heuristic originally developed as a simulation of birds flocking. Because it is a little simpler to apply and implement than genetic algorithms, it has gained researchers attention lately. In order to further speed up the approximation, parallel asynchronous version of particle swarm optimization ([14]) was applied.

The rest of this paper is divided as follows. Section II contains basic information about rough sets theory and reducts calculations. In section III particle swarm optimization has been presented. Above-mentioned section also contains com-

This research was partly supported by Polish National Science Centre (NCN) grants DEC-2011/01/B/ST6/03867 and DEC-2012/05/B/ST6/03215.

978-83-60810-58-3/$25.00

parison of synchronous and asynchronous versions of this meta-heuristic. Subsection III-C of section III contains detailed information about algorithm presented in [4]. Proposed algorithm for feature selection is described in section IV, where it is also compared with method from [4]. In the same section, experimental results: comparison of reducts and found attributes subsets, as well as parallelization gains are also presented. Section V contains conclusions and possible extensions to presented research.

II. Rough Sets

The development of rough sets theory was started by professor Z. Pawlak in 1981 [5]. Its main purpose is to deal with uncertainty of information or decision systems. *The information system* $\mathcal{A}$ is a pair of a non-empty, finite set of objects U called *universe* and a non-empty, finite set of their *attributes* A (see equation 1).

$$\mathcal{A} = (U, A) \tag{1}$$

An attribute is a function $a : U \to V_a,\ \forall_{a \in A}$, where set V_a is called a value set of a. *The decision system* is an information system extended by one distinguished attribute $d \notin A$ called *decision*.

As objects in information system are described only by attributes, two cases can occur:

- different objects can have the same values on all attributes, or
- some attributes can be superfluous.

To deal with the former case, an indiscernibility relation is used. More formally:

$$IND_{\mathcal{A}}(B) = \left\{(x, x') \in U^2 : \forall_{a \in B} a(x) = a(x')\right\} \tag{2}$$

The $IND_{\mathcal{A}}(B)$ from equation 2 is called *B-indiscernibility relation*. If $(x, x') \in IND_{\mathcal{A}}(B)$, then objects x and x' are indiscernible from each other by attributes from B. The equivalence classes of the B-indiscernibility relation are denoted $[x]_B$.

For a subset of objects $X \in U$ and a subset of attributes $B \in A$, X can be approximated only by attributes from B by constructing its *lower approximation* $\underline{B}X = \{x : [x]_B \subseteq X\}$ and its *upper approximation* $\overline{B}X = \{x : [x]_B \cap X \neq \emptyset\}$. The objects in $\underline{B}X$ are certainly in X basing on knowledge in B, whereas objects from $\overline{B}X$ can be classified only as possible elements of X on the basis of knowledge in B. The set $BN_B(X) = \overline{B}X \setminus \underline{B}X$ is called a B *boundary region of* X, and consists of objects which we cannot decisively classify into X on the basis of knowledge in B. A set is said to be *rough* (respectively *crisp*) if the boundary region is non-empty (respectively empty).

In order to speed up classification and clarify knowledge about data, redundant attributes can be removed. To do so, one can keep only those attributes, which preserve indiscernibility relation and hence set approximation. Rejected attributes were redundant since their removal has not worsened classification accuracy. There are usually several such subsets and those which are minimal in terms of cardinality are called *reducts*. It can be shown that the number of reducts of an information system with m attributes may be equal to $\binom{m}{\lceil \frac{m}{2} \rceil}$. Moreover, finding minimal reduct, i.e. a reduct the cardinality of which is the smallest among other reducts, is NP-hard [7].

In my experiments I have used classical algorithm for calculating reducts. Its detailed description can be found in [7]. The algorithm is executed in two steps: calculating a *discernibility matrix* and finding all prime implicants of a *discernibility function* induced from the discernibility matrix. Obtaining decision rules from reducts is straightforward: for each pair of object and reduct those attributes and their values are taken from object, that are also in reduct. Such pairs create conditional part of the rule. The decision part is made from decision for analysed object.

For the information system $\mathcal{A}$ with n objects, the *discernibility matrix* is a symmetric $n \times n$ matrix with entries given in equation 3.

$$\forall_{i,j \in \{1,2,\ldots,n\}}\ c_{ij} = \{a \in A : a(x_i) \neq a(x_j)\} \tag{3}$$

The *discernibility function* $f_{\mathcal{A}}$ for an information system $\mathcal{A}$ is a Boolean function of m Boolean variables $a_1^*, a_2^*, \ldots, a_m^*$, which correspond to attributes $a_1, a_2, \ldots, a_m$, defined as on equation 4, where $c_{ij}^* = \{a^* : a \in c_{ij}\}$.

$$f_{\mathcal{A}}(a_1^*, a_2^*, \ldots, a_m^*) = \bigwedge \left\{ \bigvee c_{ij}^* : i, j \in \{1, 2, \ldots, n\} \wedge c_{ij}^* \neq 0 \right\} \tag{4}$$

The set of all prime implicants determines the set of all reducts of $\mathcal{A}$.

III. Particle Swarm Optimization

Particle swarm optimization is a stochastic meta-heuristic developed by Eberhart and Kennedy in 1995 [11]. It was originally created to graphically model behaviour of bird flocking or fish schooling. Initial simulations were transformed into optimization algorithm, and later enhanced by introducing inertia weight [12].

A. Synchronous Particle Swarm Optimization

Particle swarm optimization is an algorithm similar to genetic algorithm [15]. In both cases solutions are mapped into parts of population: in case of particle swarm optimization population consists of particles, whereas in genetic algorithms there are individuals or phenotypes who form population. In both cases near-optimal solution is found as an individual which is the best fitted one, where fitness measure is the optimized function. Individuals in a new population are created by interactions between parts of the previous one.

Particle swarm optimization is initialized with a random set $P = \{p_1, p_2, \ldots, p_k\}$ of particles. Each particle $p_i, i \in \{1, 2, \ldots, k\}$ has:

- a position x_i in S dimensional space,
- a velocity v_i,
- memory of personal best position $best_i$.

There is also stored position $best_g$ of the best particles found so far.

In each population, all particles' positions are updated with formula 5 and particles' new fitnesses are calculated.

$$x_i(t+1) = x_i(t) + v_i(t) \tag{5}$$

Best positions: $best_g$ and $best_i$ are updated if necessary. Particles accelerate according to a formula 6.

$$\begin{aligned} v_i(t+1) = w(t) \cdot v_i(t) + \\ c_1 \cdot rand_1() \cdot (best_i - x_i) + \\ c_2 \cdot rand_2() \cdot (best_g - x_i) \end{aligned} \tag{6}$$

Velocities cannot be larger than some constant v_{max}. If they were, particles would fly too fast and probably miss subspaces containing optimal solutions. The v_{max} constant should be large enough, to allow particles to escape regions with sub-optimal solutions.

The $rand_1$ and $rand_2$ are uniformly distributed random functions in $[0, 1]$. Algorithm's parameters: c_1 and c_2 define particles' acceleration constants. Their high values correspond to high attraction of past sub-optimal solutions, whereas low values allow particles to roam far from target regions. The c_1 constant corresponds to personal best solution, and the c_2 determines how firmly particle follows the flock.

The w in equation 6 is called inertia [12]. It is positive linear function of time. Choosing proper inertia is crucial for providing balance between local and global exploration, thus to ensuring that optimal solution is found in small number of iterations.

In equation 6 one can distinguish three parts. The first one corresponds to particle's „memory". The second one, controlled with c_1 constant, is linked to particle's „cognition". The third part, which is governed by constant c_2, describes „social" behaviour of particle. It is responsible for collaboration among particles.

When all particles in the population are updated, particle swarm optimization algorithm checks its convergence. If the best solution found so far is good enough, then calculations are stopped. If found solution is not good enough, then the whole process of moving particles and obtaining their statistics is repeated.

B. Asynchronous Particle Swarm Optimization

In order to speed up particle swarm optimization algorithms, there were proposed many parallelization strategies. Some of them were based on communication strategies similar to ones used with genetic algorithms [13]. In [14] authors proposed significant change to particle swarm optimization algorithm: asynchrony. By making algorithm asynchronous, authors made it converge faster to some optimal solution. It is worth noting, that above-mentioned change created algorithm, whose results will probably be different from the ones obtained from the classical version.

In the asynchronous version of particle swarm optimization, after obtaining fitness for one particle, convergence check and updates are done. It leads to dynamically updated global best position, which can be modified after updating one particle and not after updating whole population as in synchronous version of algorithm.

It is noteworthy that asynchronous version of particle swarm optimization running sequentially will produce different results than classical particle swarm optimization. The cause of that difference is the above-mentioned difference in strategies for updating data of global best particle. If the global best position would be updated in the middle of processing one population, then the rest of population would move differently in asynchronous than in the synchronous version of particle swarm optimization.

Parallelization of asynchronous particle swarm optimization is straightforward, when done in master-slave architecture. The master processor is responsible for updating particles, checking convergence and scheduling fitness calculations on slave processors. Slave processor evaluates fitness of given particle's position and returns obtained value to master processor. Communication between master processor and slave ones is done with use of first-in-first-out task queue. As master processor process one particle, the one from the front of the task queue, at time and later schedules its data to slave processor, dynamic load balancing is done implicitly. If some slave processor is slower or more loaded, then it will calculate fitness slower than other ones. It will lead to scheduling more work on faster or less busy processors, because they will more often get tasks from master processor.

C. Feature Selection using Particle Swarm Optimization and Rough Sets

In [4], feature selection algorithm based on particle swarm optimization and rough sets theory was presented. In order to exploit particle swarm optimization for finding relevant attributes, some adaptations were necessary.

1) Representation of Position: For decision system with m attributes particle position was coded as a binary bit string of length m. If i-th attribute ($i \in \{1, 2, \ldots, m\}$) was chosen, then i-th bit of position string was set to 1. Otherwise, it was set to 0. More formally, there was defined bidirectional mapping from a power set $\mathcal{P}(A)$ of the set of attributes A into space of binary string of length m: $\mathcal{M} : \mathcal{P}(A) \longrightarrow \{0, 1\}^m$, such that for $R \subseteq A$, the condition from equation 7 holds.

$$\forall_{i \in \{1,2,\ldots,m\}} \mathcal{M}(R)_i = \begin{cases} 0 & a_i \notin R \\ 1 & a_i \in R \end{cases} \tag{7}$$

The movement of particle corresponds to modifying subset of attributes in order to find a better subset. If i-th bit of particle's position was set from zero to one, then the i-th attribute was added to subset. It i-th bit was set to zero, then the i-th attribute was removed. If proper representation of velocity and fitness measure were chosen, then particles flying towards the best position will correspond to finding subsets of A with most relevant attributes.

2) Representation of Velocity: The speed of particle was represented as a positive integer, varying from 1 to v_{max}. Value of the velocity shows how many bits of particle's position

should be changed in the particular moment of time to be the same as in the global best position. In other words, particles fly through problem space towards the current best position.

In order to update particles' speeds and positions, the authors of [4] proposed notion of positions difference. The *difference of positions* is equal to component-wise difference of positions seen as two vectors in m-dimensional space. For instance, if $a = [1, 0, 1, 0, 1]$ and $b = [1, 1, 0, 0, 0]$, then: $a - b = [0, -1, 1, 0, 1]$. The 1's' in difference correspond to those bits, which should be set in b to make b equal to a. Similarly, the -1's denote which bits in b should be unset. To use positions' difference in equation 6, it should be converted to integer. To do so, the authors of [4] proposed sum of all difference's components. In the previous example, it is: $|a - b| = \sum_{j=1}^{m} (a - b)_j = 1$. If updated speed was smaller than one, it was set to 1.

Formulae for updates of particle's velocity is presented on equation 8. The $i \in \{1, 2, \ldots, k\}$ on equation 8 denotes particle's index.

$$\begin{aligned} v_i(t+1) = \min(v_{max}, \max(1, \\ w(t) \cdot v_i(t) + \\ c_1 \cdot rand_1() \cdot \sum_{j=1}^{m} (best_i - x_i)_j + \\ c_2 \cdot rand_2() \cdot \sum_{j=1}^{m} (best_g - x_i)_j \\ )) \end{aligned} \tag{8}$$

As it is mentioned above, particle's position is updated to move particle towards the global best position. The two cases are possible:

1) $v_i \leq |best_g - x_i|\,,\ i \in \{1, 2, \ldots, k\}$,
2) $v_i > |best_g - x_i|\,,\ i \in \{1, 2, \ldots, k\}$.

In the first case, v_i random bits, which are different than the ones in $best_g$, are changed. That way, particle flies towards the global best position, but doing random search instead simply being the same as the best. In the second case, apart from flipping all of the bits which are different from the ones in the $best_g$, the $v_i - |best_g - x_i|$ the similar ones are also flipped. It can be interpreted as a particle flying past the best position and exploring more regions.

3) Fitness function: The fitness function used in [4] is presented on equation 9.

$$\mathcal{F}(p_i) = \alpha \cdot \gamma_R(d) + \beta \cdot \frac{m - |R|}{m} \tag{9}$$

The α and β are two parameters corresponding to the importance of classification and subset length, $\alpha \in [0, 1]$, $\beta = 1 - \alpha$. The $\gamma_R(d)$ is classification quality of condition attribute set R, relative to decision d, $|R|$ is its cardinality and $m = |A|$.

To measure classification quality of a condition attribute set R, the authors of [4] used the LEM2 algorithm ([8]) for inducing rules from set R and rule negotiation in classification ([9]). The final score was obtained by doing ten-fold cross validation.

IV. Proposed algorithm and conducted experiments

A. Proposed algorithm

Proposed algorithm is a fusion of two above-mentioned methods. It is slightly changed algorithm presented in [4], which is shortly described in section III-C. Instead of the LEM2 algorithm, exhaustive algorithm for finding reducts (see section II) has been used. Encoding particle's position, velocity and their update strategies was the same, as presented in section III-C1 and III-C2. The fitness function was also the same as the one shown in section III-C3.

The main advantage of proposed method over the one presented in [4], is usage of asynchronous particle swarm optimization. Every time particle fitness is obtained, that particle data are updated. After update, algorithm checks convergence and, if necessary, updates the global best position. Because of frequent updates, particles react more dynamically to finding new best solution. But the main trait of proposed change is to allow exploitation of parallel architecture of modern processors. I have used parallel asynchronous particle swarm optimization [14] and fused it with algorithm presented in [4].

B. Experimental results

1) Experimental setting: The sixth version of the Java language was chosen as an implementation language for proposed algorithm. The Rseslib library was used as a source of implementation of exhaustive algorithm for reducts calculation, inducing classifier (see section II for algotithm's details) and doing cross validation. As in [4], ten-fold cross validation was used. The parallel asynchronous particle swarm optimization [14] was implemented within Data Mining EXpressions Library (dmexl), which provides framework for implementing parallel data mining algorithms, especially the ones for feature selection. The dmexl library is being developed by the author of this paper.

Most of the algorithm's parameters were set to be equal to ones presented in [4]. The α was set to 0.9, and β – to 0.1, as in [4]. See section III-C3 and equation 9 for detailed description of above-mentioned parameters. Acceleration constants: c_1 and c_2 were both set to 2. As in [4], the v_{max} parameter was chosen to be equal to $\frac{m}{3}$. The value of minimal fitness, which could stop algorithm execution before reaching requested number of populations, was set to 0.85.

The inertia weight (see equation 6) was the same as in [4] and it is presented on equation 10.

$$w(t+1) = (w(t) - 0.4) \cdot \frac{P_{no} - t}{P_{no} + 0.4} \tag{10}$$

The P_{no} in equation 10 is the number of populations to simulate. The inertia weight is linear function which decreases with time, and varies between 1.4 and 0.4.

http://www.oracle.com/technetwork/java/javase/overview/index-jsp-136246.html

http://rseslib.mimuw.edu.pl/

https://github.com/mateka/dmexl

Experiments were conducted on a personal computer equipped with: Intel Core i7-4700HQ quad core CPU and 32GB of RAM. Amount of runtime memory available to java virtual machine was limited to 6GB.

2) Used data tables: Tests were conducted on set of fifteen data tables from UCI repository [16]. Selected tables are listed on table I. Statistics for selected tables are presented on table II.

TABLE I
DATA TABLES USED IN EXPERIMENTS

Table name	URI
Balloon 1	http://archive.ics.uci.edu/ml/datasets/Balloons
Balloon 2	http://archive.ics.uci.edu/ml/datasets/Balloons
Balloon 3	http://archive.ics.uci.edu/ml/datasets/Balloons
Balloon 4	http://archive.ics.uci.edu/ml/datasets/Balloons
Hayes-Roth	http://archive.ics.uci.edu/ml/datasets/Hayes-Roth
Voting	https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records
Lenses	http://archive.ics.uci.edu/ml/datasets/Lenses
Lung Cancer	http://archive.ics.uci.edu/ml/datasets/Lung+Cancer
Monk 1	http://archive.ics.uci.edu/ml/datasets/MONK's+Problems
Monk 2	http://archive.ics.uci.edu/ml/datasets/MONK's+Problems
Monk 3	http://archive.ics.uci.edu/ml/datasets/MONK's+Problems
Postoperative	https://archive.ics.uci.edu/ml/datasets/Post-Operative+Patient
Promoters	http://archive.ics.uci.edu/ml/datasets/Molecular+Biology+%28Promoter+Gene+Sequences%29
Tic Tac Toe	https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame
Zoo	http://archive.ics.uci.edu/ml/datasets/Zoo

TABLE II
BASE STATISTICS FOR DATA TABLES USED IN EXPERIMENTS

Table name	Attributes	Objects	Decision classes
Balloon 1	4	16	2
Balloon 2	4	20	2
Balloon 3	4	20	2
Balloon 4	4	20	2
Hayes-Roth	5	132	3
Voting	16	435	2
Lenses	4	24	3
Lung Cancer	56	32	3
Monk 1	7	432	2
Monk 2	7	432	2
Monk 3	7	432	2
Postoperative	8	90	3
Promoters	58	106	2
Tic Tac Toe	9	958	2
Zoo	17	101	7

For all data tables three experiments were made in RSES (see [10]). All of them were ten-fold cross validation classifications. In the first case, exhaustive reducts calculation algorithm was used. There was no rule shortening and conflicts were resolved by „simple voting". This case is denoted by *Exhaustive 1* on Table III. Two other experiments used rule shortening with rule shortening ratio set to 1.0 and conflicts were resolved by „standard voting". These are a default settings in RSES. Algorithms used with these settings were: exhaustive algorithm (*Exhaustive 2*) and LEM2 algorithm. All obtained classification accuracies are presented on table III. Even with 16GB of RAM, RSES was unable to calculate reducts and rules for Lung Cancer and Promoters tables when using exhaustive algorithms, with or without shortening obtained rules.

http://logic.mimuw.edu.pl/~rses/

TABLE III
RSES ALGORITHM ACCURACIES ON USED DATA TABLES

Table name	Exhaustive 1	Exhaustive 2	LEM2
Balloon 1	0.40	0.70	0.40
Balloon 2	0.70	1.00	1.00
Balloon 3	0.60	1.00	1.00
Balloon 4	0.60	1.00	1.00
Hayes-Roth	0.79	0.88	0.93
Voting	0.88	0.95	0.98
Lenses	0.45	0.80	0.85
Lung Cancer	–	–	0.40
Monk 1	0.53	1.00	0.99
Monk 2	0.43	0.49	0.71
Monk 3	0.97	1.00	1.00
Postoperative	0.49	0.42	0.42
Promoters	–	–	0.92
Tic Tac Toe	0.63	0.98	1.00
Zoo	0.73	0.97	1.00

3) Algorithm's accuracy: Proposed algorithm was executed sixty times on each data table. Half of experiments were sequential and thirty were parallel. In each algorithm's execution, there were 20 particles and at most 100 populations. Statistics for obtained accuracies are presented on table IV. The first thing to note, is that both: sequential and parallel version of algorithm have roughly the same accuracies (see Figure 1). Unfortunately, their accuracies are much lower than those of LEM2 from RSES. On average, proposed algorithm had more stable accuracy than exhaustive one from RSES. It is probably due to working on subsets of attributes. It is also important to remember, that in RSES rules were shortened whereas in proposed algorithm they were not. It probably led to overfitting rules in experiments with proposed algorithm.

TABLE IV
STATISTICS FOR ACCURACIES OF SEQUENTIAL AND PARALLEL VERSIONS OF PROPOSED ALGORITHM

Table name	Sequential			Parallel		
	min	max	avg	min	max	avg
Balloon 1	0.00	0.75	0.73	0.00	0.75	0.74
Balloon 2	0.00	0.80	0.80	0.00	0.80	0.79
Balloon 3	0.00	0.80	0.80	0.00	0.80	0.80
Balloon 4	0.00	0.80	0.80	0.00	0.80	0.80
Hayes-Roth	0.00	0.58	0.56	0.00	0.58	0.56
Voting	0.59	0.96	0.94	0.60	0.96	0.95
Lenses	0.00	0.77	0.71	0.00	0.77	0.71
Lung Cancer	0.05	0.58	0.47	0.03	0.58	0.45
Monk 1	0.00	0.75	0.64	0.00	0.75	0.68
Monk 2	0.00	0.67	0.67	0.00	0.67	0.67
Monk 3	0.00	0.81	0.81	0.00	0.81	0.81
Postoperative	0.00	0.71	0.71	0.00	0.71	0.71
Promoters	0.22	0.71	0.66	0.21	0.78	0.65
Tic Tac Toe	0.00	0.70	0.69	0.00	0.70	0.69
Zoo	0.00	0.61	0.61	0.00	0.61	0.61

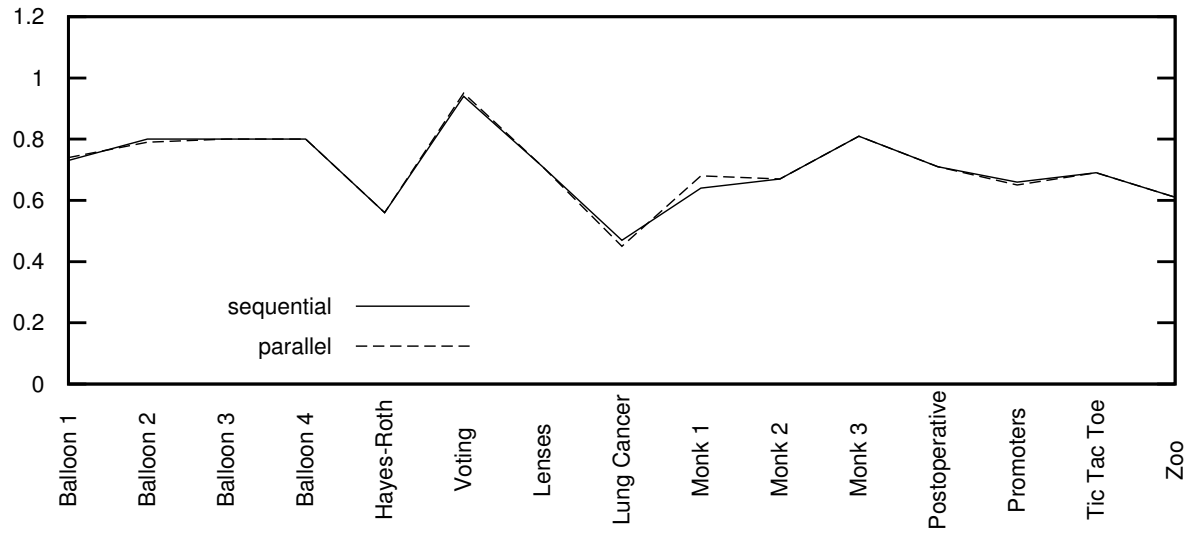


Fig. 1. Accuracies of sequential and parallel versions of proposed algorithm

The interesting conclusion can be drawn from comparing results of exhaustive algorithm without rule shortening and the proposed one. The proposed algorithm performed in all except three cases better than the exhaustive algorithm without rule shortening from RSES. Probably, due to working on smaller sets of attributes, which lead to creation of shorter reducts and rules. The next step in my research is testing proposed algorithm with rule shortening.

Comparison of obtained accuracies from proposed algorithm, exhaustive one from RSES with turned off rules shortening and LEM2 with default settings is presented on Figure 2.

It seems that number of decision classes had no impact on proposed algorithm accuracy, as it was quite stable between different experiments.

TABLE V
SPEEDUP AND EFFICIENCY

Table name	Speedup	Efficiency
Balloon 1	0.59	0.15
Balloon 2	0.58	0.15
Balloon 3	0.58	0.15
Balloon 4	0.58	0.15
Hayes-Roth	1.11	0.28
Voting	19.00	4.75
Lenses	0.60	0.15
Lung Cancer	1.45	0.36
Monk 1	3.76	0.94
Monk 2	3.97	0.99
Monk 3	3.93	0.98
Postoperative	1.00	0.25
Promoters	3.23	0.81
Tic Tac Toe	4.58	1.15
Zoo	1.74	0.43

4) Algorithm's performance: For execution times: T_1 – sequential and T_c – parallel on c processors, *seedup* $S_c = \frac{T_1}{T_c}$ and *efficiency* $E_c = \frac{T_1}{T_c \cdot c}$ are measures for expressing parallelization gains. Those measures for proposed algorithm are shown in table V. The presented values are mean values of all executions for each table. It is worth noting, the number of threads of execution for parallel version of algorithm was set to $4 \cdot c$ in order to fully utilize a given processor. That is the reason why efficiency may be greater than 1. There was an interesting case with Voting table. The sequential version of proposed algorithm have reached maximum number of populations four times, whereas the parallel version reached it only one. That is the reason why speedup for Voting table is so enormous.

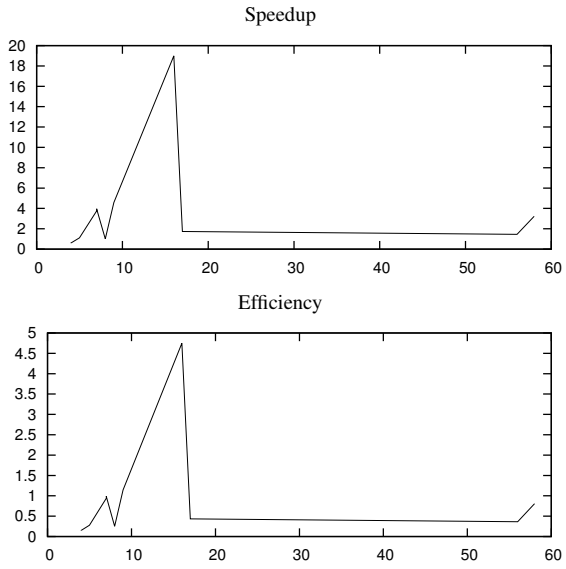


Fig. 3. Relationship between number of attributes and speedup or efficiency

The relationship between the number of attributes and speedup or efficiency is shown on Figure 3. X-axis shows

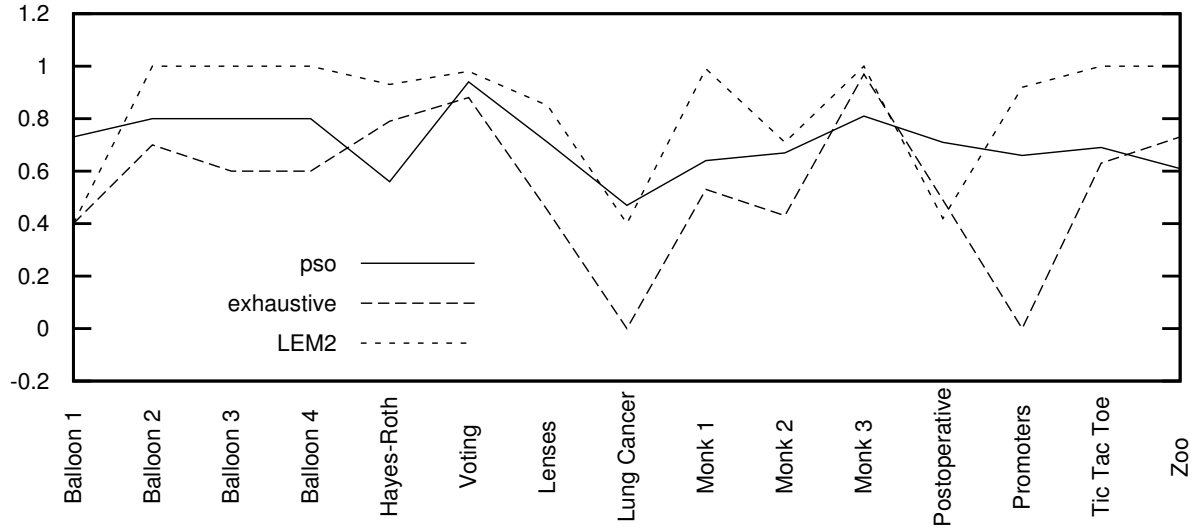


Fig. 2. Accuracies of proposed algorithm (pso), exhaustive one and LEM2

the number of attributes and y-axis shows:

- speedup for the first graph,
- efficiency for the second graph.

Although, with an increasing number of attributes speedup and efficiency are generally growing, the tendency is not clear.

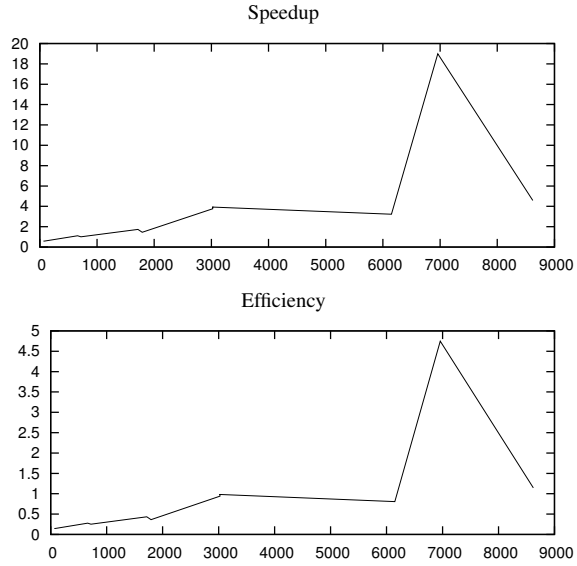


Fig. 4. Relationship between size of table and speedup or efficiency

On Figure 4 the relationship between the size of table and speedup or efficiency is shown. The horizontal axis corresponds to the table size, i.e. the number of attributes multiplied by the number of objects. Similarly to Figure 3, vertical axes correspond to speedup and efficiency. On Figure 4 it can be almost clearly seen, that, with an increasing size of data table, speedup and efficiency grow. The trend is only disturbed by the case of the Voting table, which was already described.

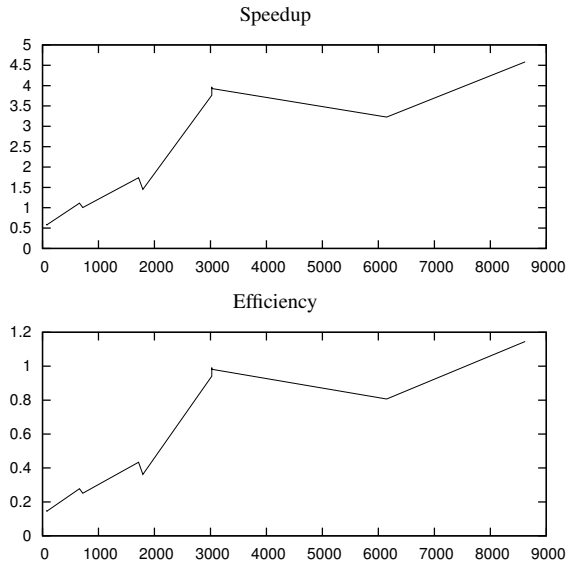


Fig. 5. Relationship between size of table and speedup or efficiency with Voting table removed

Relationship between size of table and speedup or efficiency, with removed the Voting table statistics, is shown on Figure 5. Graphs on Figure 5 clearly show upward trend in speedup and efficiency of proposed algorithm with the growth of the size of the table. When overhead of managing threads and parallelism is compensated by the table size, a parallel version of proposed algorithm performs better than the sequential one.

If there is any relation between the number of decision

classes and performance of proposed algorithm, it cannot be induced from gathered results. The main parameters, which can help to decide if a parallel or sequential version should be used, are: the number of attributes and the number of objects.

V. Conclusion

Although the presented algorithm was not able to achieve astonishing accuracy, it was on par with the ones implemented in RSES system used with default settings. As with all settings, they can be tuned to selected problem to achieve better results. It is noteworthy, that proposed algorithm performed better in most cases than exhaustive algorithm from RSES with rule shortening turned off and simple voting as a method for solving conflicts.

Particle swarm optimization is said to be simpler to apply in purely numerical optimization problems than genetic algorithms. It is due to its straightforward application and there is no need to define complex operators, as in genetic algorithms. Unfortunately, applying particle swarm optimization to feature selection and calculating reducts is not as straightforward as applying it to numerical optimization problems. Coding attributes subsets as binary strings and updating particles' speed and position turned out to be as complicated as when calculating operators in genetic algorithms.

The performance of parallel version of presented algorithm is promising. With growing size of data tables, algorithm's speedup and efficiency were raising. Although, these are optimistic results, there should be examined what impact on these coefficients would have adding more cores or processors.

The performance results are even more optimistic, as presented algorithm is a part of a *Data Mining EXpression Library* (dmexl). The dmexl library is a framework for easier development of data mining algorithms. Currently, efforts are taken into providing building blocks for feature selection algorithms. From the user's perspective, library enables writing complex algorithms as a simple expression. By selecting executor object, the user decides if algorithm should be executed sequentially or in parallel. Some examples can be seen in library's source code, which is available on the Internet: https://github.com/mateka/dmexl.

Another interesting research topic, would be using modification of proposed algorithm to execute feature selection in the context of clustering. As a classification accuracy would be unavailable, some other measure should be used to grade obtained particles. One possibility is to use a measure of coherence of resulting clusters. If this approach is sensible and applicable, further research has to be done.

References

[1] Guyon, I., Elisseeff, A.: An introduction to variable and feature selection, The Journal of Machine Learning Research, 3, 1157–1182, JMLR. org, 2003

[2] Fleuret, F.: Fast binary feature selection with conditional mutual information, The Journal of Machine Learning Research, 5, 1531–1555, JMLR. org, 2004

[3] Das, S.: Filters, wrappers and a boosting-based hybrid for feature selection, ICML, 1, 74–81, Citeseer, 2001

[4] Wang, X., Yang, J., Teng, X., Xia, W., Jensen, R.: Feature selection based on rough sets and particle swarm optimization, Pattern Recognition Letters, 28, 4, 459–471, Elsevier, 2007, http://www.dx.doi.org/10.1016/j.patrec.2006.09.003

[5] Pawlak, Z.: Information systems theoretical foundations, Information systems, 6, 3, 205–218, Elsevier, 1981 http://www.dx.doi.org/10.1016/0306-4379(81)90023-5

[6] Widz, S. and Slezak, D., Rough Set Based Decision Support – Models Easy to Interpret, Selected Methods and Applications of Rough Sets in Management and Engineering, 95-112, Peters, G., Lingras, P., Slezak, D., Yao, Y., Advanced Information and Knowledge Processing, Springer, 2012, http://www.dx.doi.org/10.1007/978-1-4471-2760-4_6

[7] Komorowski, J., Pawlak, Z., Polkowski, L., Skowron, A.: Rough sets: A tutorial, Rough fuzzy hybridization: A new trend in decision-making, 3–98, Springer Verlag, Singapore, 1999

[8] Stefanowski, J.: On rough set based approaches to induction of decision rules, Rough sets in knowledge discovery, 1, 1, 500–529, Heidelberg, Germany: Physica-Verlag, 1998

[9] Bazan, J. G.: A comparison of dynamic and non-dynamic rough set methods for extracting laws from decision tables, Rough sets in knowledge discovery, 1, 321–365, Citeseer, 1998

[10] Bazan, J. G., Szczuka, M.: RSES and RSESlib-a collection of tools for rough set computations, Rough Sets and Current Trends in Computing, 106–113, Springer Berlin Heidelberg, 2001, http://www.dx.doi.org/10.1007/3-540-45554-X_12

[11] Eberhart, R., Kennedy, J.: Particle Swarm Optimization, Neural Networks, 1995., IEEE International Conference on, 1942-1948, IEEE, 1995, http://www.dx.doi.org/10.1109/ICNN.1995.488968

[12] Shi, Y., Eberhart, R.: A modified particle swarm optimizer, Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on, 69–73, IEEE, 1998, http://www.dx.doi.org/10.1109/ICEC.1998.699146

[13] Chu, S.-C., Roddick, J. F., Pan, J.-S.: Parallel particle swarm optimization algorithm with communication strategies, submitted to IEEE Transactions on Evolutionary Computation, 2003

[14] Koh, B., George, A. D., Haftka, R. T., Fregly, B.: Parallel asynchronous particle swarm optimization, International Journal for Numerical Methods in Engineering, 67, 4, 578–595, Wiley Online Library, 2006, http://www.dx.doi.org/10.1002/nme.1646

[15] Goldberg, D.: Genetic algorithms in search, optimization, and machine learning, Addison-wesley Reading Menlo Park, 1989

[16] Bache, K., Lichman, M.: UCI Machine Learning Repository, http://archive.ics.uci.edu/ml, 2013, University of California, Irvine, School of Information and Computer Sciences

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 51–58

DOI: 10.15439/2014F464
ACSIS, Vol. 2

Global versus modular link prediction approach for discapnet: website focused to visually impaired people

O. Arbelaitz, A. Lojo, J. Muguerza, I. Perona
University of The Basque Country
Department of Computer Architecture and Technology
Donostia-San Sebastian, 20018, Spain
Email: {olatz.arbelaitz, aizea.lojo, j.muguerza, inigo.perona}@ehu.es.

***Abstract*—Web personalization becomes essential in industries and specially for the case of users with special needs such as visually impaired people. Adaptation may very much speed up the navigation of visually impaired people and contribute to diminish the existing technological gap. This work is the first stage of a web mining process carried out in discapnet: a website created to promote the social and work integration of people with disabilities where slow navigation has been detected. Based on observation in-use where behaviours emerge applying a web mining process to server log data, we designed a system to generate user navigation profiles and adapt to the web site through link prediction. Two approaches for user profiling were implemented: a global system built based on the complete database and a modular approach carried out discovering the navigation profiles within different zones. Although both approaches are effective, the modular approach outperforms. When 25% of the navigation of the new user has happened the designed system is able to propose a set of links where nearly 60% of them (2 out of 3) is among the ones the new user will be using in the future. This will definitely make the navigation easier saving a lot of time.**

I. Introduction

THE success of electronic commerce, especially for the less well-known companies, is largely dependent on the appropriate design of their website [1]. Chaffey et al. [2] stated in their work that a good website should begin with the users and understanding how they use the channel. This confirms that understanding the needs and preferences of the website audience will help to answer questions about what the content of the website should be, how it should be organized and so on. Organizations have to respond not only by adopting new technologies, but also by interpreting and using the knowledge created by Internet users.

In the last decades, the trends have led to a dramatic increase in the amount of information stored in the web, which often makes the information intractable for users. As a consequence, the general need for websites to be useful in an efficient way for users has become especially important. There is a need for easier access to the required information and adaptation to the users' preferences or needs. Web personalization thus becomes essential in industries and specially for the case of users with special needs such as visually impaired people.

However, little is known about the navigation tactics employed by screen reader users when they face problematic situations on the Web. Modelling the navigation of users is of utmost importance as it allows not only to predict interactive behaviour, but also to assess the appropriateness of the content in a link, the information architecture of a site and the design of a web page [3].

Navigating through audio web interfaces is a challenging task mainly because content is serially rendered. Content serialisation has several negative implications: users cannot get an overview of the page, entailing that users can only catch a glimpse of the page as long as they scan through the document. Consequently navigation across different web pages is a time consuming task and web page exploration is a resource intensive activity that requires a dedicated attention span [3].

The sequential access of screen readers means that visually impaired users take up to five times longer than sighted users to explore a web page [4]; the screen reader itself requires an additional cognitive effort [5] and, moreover, inter-page and intra-page navigation problems are some of the problems that need to be faced when working with visually impaired users [3].

As a consequence, in the case of users with disabilities, adaptation becomes crucial and may very much speed up the navigation and contribute to diminish the existing technological gap. In order to be able to model the user, the modelling component must collect information about a number of observable parameters such as interest, characteristics, etc. This information can be requested to the user in a previous session, but this is annoying, disruptive and can produce false assumptions. Another option is to collect this information in-use while the user is accessing the web, and therefore, to build a non invasive system able to model the users in the wild. In this way the system can learn its interests, likes, etc.

According to Pierrakos et al. [6] web personalization can be defined as the set of actions to dynamically adapt the presentation, the navigation schema and the contents of the website, based on the preferences, abilities or requirements of

978-83-60810-58-3/$25.00

the user. Nowadays, as Brusilovsky et al. [7] describe, many research projects focus on this area, mostly in the context of e-Commerce [7] and e-learning [8]. Important websites such as Google and Amazon are clear examples of this trend.

In any web environment, the contribution of the knowledge extracted from the information acquired from observation in-use is twofold:

- It can be used for web personalization (i.e. for the adaptation of the website according to the user requirements).
- It can also be used to extract knowledge about the interests of the people browsing the website or about the possible design mistakes.

Data mining for web personalization has many advantages. It is not disruptive, it is based on statistical data obtained by real navigation data (decreasing the possibility of false assumptions) and is itself adaptive (when the characteristics of the user change, collected data allows the automatic change of the interaction schema). When the user is a person with physical, sensory or cognitive restrictions, data mining is the easiest (and frequently almost the only) way to obtain information about the uses of the person.

Data mining in this context has also some drawbacks. The most important one is its impact over privacy, due to the need of storing large quantities of data about the users. Diverse laws in different countries protect user rights for privacy. Even if it is difficult to reach a balance among privacy and personalization, some appealing proposals have been recently published.

Web mining can be defined as the application of data mining techniques to data from the Internet. This process has three main stages:

- The data acquisition and pre-processing stage.
- The pattern discovery and analysis phase to find groups of web users with common characteristics related to the Internet and the corresponding patterns or user profile. Machine learning techniques are mainly applied in this phase.
- Finally, the patterns detected in the previous steps are used in the operational phase to adapt the system and make navigation more efficient for new users or to extract important information for the service providers.

This work is the first stage of a web mining process carried out in discapnet website *www.discapnet.es* where we analysed the navigation of users (web usage mining) and built user navigation profiles that provide a tool to adapt the web to new users while they are navigating (through link prediction). Being discapnet website addressed to people with disabilities, mainly to visually impaired people, link prediction will be specially important in the system. This is corroborated somehow because a preliminary analysis of the web logs showed that the time spent in link type or hub type pages is considerably longer than it would be expected to; it is longer than the one spent in pages devoted to content (content pages) and dynamic pages which are mainly related to news. This makes us suspect that the implementation of an efficient link prediction system will definitely help to make navigation easier, and as a consequence, diminish the time spent in link type pages.

So that the link prediction system is efficient, it is important for it to be based on observation in-use. This way behaviours emerge from the obtained data instead of looking for predefined models. Web logs are the most simple in-use information and the ones applicable to the wider set of users. Other more complete tools [9] capturing longitudinally low-level interaction unobtrusively limit the public to be used in the modelling process.

Summarizing, the aim of this paper is to design a link prediction system which contributes to make the navigation of users navigating in discapnet easier. The proposed system is based on observation in-use; behaviours emerge applying a web mining process to the obtained data, web server log data. The web mining process has required a thorough analysis of the environment and the data, a selection of the machine learning tools to be used and, finally, a design and evaluation of the system. After this process, we developed two approaches for user profiling: a global system built based on the complete website and a modular approach carried out discovering the navigation profiles within each zone. Both systems show to be useful for link prediction but the values of the evaluation measures are a bit higher for the modular approach. We consider that the inclusion of the described system in discapnet will contribute to improve inter-page navigation within the website and diminish the times spent in link type pages.

The paper describes the discapnet website and its main characteristics in Section II and the preprocess applied to the data in Section III. Section IV describes the machine learning techniques used for user navigation profile generation whereas Section V is devoted to describing the two profiling options implemented for discapnet. The evaluation of the two systems and their use for link prediction are presented in Section VI. Finally, Section VII summarises the conclusions and future work.

II. Discapnet website analysis

Discapnet is an initiative created to promote the social and work integration of the people with disabilities financed jointly by the Fundación ONCE [10] and Technosite. It contains two main action lines:

- An information service for organizations, professionals, people with disabilities and families.
- A platform to develop actions to promote the involvement of people with disabilities in the economic, social and cultural life.

Technosite provided us the server logs of two serves that store the activity generated in some areas of the web discapnet. The transferred data was basic anonimized server log data in Common Log Format [11] (see Figure 1). It contained all the requests served by two of the servers hosting discapnet website from the 2nd February 2012 to the 31st December 2012.

```
207.46.13.48 - - [22/Feb/2012:00:04:05 +0100] "GET /index.php?...&lang=es HTTP/1.1" 200 30055 "-" "Mozilla/5.0 (cor
207.46.19.49 - - [22/Feb/2012:00:04:07 +0100] "GET /index.php?...&lang=en HTTP/1.1" 200 29646 "-" "Mozilla/5.0 (cor
207.46.19.49 - - [22/Feb/2012:00:04:07 +0100] "GET /index.php?...&lang=es HTTP/1.1" 200 28088 "-" "Mozilla/5.0 (cor
66.249.72.32 - - [22/Feb/2012:00:04:09 +0100] "GET /index.php?...&lang=es HTTP/1.1" 200 29440 "-" "Mozilla/5.0 (cor
207.46.99.49 - - [22/Feb/2012:00:04:12 +0100] "GET /index.php?...&lang=fr HTTP/1.1" 200 28106 "-" "Mozilla/5.0 (cor
207.46.19.49 - - [22/Feb/2012:00:04:13 +0100] "GET /index.php?...&lang=en HTTP/1.1" 200 29557 "-" "Mozilla/5.0 (cor
207.46.19.49 - - [22/Feb/2012:00:06:06 +0100] "GET /index.php?...&lang=eu HTTP/1.1" 200 23380 "-" "Mozilla/5.0 (cor
73.224.15.77 - - [17/Sep/2012:00:00:00 +0200] "POST /administ...index.php HTTP/1.1" 301 261   "-" "Mozilla/5.0 (cor
13.4.215.228 - - [17/Sep/2012:10:21:58 +0200] "GET /templates/...logo.gif HTTP/1.1" 304 -     "-" "Mozilla/5.0 (cor
194.69.224.7 - - [18/Sep/2012:09:16:31 +0200] "GET /templates/....uery.js HTTP/1.1" 200 55774 "-" "Mozilla/4.0 (cor
194.69.224.7 - - [18/Sep/2012:09:16:33 +0200] "GET /images/....Button.gif HTTP/1.1" 200 368   "-" "Mozilla/4.0 (cor
194.69.224.7 - - [18/Sep/2012:09:16:33 +0200] "GET /templates/...logo.gif HTTP/1.1" 200 12530 "-" "Mozilla/4.0 (cor
194.69.224.7 - - [18/Sep/2012:09:16:33 +0200] "GET /templates/...ogo2.gif HTTP/1.1" 200 3451  "-" "Mozilla/4.0 (cor
194.69.224.7 - - [18/Sep/2012:09:16:33 +0200] "GET /templates/...piti.gif HTTP/1.1" 200 45    "-" "Mozilla/4.0 (cor
194.69.224.7 - - [18/Sep/2012:09:16:35 +0200] "GET /templates/...irun.png HTTP/1.1" 200 2450  "-" "Mozilla/4.0 (cor
```

Fig. 1. Sample lines of a log file in CLF.

Fig. 2. Appearance of the front page of discapnet website.

In this context, the next stage of our research consisted on analysing the structure and content of the site. Figure 2 shows the appearance of the front page of discapnet.

The site is divided in different areas being main ones:

- *Areas Temáticas*
- *Comunidad*
- *Actualidad*

Some of these parts, such as, *Actualidad* (actuality) and *Noticias* (news) are very dynamic and can hardly be used for link prediction because it is impossible to build the models according to news that will be generated in the future. From the rest of the zones in the website, the experts in Technosite considered that *Areas Temáticas* (excluding *Salud*) and *Canal Senior* within *Comunidad* were the most interesting zones for modelling and introducing adaptation tools. And, as a consequence the provided data was limited to these zones.

The direct consequence of the previous assertion is that the built user models and link prediction system will be mainly limited to *Areas Temáticas* (see Figure 3).

Therefore, it shouldn't be forgotten that the provided sequences might not be complete user sessions what limits the data mining process and, as a consequence, the quality of the obtained profiles.

Finally, before starting with the modelling process we evaluated the accessibility of each of the pages of discapnet. We found this an important starting point because we considered that being discapnet a website addressed to people with special requirements, accessibility of the pages might become a source of problems. Therefore, accessibility was evaluated using EvalAccess [12]; the Automatic Accessibility Evaluator developed by EGOKITUZ according to the design guidelines published by WAI [13] and devoted to help designers to produce web sites that are accessible. The study showed that the accessibility rate was in average near 90% and this means that each individual page in discapnet was designed taking into account the accessibility guidelines.

III. Data Preprocessing

After the preliminary analysis the logs must be preprocessed to extract the useful information. Web server log files follow a standard format called Common Log Format [11]. This standard specifies the fields all log files must have for each request received: remotehost, rfc931, authuser, date, request, status and bytes. The fields we used for this work are: the

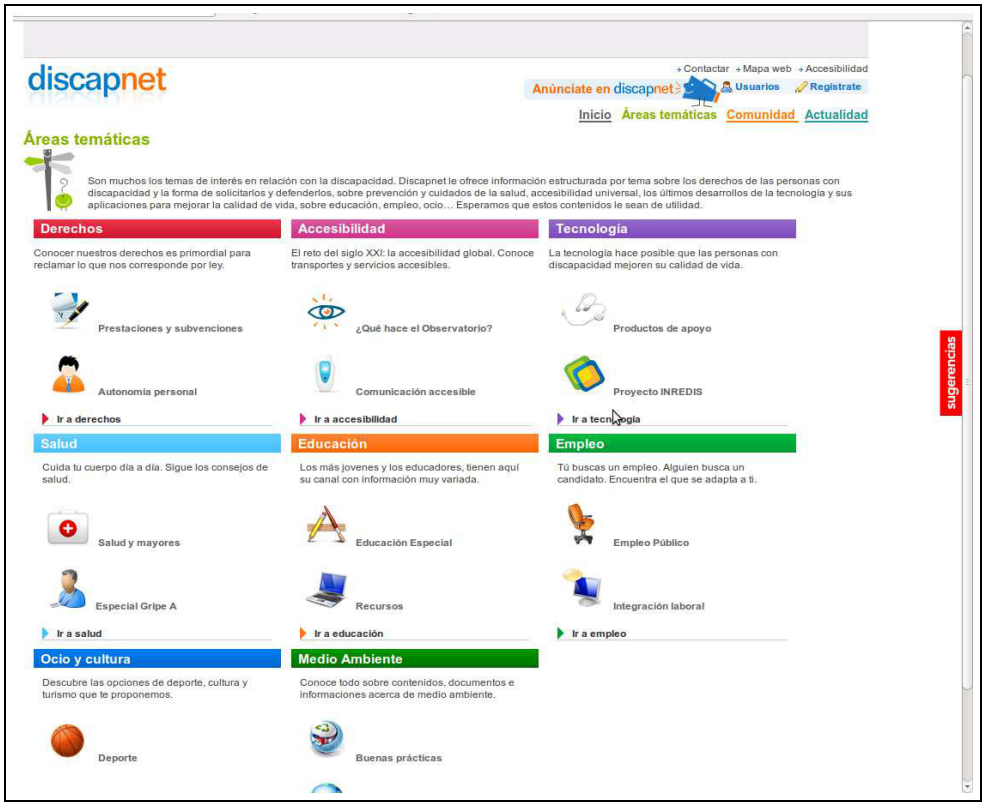


Fig. 3. Appearance of *Areas tematcias* within discapnet website.

code given by the anonymization process to remote host IP addresses, the time the request was recorded, the requested URL and the status field that informs about the success or failure when processing the request.

The log files used in this work contained 157,527,312 requests, which were reduced to 13,352,801 after the data preparation or data preprocessing phase described in the following lines. First of all we removed erroneous requests, those that had an erroneous status code (client error (4xx) and server error (5xx)). Therefore, we only took into account successfully processed requests. The next step consisted of selecting the requests directly related to the user activity. User clicks indirectly send many web browser requests to complete the requested web page with images, videos, style (css) or functionalities (scripts) for example. All these indirect requests were removed.

We then carried out session identification: we fixed the expiry time of each session to 10 minutes of inactivity [14] obtaining a total amount of 907,404 sessions.

A. Session Representation

Once the database to be used for the process has been selected in the filtering and preprocessing steps, we need to decide how to represent the information to be used in the machine learning algorithms. Being the aim of this work to detect sets of users with similar navigation patterns and to use them to make the navigation of future users easier, we represented the information corresponding to each of the sessions as a clickstream or sequence of clicks preformed in different URLs. Since we want to build user navigation profiles the order of the visited URLs will be important.

We selected the most relevant sessions (those with a minimum activity level; 3 or more clicks) and removed the longest sequences (those with their length out of the 98 percentile) with the assumption that long sequences are outliers and might be caused by some kind of robot, such as crawlers, spiders or web indexers.

IV. Pattern Discovery

This is the stage that, taking as input the user click sequences, is in charge of modelling users and producing user profiles. Most commercial tools perform statistical analysis over the collected data. They extract information about most frequently accessed pages, average view times, average lengths of paths, etc. that are generally useful for marketing purposes. But the knowledge extracted from this kind of analysis is very limited. Machine learning techniques are in general able to extract more knowledge from data. In this context unsupervised machine learning techniques have shown to be adequate to discover user profiles [6]. We have used a crisp clustering algorithm to group users that show similar navigation patterns.

A. Clustering

In this work, in order to grouping into the same segment users that show similar navigation patterns, we need a clustering algorithm that is able to deal with sequences and an adequate distance to compare sequences. Based in our experience in previous works, as clustering algorithm, we selected *PAM (Partitioning Around Medoids)* [15] [16] which is similar to k-means but uses medoids or examples as centres instead of centroids what makes it suitable for cases where the examples are represented as sequences. Furthermore, we selected a Sequence Alignment Method, Edit Distance [17][18], as a metric to compare sequences. As it happens with most clustering algorithms, *PAM* requires the K parameter to be estimated. This parameter is related to the specificity of the generated profiles, when greater its value is more specific the profiles will be. We didn't have prior knowledge of the structure of the data, that is, we have no idea of the number of different user profiles. Therefore we performed an analysis to try to find the value of K that is enough to group the sessions with common characteristics but does not force to group examples with not similar navigation patterns in the same cluster (the range of values tested for K will be described in Section VI).

B. Profile Generation

The outcome of the clustering process is a set of groups of user sessions that show similar behaviour. But we intend to model those users or to discover the associated navigation patterns or profiles for each one of the discovered groups. The model will be composed by the common click sequences appearing among the sessions in a cluster. We used SPADE (Sequential PAttern Discovery using Equivalence classes) [19], an efficient algorithm for mining frequent sequences, used to extract the most common click sequences of the cluster. SPADE uses combinatorial properties to decompose the original problem into smaller sub-problems, that can be independently solved in main-memory. All sequences are discovered in only three database scans.

In order to build the profiles of each cluster using SPADE we matched each user session with a SPADE sequence, with events containing a single user click. The application of SPADE provides for each cluster a set of URLs that are likely to be visited for the sessions belonging to it. The number of proposed URLs depends on parameters related to SPADE algorithm such as minimum support and maximum allowed number of sequences per cluster. We fixed the value for the minimum support to 0.2 and limited the amount of proposed URLs to 3 because proposing too many could disturb the user.

V. DISCAPNET USER NAVIGATION PROFILE DISCOVERY

We are aware that nowadays navigation in a website can be difficult for any type of users but this is still harder for users with special requirements. As a consequence, an adaptive system able to propose the adequate links to the user during her navigation would be specially helpful for them.

TABLE I
SIZES AND CHARACTERISTICS OF DATABASES USED FOR THE MODULAR SYSTEM AND THE GLOBAL SYSTEM.

Website zone	User Sessions	Average length	K
Accesibilidad	10,259	5.08	90
Derechos	22,561	4.78	80
Educación	1,773	4.27	27
Empleo	4,720	3.89	60
Medioambiente	852	5.05	20
Ocio y cultura	3,603	4.86	50
Tecnología	3,954	4.54	50
canal senior	338	4.60	13
Global	48,060	4.63	130

The structure of each of the subtopics within *Areas Temáticas* is very different and this will probably affect to the navigation the users do within them. Moreover a preliminary analysis of the sequences showed that nearly 50% of the user sessions extracted from the database belonged to navigations in a single zone. We considered those sessions representative of the navigation within each zone and decided to build the link prediction system based on them. After selecting the most relevant sessions and removing the longest sequences from it, the database contains 48,060 user sessions. We used two approaches to face the problem:

- The design of a global system using the data of all the analysed zones (see Figure 4).
- A modular system which builds profiles independently for each of the navigation zones (see Figure 5).

A. Global approach

The global system consists on applying the clustering and profiling processes as described in Section IV to the complete database; the 48,060 user sessions; the patterns are grouped using PAM clustering algorithm and the profile for each of the clusters discovered based on SPADE. The schema of the system is represented in Figure 4.

B. Modular approach

Being the structure of each zone different, we decided to build a modular approach to the user navigation profiling within discapnet. This means to build the profiles focusing on each of the possible analysis zones for user navigation profile discovery. With this aim, instead of working with the whole database, we worked with the user sessions located in a single web zone. That is we divided the database according to navigation zones and we worked with 8 different subsets; one for each of the zones where user navigation profile discovery will be carried out. Table I summarizes the sizes of each subset and Figure 5 shows the schema of the system where it can be observed that the profile discovery process within each zone was carried out as described in Section IV.

The set of profiles in the modular approach will be composed by the set of profiles generated for each one of the 8 zones.

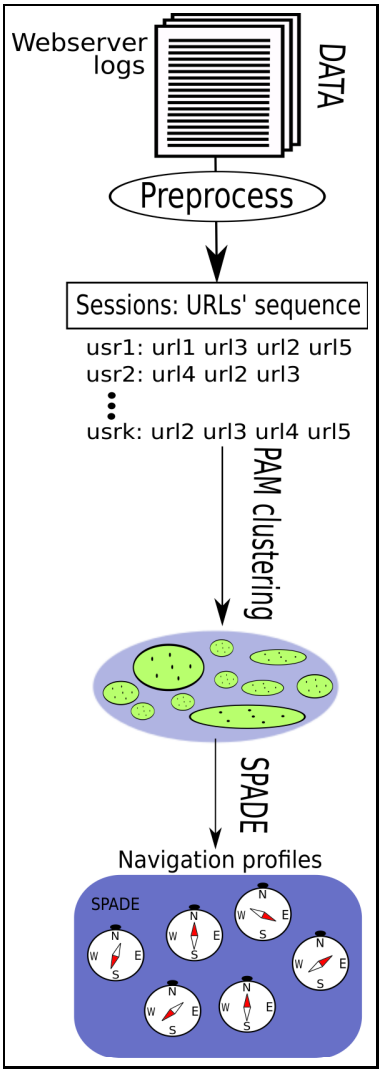


Fig. 4. Global approach to user profile discovery.

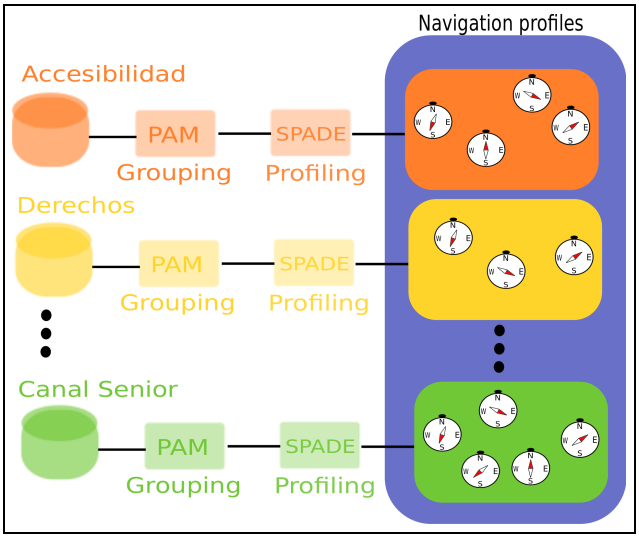


Fig. 5. Modular approach to user profile discovery.

VI. Evaluation and Link Prediction

Before the system is used in any real application the generated profiles need to be evaluated; i.e., we need to compare the generated profiles with the profiles of new users navigating the website and measure their similarity. We first generated user profiles by combining PAM with SPADE and compared these profiles to those for new users navigating the website. The evaluation procedure was exactly the same for the two approaches used to build the model: the modular approach and the global approach.

In order to carry out this evaluation we used a hold-out methodology, dividing each folder into a training set (70% of the examples), validation set (20% of the examples) and test set (10% of the examples). We used the validation set to select K (the number of clusters) and the test set to evaluate the performance of the system.

The internal structure of the data is completely unknown and we therefore tried a wide range of values for K to select the optimum number of clusters. Based on the usual exploration limit for the number of clusters, $\sqrt{n}$, in each of the databases we tried the following 5 values: $\sqrt{n}/4$, $\sqrt{n}/3$, $\sqrt{n}/2$, $2*\sqrt{n}/3$ and $\sqrt{n}$. Obviously, being the sizes of the subsets very different, the number of clusters or user profiles generated in each of them has been different; Table II shows in column K the number of profiles generated in each of the modules according to the evaluation carried out using the validation set. This number has been selected using the validation set to evaluate results the same way the test set has been used to evaluate the final system. The procedure is explained in the following lines (although the explanation is given for the test the procedure used with the validation set is exactly the same).

The generated profiles were evaluated by comparing them to new users navigating the website (test set). With this aim, the system needs to select a profile for the new users which will then be compared to their click sequence. The selection is done according to a distance calculation. This can be done at any stage of the navigation process; i.e. from the first click of the new user to more advanced navigation points. The hypothesis is that the navigation pattern of the user will be similar to the user profiles of its nearest clusters. As a result, the system will propose to the new user the set of links that models the users nearest clusters.

In order to simulate a real situation we need to take into account that when a user starts navigating only the first few clicks will be available to be used for deciding the corresponding profile. We simulated this real situation using 25% of the test sequences to select the profile for the new user according to the built model (between 1 and 2 links, because, as it is shown in column average length of Table I, the click sequences have in average near 5 links).

According to previous works found in the bibliography [20], new users might not be identical to any of the profiles

TABLE II
SUMMARY OF THE RESULTS.

		Validation			Test		
	k	Pr	Re	F.5	Pr	Re	F.5
Global system	130	0.55	0.40	0.51	0.55	0.40	0.51
Modular system	49	0.58	0.44	0.54	0.58	0.44	0.54
Improvement%		5.65	9.06	6.10	6.49	8.97	6.70

discovered in the training set; their profile might have similarities with more than one profile and, as a consequence, the diversification helps; it is better to build the profiles of the new users dynamically based on some of their nearest profiles. We propose the use of the k-Nearest Neighbour (k-*NN*) [21] supervised learning approach to calculate the distance from the click sequence (Edit Distance to the medoid) of the new users to the clusters generated in the previous phase. Due to its characteristics, the k-*NN* algorithm allows to select naturally the set of profiles generated in the training phase with higher similarity to the new user, and moreover, it showed to have good performance in previous works [20]. We used 2-*NN* to select the nearest clusters and combined the profiles of the two nearest clusters with defined profiles, weighting URL selection probabilities according to their distance. We combined these to propose profiles containing at most 3 URLs; those with the highest support values. If there are not enough URLs exceeding the minimum support value the profiles could have less than 3 URLs.

We computed performance metrics based on the results obtained for each of the new users of the test set. We compared the number of proposed links that are actually used in the test examples (hits) and the number of proposals that are not used (misses) and calculated precision (percentage of clicks used among the proposed ones), recall (percentage of clicks proposed among the used ones) and F.5-measure (a relationship between precision and recall giving more importance to precision).

The greater the number of URLs proposed as profiles the smaller will be the significance of some of them and the risk taken by the system will thus be greater. As a consequence, the values for precision will probably drop. Furthermore, by limiting the maximum number of URLs proposed for each profile to 3 the recall values will never reach 1. Since the average length of the sequences is near 5, if we propose a profile (3 URLs) based on 25% of the navigation sequence (between 1 and 2 URLs), we would be proposing less links than the really used ones, what makes impossible for the recall to achieve the highest values. We consider that in the concrete environment we are working it is really important to propose links that the user finds interesting because other proposed links would probably disturb the user. As a consequence, it is more important for the proposed links to be of good quality (precision) than guessing more of the used links (recall). This is why we used F.5-measure.

Table II shows the average results (precision, recall and F.5-measure) obtained for the test and validation sets in both cases: with the global system, and the modular system. The numbers show that the modular approach achieves better results than the global one, obtaining improvements of around 9% in recall and around 6.5% in precision and F.5-measure. Furthermore, results are similar for both, the validation set and the test set what means that the concrete data used to evaluate the system does not severely affect to the obtained performance.

The values obtained for the modular system show that if we would use the profiles for link prediction, nearly 60% (precision=0.58) of the proposed links (tending to 2 out of 3) would be among the ones used by the new user. This could probably make the user navigation easier.

Note that these results should be seen as lower bounds because, although not appearing in the user navigation sequence, the proposed links could be interesting and useful for them. Unfortunately, their usefulness/relevance could only be evaluated in a controlled experiment, by using user feedback.

Moreover, taking into account that the preliminary analysis showed that the time spent in hub pages is longer than usual we could assert that using those profiles for link prediction would save a big part of the time spent by users in their navigations.

The designed system seems to obtain near balanced values for precision and recall. Therefore analysing the recall we could state that nearly 45% of the links used by the new users would be among the ones proposed by the system (recall=0.44).

We need to take into account that this is a very strict evaluation of the models because, in a real situation, although not used during the navigation, some of the proposed links might also interest to the new users.

The main use of navigation profiles is link prediction and our system could be directly used for link prediction following the methodology described in the evaluation procedure.

VII. CONCLUSIONS

Web personalization becomes essential in industries and specially for the case of users with special needs such as visually impaired people. Adaptation may very much speed up the navigation of visually impaired people and contribute to diminish the existing technological gap. This work is the first stage of a web mining process carried out in discapnet: a website created to promote the social and work integration of people with disabilities. Based on observation in-use where behaviours emerge applying a web mining process to server log data, we designed a system to generate user navigation profiles and propose adaptations to the site through link prediction. The work was limited to the most static zones of the website.

We used *PAM (Partitioning Around Medoids)* clustering algorithm and Edit Distance to group into the same segment users with similar navigation patterns and SPADE (Sequential PAttern Discovery using Equivalence classes) to extract the user profiles from the cluster. These techniques were used to implemented two approaches: a global system built based on the complete website and a modular approach carried out discovering the navigation profiles within different zones of

the website. We then used a k-*NN (k-Nearest Neighbour)* based heuristic for link prediction.

Using a hold-out strategy and precision, recall and F.5-measure as performance measures for evaluation, we could deduce that both approaches showed to be effective for link prediction but the modular approach outperforms obtaining values of nearly 60% for precision and 45% for recall. This means that when 25% of the navigation of the new user has happened the designed system is able to propose a set of links where nearly 60% of them (2 out of 3) is among the ones the new user will be using in the future and this will definitely make the navigation easier saving a lot of time.

Being this a preliminary work, the system is open and many new ideas to be implemented in the future appeared during its development. First of all, the introduction of the designed link prediction system in the website and its evaluation in a real experiment would be the best way to discover the efficiency of the system. On the other hand, based on the web server log data provided by discapnet, other types of characteristics of the user sessions could be extracted which would allow to analyse the use of the website from another point of view mainly for problem detection.

ACKNOWLEDGMENT

This work was funded by the Department of Education, Universities and Research of the Basque Government (Eusko Jaurlaritza/Gobierno Vasco) through Grant IT-395-10, by the Science and Education Department of the Spanish Government (ModelAccess project,TIN2010-15549), by the Basque Governments SAIOTEK program (Datacc2 project, S-PE12UN064)

REFERENCES

[1] E. Turban and D. Gehrke, 2000. “Determinants of e-commerce website”. *Human Systems Management*, vol. 19, pp.111-120.

[2] D. Chaffey, F. Ellis-Chadwick, K. Johnston and R. Mayer, 2006. “Internet Marketing”. *Prentice Hall/Financial Times*.

[3] M. Vigo, S. Harper, 2013. “Challenging information foraging theory: screen reader users are not always driven by information scent”. *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pp.60-68, http://dx.doi.org/10.1145/2481492.2481499.

[4] J. Craven, P. Brophy, 2013. “Nonvisual access to the digital library: The use of digital library interfaces by blind and visually impaired people”. *In Technical report No. 145. Manchester, United Kingdom: Centre for Research in Library and Information Management*.

[5] S. Chandrashekar, T. Stockman, D. Fels, R. Benedyk, 2006. “Using Think Aloud Protocol with Blind Users: A Case for Inclusive Usability Evaluation Methods”. *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, pp.251-252, http://dx.doi.org/10.1145/1168987.1169040.

[6] D. Pierrakos, G. Paliouras, C. Papatheodorou and C.D. Spyropoulos, 2003 “Web usage mining as a tool for personalization: A survey”. *User Modeling and User-Adapted Interaction*, vol. 13, pp.311-372, http://dx.doi.org/10.1023/A:1026238916441 .

[7] P. Brusilovsky, A. Kobsa and W. Nejdl, 2007. “The Adaptive Web: Methods and Strategies of Web Personalization”. *Lecture Notes in Computer Science (Springer)*, Berlin, http://dx.doi.org/10.1007/978-3-540-72079-9 .

[8] E. García, C. Romero, S. Ventura and C.D. Castro, 2009.“An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering”. *User Modeling and User-Adapted Interaction*, vol. 19, pp.99-132, http://dx.doi.org/10.1007/s11257-008-9047-z.

[9] A. Apaolaza, S. Harper, C. Jay, 2013.“Understanding Users in the Wild”. *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, pp.1-4, http://dx.doi.org/10.1145/2461121.2461133.

[10] “Fundación ONCE for cooperation and social inclusion of people with disabilities”. Available: http://www.fundaciononce.es/EN/Pages/Portada.aspx, accessed 04-05-2014.

[11] “Common log format (clf)” 1995.*The World Wide Web Consortium (W3C)*. Available: http://www.w3.org/Daemon/User/Config/Logging.html, accessed 04-05-2014.

[12] “EvalAccess: Web Service tool for evaluating web accessibility”. Available: http://www.adm.aau.dk/rektor/aalborgeksperiment/engelsk/preface.html, accessed 04-05-2014.

[13] “WAI Guidelines and Techniques”. Available:http://www.w3.org/WAI/guid-tech.html, accessed 04-05-2014.

[14] D. He, A.Gker, 2000. “Detecting session boundaries from web user logs”. *In Proceedings of of the BCS-IRSG 22nd Annual Colloquium on Information Retrieval Research*, pp.57-66.

[15] L. Kaufman and P. J. Rousseeuw, 1990. “Finding Groups in Data: An Introduction to Cluster Analysis”. *Wiley-Interscience*, .

[16] L. Liu and M. T. Özsu, 2009. “Encyclopedia of Database Systems. In: PAM (Partitioning Around Medoids).“ *Springer US*, .

[17] D. Gusfield, 1997 “Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology”. *Cambridge University Press*, New York, NY, USA, .

[18] B. Chordia and K. Adhiya, 2011. “Grouping web access sequences using sequence alignment method”. *Indian Journal of Computer Science and Engineering (IJCSE)*, vol. 2, pp.308-314.

[19] M. J. Zaki, 2001. “Spade: An efficient algorithm for mining frequent sequences”. *Machine Learning*, vol. 42, pp.31-60, http://dx.doi.org/10.1023/A:1007652502315.

[20] O. Arbelaitz, I. Gurrutxaga, A. Lojo, J. Muguerza, J. Pérez and I. Perona, 2012. “Adaptation of the user navigation scheme using clustering and frequent pattern mining techniques for profiling”. *4th International Conference on Knowledge Discovery and Information Retrieval (KDIR)*, pp.187-192.

[21] S. Dasarathy, 1991. “Nearest neighbor (NN) norms : NN pattern classification techniques”. *IEEE Computer Society Press*, .

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 59–66
DOI: 10.15439/2014F482
ACSIS, Vol. 2

Using Fuzzy Logic and Q-Learning for Trust Modeling in Multi-agent Systems

Abdullah Aref
School of Electrical Engineering and Computer Science
Faculty of Engineering, University of Ottawa
Ottawa, Ontario, K1N 6N5, Canada

Thomas Tran
School of Electrical Engineering and Computer Science
Faculty of Engineering, University of Ottawa
Ottawa, Ontario, K1N 6N5, Canada

***Abstract*—Often in multi-agent systems, agents interact with other agents to fulfill their own goals. Trust is, therefore, considered essential to make such interactions effective. This work describes a trust model that augments fuzzy logic with Q-learning to help trust evaluating agents select beneficial trustees for interaction in uncertain, open, dynamic, and untrusted multi-agent systems. The performance of the proposed model is evaluated using simulation. The simulation results indicate that the proper augmentation of fuzzy subsystem to Q-learning can be useful for trust evaluating agents, and the resulting model can respond to dynamic changes in the environment.**

I. Introduction

A MULTI-AGENT Systems (MAS) involves multiple autonomous, self-interested, and goal-driven interacting intelligent agents [1]. An open MAS is a class of these systems in which agents can freely enter and leave at any time [2]. As each agent has only limited capabilities, it may need to rely on the services or resources from other agents in order to accomplish its goals [3]. Agents cannot assume that other agents share the same core beliefs about the system, or that other agents make accurate statements regarding their competencies and abilities. In addition, agents must accept the possibility that other agents may intentionally spread false information, or otherwise behaving in a harmful way, to achieve their own goals [1]. Therefore, agents should be equipped with a strong trust assessment model that is capable of maximizing the benefit, also referred to as utility gain (UG), of interacting with other agents. The estimation should be accurate enough that allows trust evaluating agents, also referred to as trustors (TRs), to identify the most beneficial trustee (TE) in their systems. The trust estimation model should consider all relevant factors, which affect the trust that an agent has about other agents. Failure to gather those factors would lead to compute a non-accurate trust value, which could explicitly affect agent's outcome [4]. Moreover, the model should dynamically update agents' belief sets to capture new characteristics of the environment, and should not rely on any centralized entities. Furthermore, the failure or takeover of any node must not lead to the failure of the whole system.

Trust has been defined in many ways in different domains [5]. For this work the definition used in [4] for trust in MASs, will be adapted. An agent's trustworthiness is considered as a measurement of the agent's possibility to do what it is supposed to do. In this work, we describe a trust model for MAS that combines the advantages of both: fuzzy logic and reinforcement learning for trust modeling in MAS. Moreover, we use a suspension technique in combination with reinforcement-learning to speedup the response of the model to dynamic changes in the system.

The paper is organized as follows: the related work is presented in section II followed by a general overview about fuzzy logic systems and reinforcement learning in section III. Section IV presents the details of the proposed model, while performance analysis is presented in section V. The last section presents conclusions and future work

II. Related Work

According to [3], most existing research on trust evaluation models can be divided into four main categories: direct trust evaluation models, that depends on past experience, indirect or reputation-based trust evaluation models, that depends on third-party testimonials from other agents in the same environment, socio-cognitive trust evaluation models, that depends on examining the social connections among agents to determine their trustworthiness, and organizational trust evaluation models, that depends on some organizational affiliations or endorsements issued by some trusted third-party to determine the trustworthiness of agents

FIRE [2] is a well-known decentralized trustworthiness estimation model for open MASs. The model categorizes trust components into direct experience called Interaction trust, Witness reputation, Role-based trust and Certified reputation. The model assumes that witnesses are honest and willing to cooperate and uses weighted summation to aggregate trust components.

Fuzzy logic offers the ability to handle uncertainty and imprecision effectively, and is therefore ideally suited to reasoning about trust [6]. Fuzzy inference copes with imprecise inputs and allows inference rules to be specified using imprecise linguistic terms, such as "very high" or "slightly low" [6].

FuzzyTrust [7] uses fuzzy logic inferences to estimate trust based on direct experience and witnesses testimonials taking into consideration uncertainties and incomplete information in a peer to peer system. The authors compare the performance of FuzzyTrust with the well known EigenTrust algorithm

978-83-60810-58-3/$25.00

[8], over the public domain transaction data from eBay, and demonstrated that it is more effective than EigenTrust.

A reinforcement learning (RL) based trustworthiness estimation model for buying and selling agents in an open, dynamic, uncertain and untrusted e-marketplace is described in [9] and further elaborated in [10], where buyers model the trustworthiness of the sellers as trustworthy, untrustworthy and neutral sellers. A buying agent chooses to purchase from a trustworthy seller. If no trustworthy seller is available, then a seller from the set of non-untrustworthy sellers is chosen. The seller's trustworthiness estimation is updated based on whether the seller meets the expected value for the demanded product with proper quality. A decentralized extension to the model used in [9] is describe in [11], [12], to enable indirect trustworthiness where advising agents are partitioned into trustworthy, untrustworthy and neutral sets to address buyers' subjectivity in opinions. However, the authors did not present any experimental results to justify their theoretical approach [13].

Recent survey such as [3][14] provide more insight on existing work in the field of MAS trust modeling.

III. Basic Concepts

A. Fuzzy Logic System (FLS)

FLSs have been extensively applied with success in many diverse application areas due to their similarity to human reasoning, and their simplicity [15]. An FLS provides a nonlinear mapping of input data vector into a scalar output. Such system maps crisp inputs into crisp outputs. It has four components: fuzzy logic rules, fuzzifier , an inference engine, and defuzzifier [16].

The main idea is that, the sets are based on the concept of a membership function (MFs), that defines the level to which a fuzzy variable is a member of a set. One represents full membership, whereas zero represents no membership; in other words, sets used for expressing input and output parameters are fuzzy [6]. An MF provides a measure of the level of similarity of an ingredient to the fuzzy subset. It is necessary to note that in fuzzy logic an ingredient can reside in more than one set to varying levels of association, which can't happen in crisp set theory. Triangular, trapezoidal, piecewise linear and Gaussian, are commonly used shapes for MFs [16].

Rules may be implemented by experts or can be derived from numerical data. In either case, fuzzy rules are represented as a collection of IF- THEN statements. MFs map input values into the interval [0,1] by the process known as "fuzzification" [6]. The fuzzifier maps crisp inputs into fuzzy sets, to stimulate rules which are in terms of linguistic variables. Fuzzy logic rules define the relationship between inputs and output. The inference engine, handles the way in which rules are combined. The conclusion membership levels are aggregated by superimposing the resultant membership curves. In many applications, crisp numbers must be collected at the output of an FLS. The defuzzifier maps output sets into crisp numbers[16]. Figure 1 presents the general architecture of an FLS.

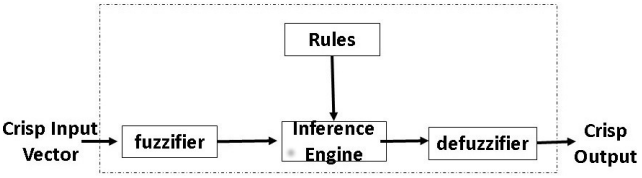


Fig. 1. Fuzzy Logic System [16]

During fuzzy inference, for each fuzzy rule, the inference engine determines the membership level for each input. Then measures the degree of relevance for each rule based on membership levels of inputs and the connectives (such as AND, OR) used with inputs in the rule. After that, the engine drives the output based on the calculated degree of relevance and the defined fuzzy set for the output variable in the rule[17].

Mamdani min-max method [18] is a well known direct inference method. where the degree of membership of rule conclusions is clipped at a level determined by the minimum of the maximum membership values of the intersections of the fuzzy value antecedent and input pairs. This ensures that the degree of membership in the inputs is reflected in the output [6]. In this work, Mamdani's method is used.

The centroid defuzzification method is an appealing defuzzification method [17]. The centroid method takes the center of gravity of the final fuzzy space in order to produce an output sensitive to all rules. In this work, the centroid defuzzification method is used.

B. Reinforcement learning

The reinforcement learning problem is the problem of learning from interaction to achieve a goal. In this problem, an agent observes a current state s of the environment, performs an action an on the environment, and receives a feedback r from the environment (reward, or reinforcement). The goal of the agent is to maximize the cumulative reward it receives in the end [10].

Temporal-difference (TD) learning algorithms can learn directly from experience without a model of the environment. TD algorithms do not require an accurate model of the environment and are incremental in a systematic sense [10]. One of the most widely used TD algorithms is known as the Q-learning algorithm. Q-learning works by learning an action-value function based on the interactions of an agent with the environment and the instantaneous reward it receives. For a state s, the Q-learning algorithm chooses an action a to perform such that the state-action value $Q(s, a)$ is maximized. If performing action a in state s produces a reward r and a transition to state s', then the corresponding state-action value $Q(s, a)$ is updated accordingly. State s is now replaced by s' and the process is repeated until reaching the terminal state [10]. The detailed mathematical foundation and formulation, as well as the core algorithm of Q-learning, can be found in [19] therefore it is not repeated here.

Q-learning is an attractive method of learning because of the simplicity of the computational demands per step and also

because of proof of convergence to a global optimum, avoiding all local optima, as long as the Markov Decision Process (MDP) requirement is met; that is the next state depends only on the current state and the taken action (it is worth noting that the MDP requirement applies to all RL methods) [15].

IV. Using Fuzzy Logic and Q-Learning for Trust Modeling in Multi-agent Systems

In this section, we propose the use of Fuzzy Logic and Q-Learning for Trust Modeling in Multi-agent Systems (FQT) as an improvement over RL based trust estimation by incorporating fuzzy subsystems to perform human-like decisions.

A. Overview

According to the proposed model, TRs classify TEs into three non-overlapping sets. The first set includes trustworthy TEs, the second set contains untrustworthy TEs and the third set includes neutral (neither trustworthy nor untrustworthy) TEs. Additionally, TRs classify witnesses in a similar way. If a TR is not satisfied by the interaction with a TE, the TR suspends the use of that TE for incoming transactions for a while. TRs suspend witnesses in a similar way.

TRs use Q-Learning to estimate the trustworthiness for TEs based on direct experience (DT). For those TEs that are not categorized as untrustworthy, the calculated DT is used as an input to the direct trust fuzzy subsystem, together with suspension period and the average of time-decayed utility gain within the last G interaction with the TE. The defuzzified output of this fuzzy subsystem is the fuzzy direct experience (FDT) of trustworthiness estimation.

For those TEs that are not categorized as untrustworthy, TRs consult witnesses for their testimonials about TEs. This is known as indirect trust (IT). Then information from both sources (direct experience and testimony of witnesses) are combined to compute total trust estimation (TT).

TRs request TEs to bid for coming transactions. The calculated TT is used as an input to the TE selection fuzzy subsystem (TSF). The second input is the difference between the bid value of the TE and the average bidding of all TEs for the same transaction. The third input is the difference between the average of time-decayed utility gain within the last H interaction with the TE and the average of time-decayed utility gain within the last H interaction with all TEs. The TR selects the TE that maximizes the outcome of TSF.

Figure 2 present the general architecture of the proposed model.

B. Fuzzy Direct Trustworthiness Estimation FDT (TR,TE):

In the proposed model, TRs use Q-learning to estimate the direct trust of TEs in a way similar to the process in [10]. If the TR is satisfied by the interaction with the TE, Eq. (1) is used to update the credibility of the TE as viewed by the TR.

$$DT_i(TR, TE) = DT_{i-1}(TR, TE)+\alpha(1-|DT_{i-1}(TR, TE)|) \quad (1)$$

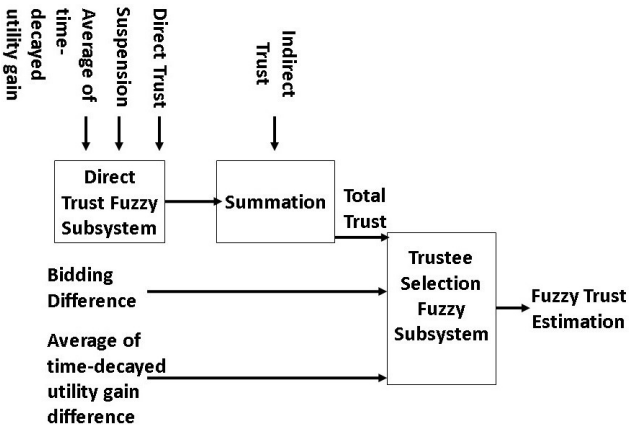


Fig. 2. Architecture of FQT

Here DT_i (TR, TE) is the direct trust estimation of the TE by the TR at time i. The value of DT (TR, TE) varies from -1 to 1. A TE is considered trustworthy if the trustworthiness estimation is above an honesty threshold (HT). The TE is considered untrustworthy if the trustworthiness estimation value falls below a fraudulent threshold (FT). TEs with trustworthiness estimation values between the two thresholds are considered neutral. The cooperation factor α is positive ($1>\alpha>0$) and the initial value of the direct trustworthy estimation is set to zero. TR will consider TE as being cooperative if the resulting UG of the transaction is greater than or equal the TR's satisfactory threshold.

If the TR is not satisfied by the interaction with the TE, Eq. (2) is used to update the credibility of the TE as viewed by TR.

$$DT_i(TR, TE) = DT_{i-1}(TR, TE)+\beta(1-|DT_{i-1}(TR, TE)|) \quad (2)$$

Here β is a negative factor called the non-cooperation factor ($0> \beta > -1$). TR will consider TE as being non-cooperative if the resulting UG of the transaction is less than the TR's satisfactory threshold. [10] described mathematical formulas to calculate the cooperation and non-cooperation factors in the context of an e-marketplace; however, we believe that those factors are application dependent and should be set by each agent independently. In general, we agree with [10] that the factors should be related to the value gain of the transaction.

Furthermore, the TR suspends the use of the TE for a period of time determined by equation (3)

$$SUS_i(TE) = SUS_{i-1}(TE) + BSI * IV \quad (3)$$

Where SUS_i(TE) is the suspension penalty associated with TE at time instant i. Basic Suspension Interval (BSI) is application dependent, it could be days in e-marketplace or seconds in a robotics system that has a short life time and Interaction Value (IV) indicates how much the TR values the interaction, not the actual utility gain of the interaction.

The value $SUS_i(TE)$ decreases with time if ther is no dissatisfactory transaction. That is $SUS_i(TE) = SUS_{i-1}(TE) - 1$, but can't be less than 0.Therefore, a large value of $SUS_i(TE)$ for a TE means a recently misbehaved TE

In the proposed model, each trust evaluation agent uses a direct trust fuzzy subsystem to find the Fuzzy Direct Trustworthiness Estimation (FDT). We define three input parameters for the fuzzy engine of the trust model and one output parameter; the input parameters are:

- The calculated DT using Q-Learning, calculated in equations 1 and 2. This parameter represents the long term relationship between the TR and the TE. A TE with a large value of DT means a TE that used to be cooperative for a relatively large number of transactions
- The suspension period of the TE, calculated in equation 3. This parameter is used to address the short term relationship between the TR and the TE. It helps the TR to address a recently malfunctioning TE that used to be honest for a relatively large number of transactions.
- AUG'$_t$The average of time-decayed utility gain within the last G interactions with the TE, calculated in equation 4. UG is the net benefit that the TR achieves from the transaction. Time decaying is used to to emphasize that UG from recent transaction weigh more compared to UG from old transactions if they have the same absolute value.

$$AUG'_t = \frac{\sum_{j=1}^{G} e^{-\lambda \Delta T_j} UG_j}{G} \quad (4)$$

Here ΔT_j = Current Time - Time of transaction j, λ is the decaying factor, $UG_j$is the utility gain for transaction j with the TE being evaluated, and G is the size of the historical window considered for calculating AUG'$_t$

The input parameters should be fuzzified before being used in the engine. We define the FDT as the defuzzified output. The individual "if…then" rules for driving the FDT is of the kind "if DT is HIGH and the SUS is LOW, and the AUG'$_t$ is HIGH then the FDT is VERY HIGH". This rule intuitively states that if the estimated direct trust is high and the suspension period is low, and the average utility gained by interacting with this TE is HIGH then the TE is expected to be honest and the transaction result is expected to be very high based on local experience of TR.

In the proposed model, we use the rules presented in Table I. Since each of the input parameters can be categorized as being Low (L), Medium (M), and High (H) and the output parameter can be categorized as being Very Low (VL), Low (L), Medium (M), Very High (VH), and High (H). We use a Mamdani min-max approach of inference and the centroid technique for the defuzzification

In table I, we insisted that a recently suspended TE will have a low direct trust value. The idea is that a TR will stop interacting with a misbehaving TE immediately, and wait utile it is clear whether this misbehaviour is accidental or it is a behavioural change. Because suspension is temporary, and because TR uses information from witnesses, the effect of accidental misbehavior will phase out, but the effect of a behaviour change will be magnified

TABLE I
DIRECT TRUST FUZZY SUBSYSTEM RULES

Rule	DT	Suspension	AUG'	Output
1	L	L	L	L
2	M	L	L	L
3	H	L	L	M
4	L	M	L	L
5	M	M	L	L
6	H	M	L	M
7	L	H	L	VL
8	M	H	L	VL
9	H	H	L	VL
10	L	L	M	L
11	M	L	M	L
12	H	L	M	H
13	L	M	M	L
14	M	M	M	L
15	H	M	M	H
16	L	H	M	VL
17	M	H	M	VL
18	H	H	M	VL
19	L	L	H	L
20	M	L	H	M
21	H	L	H	VH
22	L	M	H	L
23	M	M	H	L
24	H	M	H	H
25	L	H	H	VL
26	M	H	H	VL
27	H	H	H	VL

C. Indirect Trustworthiness Estimation IT(TR, TE):

To estimate indirect trust, a TR consults other witnesses who interacted previously with the TE. To reduce the effect of fraudulent witnesses, a TR excludes reports from any witness where the mean of the differences between the witness's trustworthiness estimation and the TR's trustworthiness estimation of TEs other than the one under consideration is above the witnesses differences threshold (WDT). An honest witness (WT) reports its testimony (RT) about a TE as

$$RT(WT,\, TE) = FDT(WT,\, TE) \quad (5)$$

where FDT (WT, TE) is the WT fuzzy direct experience of trustworthiness estimation of the TE.

A TR will calculate the indirect trust (IT) component as

$$IT(TR,\, TE) = \frac{\sum_{k=1}^{N} weight_k * RT(WT_k,\, TE)}{N} \quad (6)$$

where N is the number of consulted witnesses. RT (WT_k, TE) is the testimony of witness k about TE, and $weight_k$ is the weight assigned by the TR to testimony of WT_k. The calculation of the weight factor, or the adaptation of a calculation technique from the literature, is considered a future work.

TRs track the credibility of their witnesses. Each TR updates its rating for the witnesses after each interaction as follows

- If the transaction was satisfactory for the TR and the witness WT had recommended TE or

- If the transaction was NOT satisfactory and WT's opinion was "not recommend".

Then the trustworthiness estimation of WT is incremented as in equation 7

$$DT(TR, WT) = DT(TR, WT) + \gamma(1 - |DT(TR, WT)|) \quad (7)$$

- Otherwise, the trustworthiness estimation of WT is decremented as in equation 7

$$DT(TR, WT) = DT(TR, WT) + \zeta(1 - |DT(TR, WT)|) \quad (8)$$

where γ and ζ are positive and negative factors respectively and chosen by the TR as cooperation and noncooperation factors. The value of DT (TR, WT) varies from -1 to 1. A witness is considered trustworthy if the trustworthiness estimation is above the witnesses' honesty threshold (WHT). A witness is considered untrustworthy if the trustworthiness estimation falls below the witnesses' fraudulence threshold (WFT). Witnesses with trustworthiness estimation values in between the two thresholds are considered neutral.

When a TR wants to interact with a TE at instant i, the TR avoids any WT that is untrustworthy.

D. Total Trustworthiness Estimation TT(TR,TE):

The proposed trust model takes into consideration TRs' direct trust of TE(s), testimonials of witnesses, and credibility of witnesses. Therefore, the total trust estimate can be calculated using Eq. (9)

$$TT(TR, TE) = x*FDT(TR, TE)+(1-x)*IT(TR, TE) \quad (9)$$

Here FDT(TR,TE) is the fuzzy direct experience estimation component of the TR for the TE, IT(TR,TE) is the indirect trust estimation component of the TR for the TE and x is a positive factor, chosen by the TR, which determines the weight of each component in the model.

E. Trustee Selection

In the proposed model, TRs request TEs to bid for the coming interactioneach. Each TR uses a fuzzy engine to select a profitable TE. We define three input parameters for the fuzzy engine of the trust model and one output parameter; the input parameters are

- The total trust estimation (TT), as calculated by combining information from direct experience and testimony of witnesses, detail calculations described later in this section.
- Bidding Difference (BD): The difference between the promised UG, i.e. bidding value, of the TE (B_t) and the average bidding values of all TEs bidding for the same transaction.This parameter is used to differentiate a TE that promises high UG while the average promise is relatively low, from one that promises high UG while almost every TE promises high UG. In both cases, the TE promises high UG. but this value in more important in the first case compared to the second case

$$BD = B_t - \frac{\sum_{l=1}^{M} B_l}{M} \quad (10)$$

- Average UG Difference (DAUG'$_t$): The difference between the average of time-decayed UG within the last H interactions with the TE and the average of time-decayed utility gain within the last H interactions with all TEs. Time decaying is used to emphasize that recent transaction weighs more compared to old transactions if they have the same value of UG.

$$DAUG'_t = \frac{\sum_{p=1}^{H} e^{-\lambda\Delta T} UG_p}{H} - \frac{\sum_{q=1}^{H} e^{-\lambda\Delta T} \bar{UG}_q}{H} \quad (11)$$

Here ΔT = Current Time - Time of the transaction. UG$_p$is the utility gain for transaction p with the TE being evaluated, $\bar{UG}_q$ is the utility gain for transaction q, regardless of the TE, and H is the size of the historical window considered for calculating AUG'$_t$

The input parameters should be fuzzified before being used in the engine. We define the fuzzy estimated utility gain (FUG) as the defuzzified output parameter. The individual "if... then" rules for driving the fuzzy estimated utility gain FUG is of the kind "if the estimated total trust is HIGH and the DAUG'$_t$ is HIGH and the BD is HIGH then the Fuzzy UG is VERY HIGH". This rule intuitively states that if the estimated trust is high and utility gained by interacting with this TE is higher than the overall average utility gain, and the TE is promising higher utility gain compared to other bidding TEs, then the TE is expected to be honest and the transaction result is expected to be very high.

In the proposed model, we use the rules presented in Table II for TSF. Since each of the input parameters can be categorized as being Low (L), Medium (M), and High (H) and the output parameter can be categorized as being Very Low (VL), Low (L), Medium (M), Very High (VH), and High (H). Here, again, we use a Mamdani min-max approach to inference and the centroid technique for the defuzzification.

TR evaluates the trustworthiness of TEs that are not untrustworthy. TEs whose trustworthiness cannot be determined (due to no available rating) are placed in the Unknown Trust (UT) set. Those, whose trustworthiness has been determined, are placed in the Known Trust (KT) set. On one side, selecting a TE from the set KT is likely to give a more predictable value for the expected UG. However, the TR has not learnt enough about the TE population, therefore, it may get a non-optimal performance. On the other side, selecting a TE from the set UT allows TR to explore more about the TE population, although it may risk losing utility if it encounters a bad TE [2]. To encourage honest bidding when selecting a TE from UT set, a random TE with the second highest bidding value is selected.

Obviously, if one of the two sets is empty, TR can only select from the other set. Otherwise, it needs to determine

TABLE II
TRUSTEE SELECTION RULES

Rule	TT	BD	DAUG'$_t$	Output
1	L	L	L	VL
2	M	L	L	L
3	H	L	L	M
4	L	M	L	L
5	M	M	L	M
6	H	M	L	M
7	L	H	L	L
8	M	H	L	M
9	H	H	L	M
10	L	L	M	L
11	M	L	M	M
12	H	L	M	M
13	L	M	M	L
14	M	M	M	M
15	H	M	M	M
16	L	H	M	M
17	M	H	M	H
18	H	H	M	H
19	L	L	H	L
20	M	L	H	M
21	H	L	H	H
22	L	M	H	L
23	M	M	H	M
24	H	M	H	H
25	L	H	H	M
26	M	H	H	H
27	H	H	H	VH

TABLE III
VALUES OF USED PARAMETERS

Parameter	Value
Total number of Trustees	10
Total Number of trustors	100
Number of Good Trustees	2
Number of Bad Trustees	3
Number of Ordinary Trustees	3
Number of Intermittent Trustees	2
Number of Categories for Trustees	4
Maximum utility gain	10
trustee cooperation factor	0.1
trustee non-cooperation factor	-0.3
Witnesses cooperation factor	0.1
Witnesses non-cooperation factor	-0.3
Direct trust fraction	0.5
Degree of decay	0.1
Trustees' honesty threshold	0.5
Trustees' fraudulent threshold	-0.5
Witnesses' honesty threshold	0.5
Witnesses' fraudulence threshold	-0.5
Trustor satisfactory threshold	0
Witnesses differences threshold	0.5

which action it should take. The exploit-vs explore dilemma can be addressed by using the Boltzmann exploration strategy [20]. Using this strategy, an agent tends to explore its environment first and then gradually move towards exploitation when it learns more about the environment. When exploiting, TR selects the TE with the highest FUG.

V. PERFORMANCE EVALUATION

It is often difficult to find suitable real world data set for comprehensive evaluation of trust models, since the effectiveness of various trust models needs to be assessed under different environmental conditions and misbehaviors [3]. Therefore, in trust modeling for MASs research field, most of the existing trust models are assessed using simulation or synthetic data [3]. One of the most popular simulation test-beds for trust models is the agent reputation and trust (ART) test-bed proposed in [21]. However, even this test-bed does not claim to be able to simulate all experimental conditions of interest. For this reason, many researchers design their own simulation environments when assessing the performance of their proposed trust models [3].

A. Simulation Environment

We use simulation to evaluate the performance of the proposed model for distributed, multi-agent environment using the discrete-event multi-agent simulation toolkit MASON [22] with TEs, that provide services, and TRs, that consume services. For the Fuzzy subsystems, we used the jFuzzyLogic Java package [23]. As with [2], we assume that the performance of a TE in a particular service is independent from that in another service. Therefore, without loss of generality, and in order to reduce the complexity of the simulation environment, it is assumed that there is only one type of service in the system simulated and all TEs offer the same service with, possibly, different performance. In order to study the performance of the proposed trust model for TE selection, we compare the proposed model with the well known FIRE trust model [2]. Each simulation experiment is repeated 10 times with different seed values for the random number generators, and the average of the 10 experiments is presented as the simulation result. Network communication effects are not considered in this simulation. Each agent can reach each other agent. The simulation step is used as the time value for interactions. Transactions that take place in the same simulation step are considered simultaneous. Locating TEs and witnesses are not part of the proposed model; therefore, TRs locate TEs and witnesses through the system. TRs evaluate the trustworthiness of the TE(s), and then selects one to interact with.

Having selected a TE, the TR then interact with the selected TE and gains some utility from the transaction (UG). The value of UG is in [−10, 10] and depends on the level of performance of the TE in that transaction. A TE can serve many users at a time. A TR does not always use the service in every round. The probability it needs and requests the service, called its activity level, is selected randomly when the agent is created.

After each transaction, the TR updates the credibility of the TE participated in the transaction. It is assumed that TEs may be selfish, liars, non cooperative or simply malfunctioning. In order to compare our work with FIRE [HuynhJS2006], honest witnesses assumed.

TEs can be in one of four types: good, ordinary, bad, and intermittent. Each of them, except the last, has a mean level of performance. The actual performance follows a normal distribution around this mean which is in the reange of (5,10] for good TEs, [0, 5] for ordinary TEs and [-10,0) for bad

TABLE IV
DIRECT TRUST FUZZY SUBSYSTEM INPUT AND OUTPUT MFS

	DT	Suspension	AUG'	Output
VL	- -	- -	- -	(PWL) -1.0, -0.5, -0.3
L	(PWL) -1.0,-0.5,0.01	(PWL) 0.0,0.25,0.5	(PWL) -10, -0.1, 1.0	(TR) -0.35, -0.1, 0.1
M	(TR) 0.0, 0.2, 0.5	(TR) 0.4, 0.7, 1.0	(TR) 0, 2.5, 5	(TR) 0.0, 0.2, 0.4
H	(PWL) 0.4, 0.7, 1	(PWL) 0.9, 1, 999	(PWL) 4, 6, 10	(PWL) 0.3, 0.5, 0.6
VH	- -	- -	- -	(PWL) 0.5, 0.7, 1.0

legend PWL: piece-wise linear TR:triangular

TABLE V
TRUSTEE SELECTION INPUT AND OUTPUT MFS

	TT	BD	DAUG'$_t$	Output
VL	- -	- -	- -	(PWL) 0.0, 0.25, 0.4
L	(PWL) -1.0,-0.1,0.1	(PWL) -20.0, -0.1, 0.1	(PWL) -20, -0.1, 0.1	(TR) 0.3, 0.5, 0.9
M	(TR) 0.0, 0.2, 0.6	(TR) 0.0, 1.0, 3.0	(TR) -0.1, 0.5, 2	(TR) 0.8, 1.0, 1.2
H	(PWL) 0.5, 0.7, 1	(PWL) 2, 4, 20	(PWL) 1.5, 2.5, 20	(PWL) 1.1, 1.4, 1.6
VH	- -	- -	- -	(PWL) 1.5, 1.6, 2.0

legend PWL: piece-wise linear TR:triangular

TEs. Intermittent trstees, on the other hand, yield (random) performance levels in the range [-10, 10] and they can result in positive UG some times and negative UG other times.

Since agents are owned and controlled by various stake-holders, the performance of an agent may not be consistent over time. A TE may change its behavior. In this simulation study, the performance of a TE can be changed by a randomly selected amount with a probability selected randomly when the agent is created. When bidding, an honest (good or ordinary) TE bids its utility gain value, this value is considered the value of the transaction with the corresponding TR. A bad (unhonest) TE bids a positive value for its utility gain, but the utility gain that the corresponding TR can get is the true utility gain that the bad TE can afford (negative value).

Table III presents the number of agents, and other parameters used in the proposed model and those used in the environment.

The membership functions for the input, output parameters used for direct trust fuzzy subsystem in our evaluation are summarized in Table IV, the membership functions for the input, and output parameters used for the TE selection fuzzy subsystem in our evaluation are summarized in Table V.

B. Experimental results

1) Performance in a static environment: The first thing to test is whether the proposed model helps TRs select profitable TEs (i.e. those yielding positive UG) from the population and, by so doing, helps them gain better utility than when using FIRE trust model. In this section, we use a static environment, which means that each TR attempt to make a transaction each step, and witnesses and TEs do not change their honesty levels.

Figure 3 describes the average UG per transaction as the number of transactions increases from 5 to 50 in the static environment. The charted UG is calculated as the averaged value for 10 different runs of the experiment. For each run, the summation of UG that all TRs accumulated at the end of each fifth simulation step is divided by the number of TRs (note that in the static environment, each TR interact in each simulation step.). The figure shows that selecting providers using the proposed model perform closely to FIRE despite the fact that FIRE make use of rule-based trust that can't be assumed to be available all the time. Moreover, the performance of both

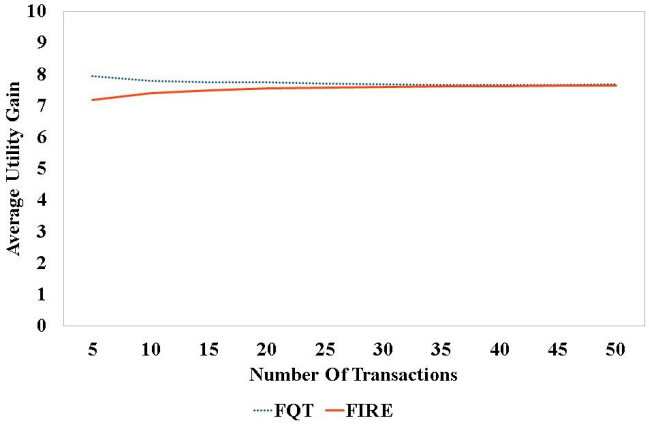


Fig. 3. Performance in Static Environment

models stabilize after a while. For FIRE, this is consistent with the results obtained in [2]. This stabilization in the performance of the two models indicates that they both learned to interact with the most beneficial TEs in the system

2) Performance in a dynamic environment: A trust model designed for MAS should be able to function properly in a dynamic environment. In this section we test the performance of the proposed model in a changing environment, as described below. As with static environment, we compare the performance of the proposed model with the case of using FIRE.

Specifically, the same experiments will be run, but with each of the following conditions: each TE may alter its average level of performance at maximum 1.0 UG unit with a probability of 0.10 each simulation step. A TR uses the service with probability in the range [0.25 - 1.0], intermittent TEs flip their honesty randomly and TEs may leave the system and new TEs may join the system with probability 0.5

Figure 4 describes the average UG per transaction as the number of transactions increases from 5 to 50 in the dynamic environment. The charted average UG is calculated as the averaged value for 10 different runs of the experiment. For each run, the summation of UG that all TRs accumulated when the total number of transaction in the system equals a multiple of five of the number of TRs is divided by the number of TRs. This value is averaged for 10 different runs of the experiment.

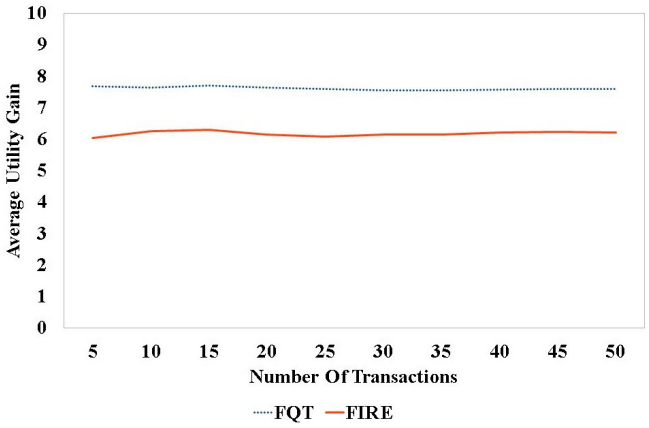


Fig. 4. Performance in Dynamic Environment

The figure shows that selecting TEs using the proposed model performs consistently better than using FIRE in terms of UG in a dynamic environment, which indicates that the proposed model responds better to dynamic changes compared to FIRE. Moreover, the performance of both models stabilize after a while. For FIRE, this is consistent with the results obtained in [2]. However, FIRE is not able to respond to the dynamics of the system as fast as the proposed model. This is due to the use of the fuzzy subsystems and due to picking the TE with the second highest bid.

VI. Conclusion and Future Work

In this paper, we presented a trust model for MASs that combines the use of Q-learning to estimate trustworthiness and incorporate two fuzzy subsystems for TE selection to enhance the utility gain estimation. The presented model allows direct and indirect sources of trust information to be integrated to provide a collective trust estimation. In addition, the proposed model incorporates fuzzy subsystems to account for suspension periods, average utility gain, bidding differences, and the relative average utility gain of a TE compared to the overall utility gain. The proposed model has been simulated using MASON with the use of the jFuzzyLogic package. The results indicate that the model can help TRs enhance their utility gain and that the proposed model can respond better to dynamic changes in the environment. In short, we believe the proposed model can provide a trust measure that is sufficiently useful to be used in MASs

Dynamically determining parameter values for the fuzzy subsystems, enabling TEs to actively promote their honesty, bootstrapping trust for new TEs and using Q-learning to dynamically select the proper action in each rule of the fuzzy subsystem are considered as future work.

References

[1] C. Burnett, T. J. Norman, and K. Sycara, "Trust decision-making in multi-agent systems," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume One*, ser. IJCAI'11. AAAI Press, 2011, pp. 115–120.

[2] T. D. Huynh, N. R. Jennings, and N. R. Shadbolt, "An integrated trust and reputation model for open multi-agent systems," *Autonomous Agents and Multi-Agent Systems*, vol. 13, no. 2, pp. 119–154, Sep. 2006.

[3] H. Yu, Z. Shen, C. Leung, C. Miao, and V. Lesser, "A survey of multi-agent trust management systems," *Access, IEEE*, vol. 1, pp. 35–50, 2013.

[4] B. Khosravifar, J. Bentahar, M. Gomrokchi, and R. Alam, "Crm: An efficient trust and reputation model for agent computing," *Know.-Based Syst.*, vol. 30, pp. 1–16, Jun. 2012.

[5] S. D. Ramchurn, D. Huynh, and N. R. Jennings, "Trust in multi-agent systems," *Knowl. Eng. Rev.*, vol. 19, no. 1, pp. 1–25, Mar. 2004.

[6] N. Griffiths, K.-M. Chao, and M. Younas, "Fuzzy trust for peer-to-peer systems," in *Distributed Computing Systems Workshops, 2006. ICDCS Workshops 2006. 26th IEEE International Conference on*, July 2006, pp. 73–73.

[7] S. Song, K. Hwang, R. Zhou, and Y.-K. Kwok, "Trusted p2p transactions with fuzzy reputation aggregation," *Internet Computing, IEEE*, vol. 9, no. 6, pp. 24–34, Nov 2005.

[8] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina, "The eigentrust algorithm for reputation management in p2p networks," in *Proceedings of the 12th International Conference on World Wide Web*, ser. WWW '03. New York, NY, USA: ACM, 2003, pp. 640–651.

[9] T. Tran and R. Cohen, "Improving user satisfaction in agent-based electronic marketplaces by reputation modelling and adjustable product quality," in *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 2*, ser. AAMAS '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 828–835.

[10] T. Tran, "Protecting buying agents in e-marketplaces by direct experience trust modelling," *Knowledge and Information Systems*, vol. 22, no. 1, pp. 65–100, 2010.

[11] K. Regan and R. Cohen, "Indirect reputation assessment for adaptive buying agents in electronic markets," *Business Agents and the Semantic Web workshop*, vol. 1, 2005.

[12] K. Regan, R. Cohen, and T. Tran, "Sharing models of sellers amongst buying agents in electronic marketplaces," *Decentralized Agent Based and Social Approaches to User Modelling workshop*, vol. 1, 2005.

[13] S. Beldona, "Reputation based buyer strategies for seller selection in electronic markets," Ph.D. dissertation, Electrical Engineering & Computer Science, University of Kansas, 2008.

[14] I. Pinyol and J. Sabater-Mir, "Computational trust and reputation models for open multi-agent systems: A review," *Artif. Intell. Rev.*, vol. 40, no. 1, pp. 1–25, Jun. 2013.

[15] S. Georgoulas, K. Moessner, A. Mansour, M. Pissarides, and P. Spapis, "A fuzzy reinforcement learning approach for pre-congestion notification based admission control," in *Proceedings of the 6th IFIP WG 6.6 International Autonomous Infrastructure, Management, and Security Conference on Dependable Networks and Services*, ser. AIMS'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 26–37.

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

[17] C. Pappis and C. Siettos, "Fuzzy reasoning," in *Search Methodologies*, E. Burke and G. Kendall, Eds. Springer US, 2005, pp. 437–474.

[18] E. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," *International Journal of Man-Machine Studies*, vol. 7, no. 1, pp. 1 – 13, 1975.

[19] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 1998.

[20] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *J. Artif. Int. Res.*, vol. 4, no. 1, pp. 237–285, May 1996.

[21] K. K. Fullam, T. B. Klos, G. Muller, J. Sabater, A. Schlosser, Z. Topol, K. S. Barber, J. S. Rosenschein, L. Vercouter, and M. Voss, "A specification of the agent reputation and trust (art) testbed: Experimentation and competition for trust in agent societies," in *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, ser. AAMAS '05. New York, NY, USA: ACM, 2005, pp. 512–518.

[22] S. Luke, C. Cioffi-Revilla, L. Panait, K. Sullivan, and G. Balan, "Mason: A multiagent simulation environment," *Simulation*, vol. 81, no. 7, pp. 517–527, Jul. 2005.

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

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 67–74
DOI: 10.15439/2014F256
ACSIS, Vol. 2

Minimizing Size of Decision Trees for Multi-label Decision Tables

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

***Abstract*—We used decision tree as a model to discover the knowledge from multi-label decision tables where each row has a set of decisions attached to it and our goal is to find out one arbitrary decision from the set of decisions attached to a row. The size of the decision tree can be small as well as very large. We study here different greedy as well as dynamic programming algorithms to minimize the size of the decision trees. When we compare the optimal result from dynamic programming algorithm, we found some greedy algorithms produce results which are close to the optimal result for the minimization of number of nodes (at most 18.92% difference), number of nonterminal nodes (at most 20.76% difference), and number of terminal nodes (at most 18.71% difference).**

I. Introduction

NOW a days, multi-label decision tables have gained attention in problem of semantic annotation of images [1], music categorization into emotions [2], functional genomics [3], and text categorization [4]. Furthermore, it is natural to have such data sets in optimization problems such as finding a Hamiltonian circuit with the minimum length in the traveling salesman problem [5], finding nearest post office in the post office problem [5]; in this case we give input with more than one optimal solutions.

In multi-label decision tables, each row is labeled with a set of decisions. It is common to have such tables in our real life because we do not have enough number of attributes of the domain to separate rows. Thus we have objects with equal values of conditional attributes but with different decisions. In literature, often, decision trees and other classifiers for multi-label data are considered for prediction (multi-label classification problem) [6], [7], [8], [9]. However, in this paper our aim is to study decision trees for multi-label decision tables for knowledge representation, and as algorithms for problem solving.

In [10] we studied a greedy algorithm for construction of decision trees for multi-label decision tables using the heuristic based on the number of boundary subtables. Besides, in [11] we have studied this algorithm in the cases of most common decision, and generalized decision approaches (in that paper, we considered decision tables with one-valued decisions as multi-label decision tables where sets of decisions attached to rows have one element).

This paper is a continuation of the conference publication [11]. We have introduced new greedy heuristics 'misclassification error', 'absent', 'combined' whose performances are as good as the previous one. Also we adapt, 'multi-label entropy', 'sorted entropy' heuristics from literature. We compared the performance among themselves for the cost function of number of nodes, number of nonterminal nodes and number of terminal nodes. We have done experiments using modified data sets from UCI Machine Learning Repository [12]. Based on the results of the experiments, we have presented rankings among the algorithms in the form of critical difference diagram [13]. Furthermore, we have shown the average relative difference between greedy and optimal results to describe how close they are. Hence, our goal is to choose some of the greedy heuristics which are close to the optimal results.

This paper consists of six sections. Section II contains the related background study of this problem. Section III contains the important definitions related to our study. After that, in Sect. IV, we presented the dynamic and greedy algorithms for construction of decision trees, then in Sect. V we gave comparison among algorithms using Friedman test. Section VI contains results of experiments and Sect. VII concludes the paper.

II. Related Work

In literature, often, problems that are connected with multi-label data are considered for classification: multi-label learning [14], multi-instance learning [9] etc. In multi-label learning, the output for each instances can be a set of decisions, whereas in our framework, we chose only one decision as output for each instance. In multi-instance learning, bag of instances are labeled rather than individual example which is far away from our problem. There is also semi-supervised learning [15] where some examples are labeled but some are not labeled, but we deal with examples that are labeled with multiple decisions.

Furthermore, some learning problems deal with many-valued data sets in different ways such as partial learning [16], ambiguous learning [17], and multiple label learning [18]. These problems consider only one label as correct and others as incorrect, but we consider all labels that are attached to an object as correct labels for that object. In [16], [18], the authors showed probabilistic methods to solve the learning problem whereas in [17], the author used standard heuristic approach to exploit inductive bias to disambiguate label information.

978-83-60810-58-3/$25.00

Additionally, these papers only focus on classification results rather than optimization of data model.

In this paper, we consider the problem of knowledge representation and optimization of data model. Therefore, our goal is to choose a data model which will be optimized and will give us one arbitrary decision from the set of decision attached with each row.

III. Main Definitions

A. Multi-label Decision Tables

A *multi-label decision table* T is a rectangular table filled by nonnegative integers. Columns of this table are labeled with conditional attributes $f_1, \ldots, f_n$. If we have strings as values of attributes, we have to encode the values as nonnegative integers. We do not have any duplicate rows, and each row is labeled with a nonempty finite set of natural numbers (set of decisions). We denote the number of rows in the table T by $N(T)$. We denote row i by r_i where $i = 1, \ldots, N(T)$. For example, r_1 means the first row, r_2 means the second row and so on.

TABLE I
A multi-label decision table T^0

	f_1	f_2	f_3	
r_1	0	0	0	{1}
r_2	0	1	1	{1,2}
r_3	1	0	1	{1,3}
r_4	1	1	0	{2,3}
r_5	0	0	1	{2}

If there is a decision which belongs to the set of decisions attached to each row of T, then we call it a *common decision* for T. We will say that T is a *degenerate* table if T does not have rows or it has a common decision. For example, T' is a degenerate table as shown in Table II, where the common decision is 1.

TABLE II
A degenerate multi-label decision table, T'

	f_1	f_2	f_3	
r_1	0	0	0	{1}
r_2	0	1	1	{1,2}
r_3	1	0	1	{1,3}

A table obtained from T by removing some rows is called a subtable of T. There is a special type of subtable called *boundary subtable*. The subtable T' of T is a *boundary subtable* of T if and only if T' is not degenerate but each of its proper subtable is degenerate. We denote the number of boundary subtables of the table T by $nBS(T)$. Below are examples of all boundary subtables of T_0:

T_1:

	f_1	f_2	f_3	d
r_2	0	1	1	$\{1,2\}$
r_3	1	0	1	$\{1,3\}$
r_4	1	1	0	$\{2,3\}$

T_2:

	f_1	f_2	f_3	d
r_1	0	0	0	$\{1\}$
r_4	1	1	0	$\{2,3\}$

T_3:

	f_1	f_2	f_3	d
r_3	1	0	1	$\{1,3\}$
r_5	0	0	1	$\{2\}$

T_4:

	f_1	f_2	f_3	d
r_1	0	0	0	$\{1\}$
r_5	0	0	1	$\{2\}$

The subtable of T which consists of rows that have values $a_1, \ldots, a_m$ at the intersection with columns $f_{i_1}, \ldots, f_{i_m}$ is denoted by $T(f_{i_1}, a_1), \ldots, (f_{i_m}, a_m)$. Such nonempty subtables (including the table T) are called separable subtables of T. For example, if we consider subtable $T^0(f_1, 0)$ for table T^0 (see Table I), it will consist of rows $1, 2$, and 5. Similarly, $T^0(f_1, 0)(f_2, 0)$ subtable will consist of rows 1, and 5 (see Table III).

TABLE III
Example of subtables of multi-label decision table T^0

$T^0(f_1, 0)$ =

	f_1	f_2	f_3	
r_1	0	0	0	{1}
r_2	0	1	1	{1,2}
r_5	0	0	1	{2}

$T^0(f_1, 0)(f_2, 0)$ =

	f_1	f_2	f_3	
r_1	0	0	0	{1}
r_5	0	0	1	{2}

The set of attributes (columns of table T), such that each of them has different values is denoted by $E(T)$. For example, if we consider table T^0, $E(T^0) = \{f_1, f_2, f_3\}$. Similarly, $E(T^0(f_1, 0)) = \{f_2, f_3\}$ for the subtable $T^0(f_1, 0)$, because the value for the attribute f_1 is constant (=0) in subtable $T^0(f_1, 0)$. For $f_i \in E(T)$, we denote the set of values from the column f_i by $E(T, f_i)$. As an example, if we consider table T^0 and attribute f_1, then $E(T^0, f_1) = \{0, 1\}$.

The minimum decision which belongs to the maximum number of sets of decisions attached to rows of the table T is called the *most common decision* for T. For example, the most common decision for table T^0 is 1. Even though both 1 and 2 appears 3 times in the sets of decisions, but 1 is the minimum decision, so we choose 1 as the most common decision. We denote the number of rows for which the set of decisions contains the most common decision for T by $N_{mcd}(T)$. For the table T^0, $N_{mcd}(T^0) = 3$.

B. Decision Tree

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

Let Γ be a decision tree over T and v be a node of Γ. There is one to one mapping between node v and subtable of T i.e. for each node v, we have a unique subtable of T. We define a subtable $T(v)$ of T corresponding to the node v. If node v is the root of Γ then $T(v) = T$ i.e. the subtable $T(v)$ is the same as T. Otherwise, $T(v)$ is the subtable $T(f_{i_1}, \delta_1) \dots (f_{i_m}, \delta_m)$ of the table T where attributes $f_{i_1}, \dots, f_{i_m}$ and numbers $\delta_1, \dots, \delta_m$ are respectively node and edge labels in the path from the root to node v.

We will say that Γ is a decision tree for T, if for any node v of Γ:

- if $T(v)$ is degenerate then v is labeled with the common decision for $T(v)$,
- if $T(v)$ is not degenerate then v is labeled with an attribute $f_i \in E(T(v))$, and if $E(T(v), f_i) = \{a_1, \dots, a_k\}$, then k outgoing edges from node v are labeled with $a_1, \dots, a_k$.

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

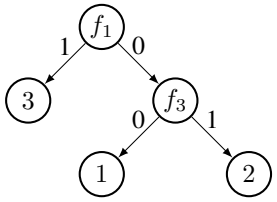


Fig. 1. A decision tree for the multi-label decision table, T^0

The number of nodes in the decision tree Γ is denoted by $N(\Gamma)$. The number of terminal and nonterminal nodes in the decision tree Γ are denoted by $N^t(\Gamma)$, and $N^n(\Gamma)$ respectively.

C. Impurity Functions and Uncertainty Measures

In greedy algorithm, we need to choose attributes to divide the decision table into smaller subtables until we get degenerate table which then be used to label the terminal node. To choose which partition to consider for tree construction, we need to evaluate the quality of partition. Impurity function is the criterion for the evaluation of quality of partition. We assume that, the smaller the impurity function value, the better is the quality of partition. We can calculate impurity function based on uncertainty measure value for the considered subtables corresponding to the partitioning. The uncertainty measure evaluates the uncertainty of the considered subtable. If we have a common decision, then there is no uncertainty in the data, hence we get uncertainty measure as zero, otherwise we will get positive values for it. We have used six different uncertainty measures, and four different impurity function types for our experiments.

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

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

- Misclassification error: $me(T) = N(T) - N_{mcd}(T)$. It measures the difference between total number of rows and number of rows with the most common decision.
- Sorted entropy: $entSort(T) = -\sum_{i=1}^{m} p_i^{'} \log_2 p_i^{'}$ (see [17]). First we sort the probability for each decision. Then, for each row, we keep the decision having maximum probability and discard others. After that, we calculate entropy for this modified decision table.
- Multi-label entropy: $entML(T) = 0$, if and only if T is degenerate, otherwise, it is equal to $-\sum_{i=1}^{m}(p_i \log_2 p_i + q_i \log_2 q_i)$, where, $q_i = 1 - p_i$. (see [19]). It measures entropy for multi-label decision table.
- $nBS(T)$: number of boundary subtables in T. We calculate number of boundary subtables using brute force approach by checking all possible subtables of T (see [10]).
- Absent: $abs(T) = q_1 \cdots q_m$, where $q_i = 1 - p_i$. It measures the absent probability q_i, and multiplies all of q_i's.
- Combined: $comb(T) = me(T) + B_2 + B_3$, where B_2 and B_3 are the number of boundary subtables with two rows and three rows, respectively. It is the combination of uncertainty measures.

2) Impurity Functions: Let $f_i \in E(T)$, and $E(T, f_i) = \{a_1, \dots, a_t\}$. The attribute f_i divides the table T into t subtables: $T_1 = T(f_i, a_1), \dots, T_t = T(f_i, a_t)$. We now define an impurity function I which gives us the impurity $I(T, f_i)$ of this partition. Let us fix an uncertainty measure U from the set $\{me, entSort, entML, nBS, abs, comb\}$, and type of impurity function from {weighted sum (ws), weighted max (wm), divided weighted sum (Div_ws), multiplied weighted sum ($Mult\_ws$)}. Then:

- wm: $I(T, f_i) = \max_{1 \leq j \leq t} U(T_j) N(T_j)$. For this type, we take the maximum among all the uncertainties of tables $T_1, \dots, T_t$ multiplied by the weights of its number of rows.
- ws: $I(T, f_i) = \sum_{j=1}^{t} U(T_j) N(T_j)$. For this type, we take the sum over all the uncertainties of tables $T_1, \dots, T_t$ multiplied by the weights of its number of rows.
- Div_ws: $I(T, f_i) = (\sum_{j=1}^{t} U(T_j) N(T_j)) / \log_2 t$. For this type, we divide the weighted sum impurity type (wt_sum) by the logarithmic function of number of branches ($\log_2 t$).

- *Mult_ws*: $I(T, f_i) = (\sum_{j=1}^{t} U(T_j)N(T_j)) \cdot \log_2 t$. For this type, we multiply the weighted sum impurity type (wt_sum) by the logarithmic function of number of branches ($\log_2 t$)

As a result, we have 24 (4 types multiplied by 6 uncertainty measures) impurity functions.

IV. Algorithms for Decision Tree Construction

In this section, we consider dynamic programming algorithm and greedy algorithms. Dynamic programming algorithm gives us optimal solution whereas greedy algorithms give us suboptimal solutions. As dynamic programming is highly time consuming, we need to choose some greedy algorithms which will be fast enough as well as their performances will be comparable to the optimal one.

A. Dynamic Programming Algorithm

We now describe an algorithm A_d which, for a given multi-label decision table constructs a decision tree with minimum size (number of nodes, or number of nonterminal nodes, or number of terminal nodes). This algorithm is based on dynamic programming approach [20], [5], and the complexity of this algorithm in the worst case is exponential.

Let T contains n conditional attribute $f_1, \ldots, f_n$. The set of all separable subtables of the table T including the table T is denoted by $S(T)$. The first part of the algorithm A_d constructs the set $S(T)$ (see Algorithm 1). For each subtable from $S(T)$, the second part of the algorithm A_d constructs a decision tree with minimum size (see Algorithm 2). Note that here size refers to either the number of nodes in the tree, or the number of nonterminal nodes in the tree, or the number of terminal nodes in the tree.

Algorithm 1 Construction of the set of separable subtables $S(T)$

Require: A multi-label decision table T with conditional attributes $f_1, \ldots, f_n$.
Ensure: The set $S(T)$
Assign $S(T) = \{T\}$, and mark T as not treated;
while (true) **do**
  if No untreated tables in $S(T)$ **then**
    Return $S(T)$;
  else
    Choose a table T_s in $S(T)$ which is not treated;
    if $E(T_s) = \phi$ **then**
      Mark the table T_s as treated;
    else
      Add to the set $S(T)$ all subtables of the form $T_s(f_i, \delta)$, where $f_i \in E(T_s)$, and $\delta \in E(T_s, f_i)$ which were not in $S(T)$, mark the table T_s as treated, and new subtables $T_s(f_i, \delta)$ as untreated.
    end if
  end if
end while

After that, A_d returns the minimum size of the optimal tree which corresponds to the table T.

Algorithm 2 Construction of a decision tree with minimum size for each table from $S(T)$

Require: A multi-label decision table T, with conditional attributes $f_1, \ldots, f_n$, and the set $S(T)$.
Ensure: Decision tree $A_d(T)$ for T.
while (true) **do**
  if T has been assigned a decision tree **then**
    Return this tree as $A_d(T)$;
  else
    Choose a table T_s in the set $S(T)$ which has not been assigned a tree yet and which is either degenerate or all separable subtables of the table T_s already have been assigned decision trees.
    if T_s is degenerate **then**
      Assign to the table T_s the decision tree consisting of one node. Mark this node with the common decision for T_s;
    else
      For each $f_i \in E(T_s)$ and each $\delta \in E(T_s, f_i)$, we denote the decision tree assigned to the table $T_s(f_i, \delta)$ by $\Gamma(f_i, \delta)$. We now define a decision tree Γ_{f_i} with a root labeled by the attribute f_i where $f_i \in E(T_s)$, and $E(T_s, f_i) = \{\delta_1, \ldots, \delta_r\}$. The root has exactly r edges $d_1, \ldots, d_r$ which are labeled by the numbers $\delta_1, \ldots, \delta_r$, respectively. The roots of the decision trees $\Gamma(f_i, \delta_1), \ldots, \Gamma(f_i, \delta_r)$ are ending points of the edges $d_1, \ldots, d_r$, respectively. Assign to the table T_s one of the trees $\Gamma_{f_i}, f_i \in E(T_s)$, having minimum size.
    end if
  end if
end while

B. Greedy Algorithms

Let I be an impurity function. For a given multi-label decision table T, the greedy algorithm A_I constructs a decision tree $A_I(T)$ for T (see Algorithm 3).

It constructs decision tree sequentially in a top-down fashion. It greedily chooses one attribute at each step based on uncertainty measure and type of the impurity function. We have total 24 algorithms. The complexities of these algorithms are polynomially bounded above by the size of the table. In case of 'number of boundary subtables' uncertainty measure, we will only consider those tables where the maximum number of decisions are bounded.

V. Comparison of Algorithms

To compare the algorithms statistically, we use Friedman test with the corresponding Nemenyi post-hoc test as suggested in [13]. Let we have k greedy algorithms $A_1, \ldots, A_k$ for constructing trees and M decision tables $T_1, \ldots, T_M$. For each decision table T_i, $i = 1, \ldots, M$, we rank the algorithms $A_1, \ldots, A_k$ on T_i based on their performance scores (from the point of view of cost functions: number of nodes, or number of nonterminal nodes, or number of terminal nodes of constructed

Algorithm 3 Greedy algorithm A_I

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

trees), where we assign the best performing algorithm as the rank 1, the second best as the rank 2, and so on. We break ties by computing the average of ranks. Let r_i^j be the rank of the j-th of k algorithms on the decision table T_i. For $j = 1, \ldots, k$, we correspond to the algorithm A_j the average rank

$$R_j = \frac{1}{M} \cdot \sum_{i=1}^{M} r_i^j.$$

For a fixed significant level α (in our work $\alpha = 0.05$), the performance of two algorithms is significantly different if the corresponding average ranks differ by at least the critical difference

$$CD = q_\alpha \sqrt{\frac{k\,(k+1)}{6M}}$$

where q_α is a critical value for the two-tailed Nemenyi test depending on α and k (see [13]).

We can also compare performance scores of algorithms $A_1, \ldots, A_k$ with optimal results obtained by dynamic programming algorithm. For $j = 1, \ldots, k$ and $i = 1, \ldots, M$, we denote, by N_{ij} the number of nodes of the decision tree constructed by the algorithm A_j on the decision table T_i. For $i = 1, \ldots, M$, we denote the minimum possible number of nodes of a decision tree for T_i by N_i^{opt}. Thus, we can compute the average relative difference in percentage for number of nodes as

$$ARD_j^N = \frac{1}{M} \sum_{i=1}^{M} \frac{N_{ij} - N_i^{opt}}{N_i^{opt}} \times 100\%.$$

Similarly, for number of nonterminal nodes ($N_n(\Gamma)$), we have

$$ARD_j^{N^n} = \frac{1}{M} \sum_{i=1}^{M} \frac{N_{ij}^n - N^{n\,opt}_i}{N^{n\,opt}_i} \times 100\%.$$

Similarly, for number of terminal nodes ($N_t(\Gamma)$), we have

$$ARD_j^{N^t} = \frac{1}{M} \sum_{i=1}^{M} \frac{N_{ij}^t - N^{t\,opt}_i}{N^{t\,opt}_i} \times 100\%.$$

VI. Experimental Results

We consider 16 decision tables from UCI Machine Learning Repository [12]. There were missing values for some attributes which were replaced with the most common values of the corresponding attributes. Some conditional attributes have been removed that take unique value for each row. To convert such tables into multi-label decision table format, we removed the more conditional attributes from these tables. As a result we obtained inconsistent decision tables which contained equal rows with different decisions. Each group of identical rows was replaced with a single row from the group which is labeled with the set of decisions attached to rows from the group. The information about obtained multi-label decision table can be found in Table IV. Modified decision table has been renamed in Table IV by the name of initial table plus an index equal to the number of removed conditional attributes. Table IV also contains the number of rows (column "Rows"), the number of attributes (column "Attr"), and the spectrum of the corresponding decision table (column "Spectrum"). Spectrum of a multi-label decision table is a sequence $\#1$, $\#2$,..., where $\#i$, $i = 1, 2, \ldots$, is the number of rows labeled with sets of decisions with the cardinality equal to i.

TABLE IV
Characteristics of multi-label decision tables

Decision table T	Rows	Attr	Spectrum					
			$\#1$	$\#2$	$\#3$	$\#4$	$\#5$	$\#6$
balance-scale-1	125	3	45	50	30			
breast-cancer-1	193	8	169	24				
breast-cancer-5	98	4	58	40				
cars-1	432	5	258	161	13			
flags-5	171	21	159	12				
hayes-roth-data-1	39	3	22	13	4			
lymphography-5	122	13	113	9				
mushroom-5	4078	17	4048	30				
nursery-1	4320	7	2858	1460	2			
nursery-4	240	4	97	96	47			
spect-test-1	164	21	161	3				
teeth-1	22	7	12	10				
teeth-5	14	3	6	3	0	5	0	2
tic-tac-toe-4	231	5	102	129				
tic-tac-toe-3	449	6	300	149				
zoo-data-5	42	11	36	6				

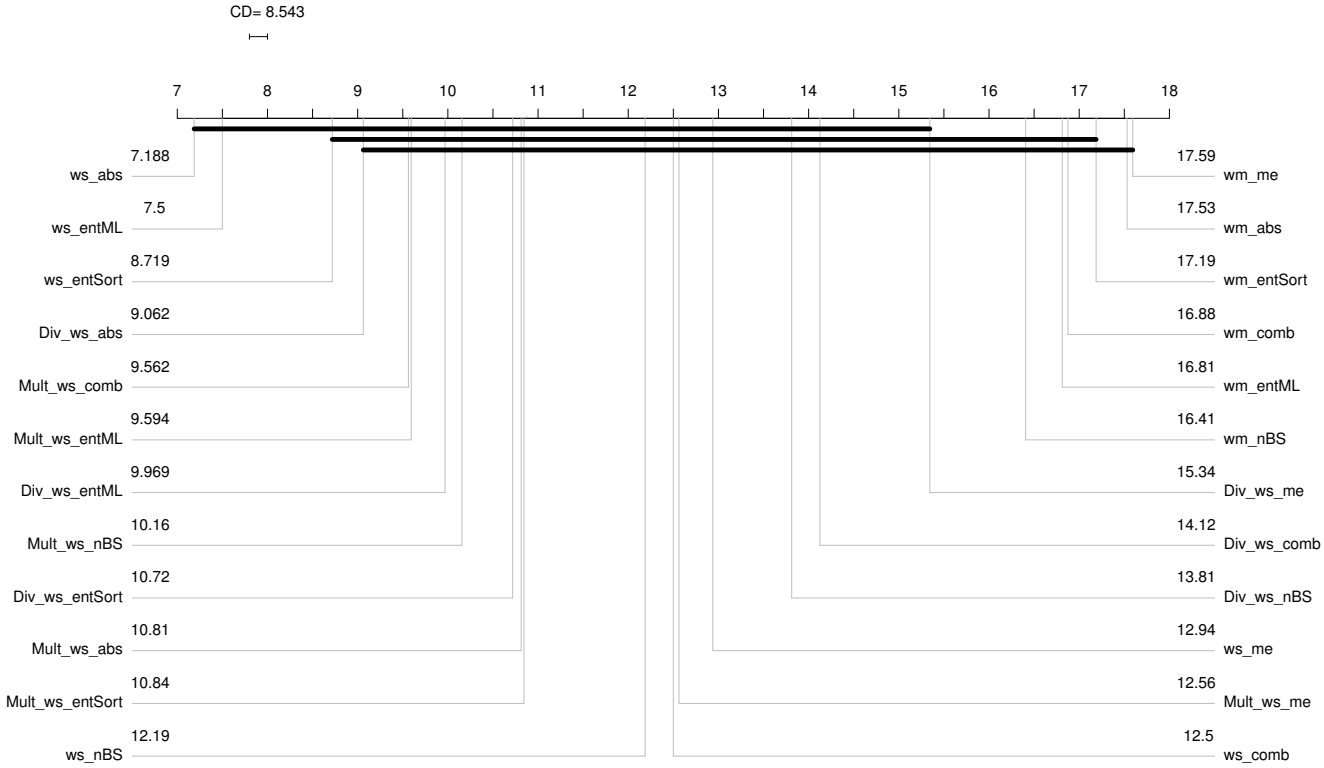


Fig. 2. CDD for number of nodes of decision trees constructed by greedy algorithms

We have six uncertainty measures (me, nBS, abs, $comb$, $entSort$, $entML$) and four types of impurity functions (ws, wm, Div_ws, $Mult\_ws$), so total 24 greedy algorithms have been compared. In the critical difference diagram (CDD), we showed the names of the algorithms as combined name of heuristic and impurity function types separated by '_'. For example, if the algorithm name is wm_nBS, this means it uses wm as a type of impurity function and nBS as uncertainty measure.

Figure 2 shows the CDD containing average (mean) rank for each algorithms on the x-axis for significant level of $\alpha = 0.1$. When Nemenyi test cannot identify significant difference between some algorithms, the algorithms are clustered (connected). It is clear from Figure 2 that, 18 algorithms from the 24 algorithms are clustered in the first group (left most algorithms is the best ranked algorithm) which are the leaders among all greedy algorithms for minimization of number of nodes in decision trees, and the best ranked algorithm is ws_abs. We have shown the best three algorithms having minimum ARD in Table V. If we look at the ARD table, we can see the best algorithm that is closer to the optimal results is ws_abs, and the average relative difference is only 18.92%.

Also, we can see from Figure 3 that, 17 algorithms among 24 algorithms are leaders for minimization of number of nonterminal nodes in decision trees, and the best ranked algorithm is ws_abs. We have shown the best three algorithms having minimum ARD for minimization of number of nonterminal nodes in Table VI, and the ARD for ws_abs is only 20.76% relative to the optimal results.

Now, for the minimization of number of terminal nodes, we can see from Figure 4 that 19 algorithms from 24 algorithms are in the best group, and the best ranked algorithm is $Mult\_ws\_entML$. Also from ARD Table VII, we can see that $Mult\_ws\_entML$ is closer to the optimal results by only 18.71%.

TABLE V
ARD IN PERCENTAGE BETWEEN RESULTS OF GREEDY AND DYNAMIC ALGORITHMS FOR TOTAL NUMBER OF NODES

Algorithm	ARD
ws_abs	18.92%
$Mult\_ws\_entML$	19.73%
ws_entML	20.58%

TABLE VI
ARD IN PERCENTAGE BETWEEN RESULTS OF GREEDY AND DYNAMIC ALGORITHMS FOR NUMBER OF NONTERMINAL NODES

Algorithm	ARD
ws_abs	20.76%
ws_entML	22.6%
$ws\_entSort$	23.7%

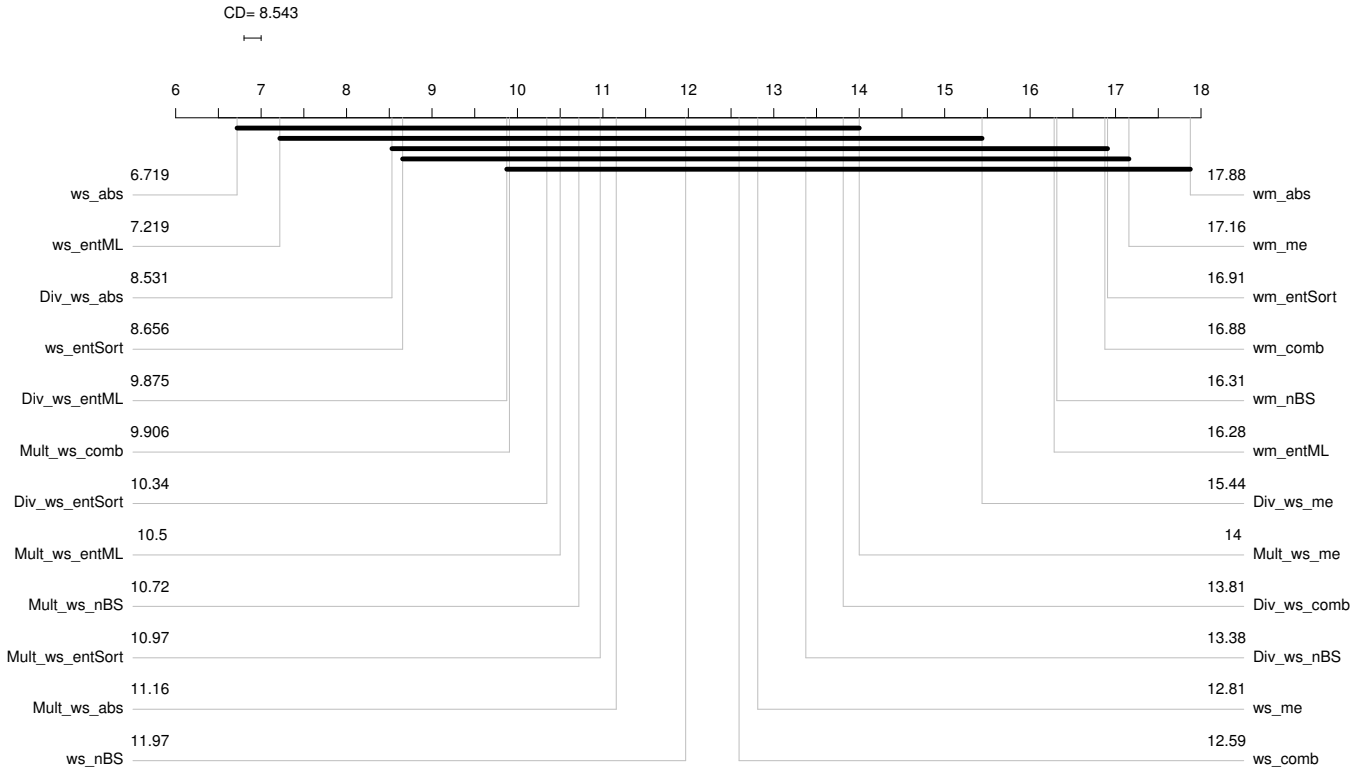


Fig. 3. CDD for nonterminal nodes of decision trees constructed by greedy algorithms

TABLE VII
ARD IN PERCENTAGE BETWEEN RESULTS OF GREEDY AND DYNAMIC ALGORITHMS FOR NUMBER OF TERMINAL NODES

Algorithm	ARD
$Mult\_ws\_entML$	18.71%
ws_abs	21%
$ws\_entSort$	22.49%

VII. CONCLUSION

In this paper, we studied greedy algorithms for decision tree construction which are based on various impurity functions. We compared these algorithms to find out which algorithm gives us minimum number of nodes, minimum number of nonterminal nodes and minimum number of terminal nodes. We also considered the average relative difference between optimal results and results obtained by the greedy algorithms. We found that, for the best greedy algorithm, it is at most 18.92% ARD for the minimization of number of nodes, at most 20.76% ARD for the minimization of number of nonterminal nodes, and 18.71% ARD for the minimization of terminal nodes, which are promising results. In future, our goal is to compare the above best greedy algorithm for the problem of prediction i.e. to minimize the prediction error.

ACKNOWLEDGEMENT

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

REFERENCES

[1] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004. doi: 10.1016/j.patcog.2004.03.009

[2] A. Wieczorkowska, P. Synak, R. A. Lewis, and Z. W. Ras, "Extracting emotions from music data," in *ISMIS*, 2005. doi: 10.1007/11425274 pp. 456–465.

[3] H. Blockeel, L. Schietgat, J. Struyf, S. Dzeroski, and A. Clare, "Decision trees for hierarchical multilabel classification: A case study in functional genomics," in *PKDD 2006, Berlin, Germany, Proceedings*, ser. LNCS, J. Fürnkranz, T. Scheffer, and M. Spiliopoulou, Eds. Springer, 2006, vol. 4213, pp. 18–29.

[4] Z.-H. Zhou, K. Jiang, and M. Li, "Multi-instance learning based web mining," *Appl. Intell.*, vol. 22, no. 2, pp. 135–147, 2005. doi: 10.1007/s10489-005-5602-z

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

[6] F. D. Comité, R. Gilleron, and M. Tommasi, "Learning multi-label alternating decision trees from texts and data," in *MLDM*, 2003. doi: 10.1007/3-540-45065-3 pp. 35–49.

[7] E. Loza Mencía and J. Fürnkranz, "Pairwise learning of multilabel classifications with perceptrons," in *IJCNN*, 2008. doi: 10.1109/IJCNN.2008.4634206 pp. 2899–2906.

[8] G. Tsoumakas, I. Katakis, and I. P. Vlahavas, "Mining multi-label data," in *Data Mining and Knowledge Discovery Handbook*, 2010, pp. 667–685.

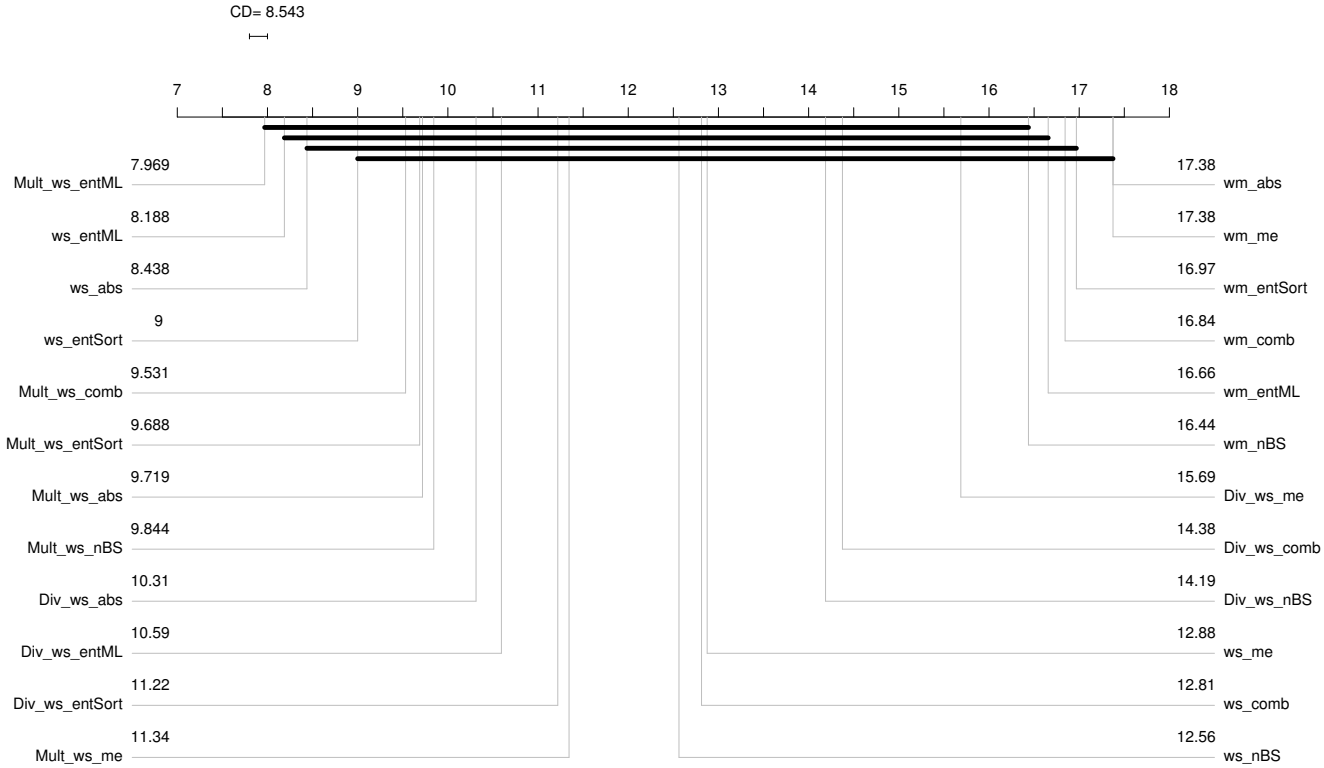


Fig. 4. CDD for terminal nodes of decision trees constructed by greedy algorithms

[9] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li, "Multi-instance multi-label learning," *Artif. Intell.*, vol. 176, no. 1, pp. 2291–2320, 2012. doi: 10.1016/j.artint.2011.10.002

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

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

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

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

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

[15] X. Zhu and A. B. Goldberg, *Introduction to Semi-Supervised Learning*, ser. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2009.

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

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

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

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

[20] M. Moshkov and I. Chikalov, "On algorithm for constructing of decision trees with minimal depth," *Fundam. Inform.*, vol. 41, no. 3, pp. 295–299, 2000. doi: 10.3233/FI-2000-41302

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 75–83
DOI: 10.15439/2014F261
ACSIS, Vol. 2

The inverse infection problem

András Bóta
University of Szeged,
Institute of Informatics,
P. O. Box 652.,
6701 Szeged, Hungary
Email: bandras@inf.u-szeged.hu

Miklós Krész
University of Szeged,
Gyula Juhász Faculty of Education,
Boldogasszony sgt. 6,
6720 Szeged, Hungary
Email: kresz@jgypk.u-szeged.hu

András Pluhár
University of Szeged,
Institute of Informatics,
P. O. Box 652.,
6701 Szeged, Hungary
Email: pluhar@inf.u-szeged.hu

***Abstract*—The applications of infection models like the Linear Threshold or the Domingos-Richardson model requires a graph weighted with infection probabilities. In many real-life applications these probabilities are unknown; therefore a systematic method for the estimation of these probabilities is required. One of the methods proposed to solve this problem, the Inverse Infection Model, was originally formulated for estimating credit default in banking applications. In this paper we are going to test the capabilities of the Inverse Infection Model in a more controlled environment. We are going to use artificially created graphs to evaluate the speed and the accuracy of estimations. We are also going to examine how approximations and heuristics can be used to improve the speed of the calculations. Finally, we will experiment with the amount of a priori information available in the model and evaluate how well this method performs if only partial information is available.**

I. Introduction

THE STUDY of infection processes has roots in two seemingly different fields of research: sociology and the medical sciences. In the latter, it was used to model the spread of epidemics [9]. Applications focused on prevention, and the identification of the “choke points" during an epidemic. In the former, the spreading of information or opinions came into focus. One of the earliest models in sociometry, Granovetter’s Linear Threshold [12] model is still considered to be a viable description of information diffusion.

In economics, Domingos and Richardson developed the Independent Cascade model (IC) [10] for the purpose of viral marketing. They proposed the influence maximization problem, that is to find the set of k initial infectors for any k that results in the largest expected infection. Kempe et al. [15], [16] proved the influence maximization problem was NP-hard, proposed a greedy algorithm for it, and also showed that the generalization of the IC model is in fact an equivalent of the Linear Threshold model. They also used random simulations to approximate the vertex infection probabilities, and they choose an arbitrary constant for edge infection probabilities. This result stresses the importance of the exact computation of vertex infection probabilities. This problem was proven to be #P-complete by Cao [4].

Computing the maximal infection or the exact probabilities of infection with any kind of model requires a weighted network, that is the edge infection probabilities must be available. This information is usually not known beforehand. In most real-life applications, the edges are considered to be some constant, or estimated using intuition guided trial-and-error method based on known edge or vertex attributes. Recently, a few papers were published in this topic discussing systematic approaches for the estimation of edge infection probabilities. In some of them [11], [17], the steps or iterations of the infection process are assumed to be known, which is realistic in the case of twitter or blog-based networks.

The Inverse Infection Problem, an application-driven approach was proposed recently by the authors [1] for the prediction of credit default in bank transaction networks. Unlike the above mentioned methods, this does not require information on the individual steps of the infection process. Instead it builds on other available data, such as estimations of the probabilities of default for individual companies and additional information characterizing the connection between the companies.

Based on the good results of this method in applications, our goal in this paper is to provide a solid foundation to the Inverse Infection Problem in a more controlled environment. Our method is based on the Generalized Cascade (GC) model [3], a generalization of the Independent Cascade model. To compute the edge infection probabilities themselves we will use a meta-heuristic: the Particle Swarm Algorithm of Kennedy and Mendes [14].

The paper is constructed as follows. In the next section we will give a short introduction into infection mechanisms, define the GC model, and the Inverse Infection Problem. In Section III we will describe several ways to accurately estimate the infection probabilities, including gradient-based methods and Particle Swarm Optimization. Then we will discuss various options to customize the estimations including heuristics of the GC model, choices for attribute functions and the number of patterns required to accurately estimate the infection probabilities.

II. Problem definition

The process of infection takes place on a graph G, where $V(G)$ denotes the set of vertices, and $E(G)$ denotes the set of edges. While most traditional models require directed edges, depending on the application, they can be easily modified to handle undirected ones. We also need to know the edge

978-83-60810-58-3/$25.00

infection probabilities, that is a weight $w_e \in [0,1]$ for each edge e.

The notion of states is important. Each vertex of the network has a state of infection. The number of states and the transitions between them are governed by the specific model. One of the most basic approaches, the SIR model, [9] has three states: Susceptible, Infected and Recovered. Infected nodes infect susceptible ones, but after a certain period, which is usually a parameter of the model, they may recover, no longer infectious. Models in epidemics have a variety of states and the transitions between them are often more complicated. Models in economics or models describing information diffusion can be considered simpler. In the case of the Independent Cascade model, there are three states loosely corresponding to the ones in the SIR model and the infection period only lasts for one iteration. These three states are: susceptible, just infected (and still infectious), infected (but no longer infectious).

Most infection processes are also iterative, that is the process takes place in discrete time steps. Those models, that allow nodes to become susceptible again some time after becoming infected, may not terminate. It is easy to see, that the IC model terminates in finite steps.

A. Infection Models

Any infection model can be described as a process, that has two inputs: the first one is a weighted graph, where the edge weights are probabilities. The second input is the set of initial infectors $A_0 \subset V(G)$. These nodes are considered as infected at the beginning of the process. The process terminates at iteration t, and results in the set of infected nodes $A = \bigcup_{i=0}^{t} A_i$.

The specific way one vertex infects another varies depending on the model. In the case of the IC model [10], let $A_i \subseteq V(G)$ be the set of nodes newly activated in iteration i. In the next iteration $i+1$, each node $u \in A_i$ tries to activate its inactive neighbors $v \in V \setminus \cup_{0 \leq j \leq i} A_j$ according to the edge infection probability $w_{u,v}$, and v becomes active in iteration $i+1$, if the attempt is successful. If more than one node is trying to activate v in the same iteration, the attempts are made independently of each other in an arbitrary order within iteration $i+1$. If $A_t = \emptyset$, the process terminates in iteration t. It is easy to see, that the process always terminates in a finite number of iterations.

B. Generalized Cascade Model

Following the works of Bóta et al. [3], we can generalize this model in the following way. Instead of using vertex sets for representing the initial infectors, we work with two probability distributions. The *a priori* distribution defines the probability, that a vertex becomes infected on its own, independently of other vertices at the beginning of the process. The *a posteriori* defines the probability, that a vertex becomes infected at the end of the process. For all vertices $v \in V(G)$, we will denote the a priori probability of infection as p_v, the a posteriori as p'_v.

In some applications, an estimate of one or both of the above described probability distributions is available. For example, in the case of the banking application [1], [7] an accurate estimation of the probability of default for each company was given by standard models used by the bank[1]. Another application in telecommunications uses estimations for the probability of churn using similar methods. If such estimations are not available we can resort to a crude but effective method. Suppose we can observe the beginning and the end of the infection process k times. By counting the frequencies of infection, for all vertices v, how many times did v belong to A_0 or A we can construct the respective probability distributions. The accuracy of the estimation obviously depends on k, but k does not have to be a large number. We will show in section IV.D, that 6-8 observations are enough to produce outputs with acceptable quality.

Based on these remarks and formulations, we can define the Generalized Cascade model [2]:

The Generalized Cascade Model: *Given an appropriately weighted graph G and the a priori infection distribution p_v, the model computes the a posteriori distribution p'_v for all $v \in V(G)$.*

The infection process itself is the IC model, although other models might also be used for different applications. We have chosen the Independent Cascade model as the basis of our method, because it performs well in modeling infection-like processes in business applications [7]. Alternatively, this model can be considered as a general framework of infection.

Unfortunately, the computation of the a posteriori distribution in the IC model is #P-complete. There are several existing heuristics to provide estimations of p'_v [5], [6], including the ones the authors proposed in [2]. Two of these are Monte Carlo based simulations. *Complete Simulation* is a direct adaptation of the idea of Kempe et al. to the framework of the GC model. The basis of the idea is the notion of reachability. By selecting the edges $(u,v) \in E(G)$ independently of each other according to their infection probabilities $w_{u,v}$, they construct an unweighted graph which is a realization of the infection process. Any vertex, that can be reached from any initially infector is considered to be infected. We can adapt this process into the GC model by computing a large number of individual runs of the model and counting the frequencies of infections (both a priori and a posteriori). The process has an unfortunate property: the frequency (or sample size) must be high enough to reduce the standard deviation characteristic of Monte Carlo based methods.

The *Edge Simulation* method decreases the standard deviation of the previous method. In each run, a subgraph containing all of the vertices able to infect the individual vertex v is constructed for all vertices $v \in V(G)$. This way the a posteriori infection of v can be computed directly in each run. The results of individual runs are averaged. The authors have proposed two additional heuristics: In the *Neighborhood Bound Heuristic* a tree is constructed from the 2-neighborhood

[1]The BASEL II default probabilities were computed using vertex attributes.

of a given vertex v representing all possible routes of infection. Both the tree and the a posteriori infection of v can be computed in a short time, resulting in a very fast heuristic. The *Aggregated Linear Effect* model is a linear approximation of the mechanism of the IC model. A more detailed description of these methods can be found in [2].

C. Inverse Infection Problem

Based on the framework of the Generalized Cascade model, we can define the Inverse Infection Problem.

Inverse Infection Problem: *Given an unweighted graph G, the a priori and the a posteriori probability distributions p_v and p'_v, compute the edge infection probabilities w_e for all $e \in E(G)$.*

Directly estimating each individual edge in a graph is computationally infeasible even in small graphs. However, in real-life applications the probability of infection between vertices is a combination of other properties of the edges, vertices and the graph itself. We are going to take advantage of this fact to simplify computations and make the problem solvable in reasonable time. We are going to assume, that on each edge there are several *attributes*[2], and the infection probability of the edge is a parametrized*function* of these attributes[3]. This way, only the *coefficients* of this function have to be estimated, which is a small number compared to the number of edges.

There are multiple ways to define these functions. In this paper, we are going to consider, that there is a polynomial function f_i on each individual attribute $a_i, i = 1, \ldots, \ell$, where ℓ is the number of attributes. Then, a normalized sum or product is calculated from each $f_i(a_i)$ resulting in the infection probability w_e. The degree of these polynomials should be low, but we allow different polynomials on different edges, with possibly different degrees. If the maximum degree of these polynomials is f_{max}, then there are at most $(f_{max}+1)\ell$ coefficients to estimate.

III. Estimation with learning methods

To provide a solution for the Inverse Infection Problem, we have developed the following learning algorithm. The problem definition states, that the a posteriori distribution is required as an input of any algorithm. In the case of a learning algorithm, it is considered as a test or reference dataset. By taking the a priori distribution we compute an estimation of the a posteriori distribution with some initially random starting coefficients. Then, an error function calculates the difference between the reference set and the newly calculated infection values. Our goal is finding the global minimum of this error function: the difference between the a posteriori vertex infections. Using an optimization algorithms, we can efficiently estimate the coefficients of the attribute-functions and thus the edge weights.

[2]Vertex attributes can be easily converted into edge attributes.

[3]It is possible, that some of these attributes have no influence on the infection probability, but we expect the method to ignore the effect of these.

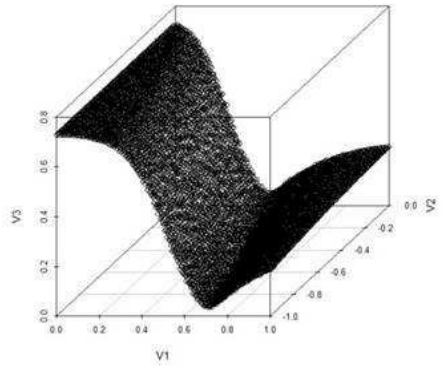


Fig. 1. The error surface of an IIP with one edge attribute and $f_1(a_1) = c_0 * a_1 + c_1$ as the attribute function. Root mean squared error was used for the evaluation.

A. Previous approaches and experiences

We have tried several optimization algorithms, including more simple ones, like grid search, gradient-based methods and meta-heuristics. Our first analysis was performed on banking data where we used grid search for optimization. While this early approach provided promising results [7], it was clear that further refinement of the optimization algorithm was required. Later, we have implemented a multi agent gradient based method, and compared the performance of it with our previous results [3]. The gradient method provided more accurate estimations, but highlighted several unfortunate properties of the problem itself.

Our first observation was that the error function was noisy. This comes from using Monte Carlo methods to approximate the IC model, since the deviation of these simulations makes different runs with the same coefficient values have different results. The noise can be reduced by increasing the frequency parameter of the simulation, but this also increases the time complexity of the method [3].

The second observation was that the problem is underdetermined. Different edge weight configurations can result in the same infection pattern, the same a posteriori distribution. This results in alleys and plateaus on the error surface. In the case of the example on Figure 1, the global minimum is in the middle of the alley. Even in this simple example neither algorithms are able to reliably find the best solution.

Grid search had serious performance issues both in finding the global minimum and in time complexity. Due to its search pattern, its precision is simply not enough to tackle with this surface, and it also scales exponentially with the number of coefficients. The gradient method also performs poorly: it easily gets lost on the alleys and plateaus especially if they are noisy as well. As a consequence, it rarely finds a solution close to the global minimum, and the number of steps it takes to find a solution at all can be quite high.

We have tried several error functions, mainly vector distance

measurements, and ROC evaluation. One of our first experiences was, that the latter is not enough to properly guide the optimization method to the global minimum, so we have shifted our attention to other measurements, and finally settled on the root mean squared error. In this work we are going to use the RMSE as an error measurement, that is we are looking for the minimum of

$$\sqrt{\frac{1}{|V(G)|}\sum_{v\in V(G)}(\vec{p}'_v - \vec{p}_v)^2}, \tag{1}$$

where $\vec{p}'_v$ denotes the estimated a posteriori infection of vertex v.

B. Particle-Swarm Optimization

In order to handle the above mentioned problems, we have decided to implement the particle swarm optimization algorithm of Kennedy [13]. This is an iterative method based on the interaction of multiple agents or "particles". Each agent corresponds to a different coefficient configuration, representing a coordinate in the parameter space of the problem[4].

Apart from the coordinates themselves, the agents also have a velocity. In each iteration the position of an agent is updated by adding its velocity. The velocity of the agent is computed using the best solution the agent has found and the best solutions of the neighboring agents; the goodness of the solution is measured by evaluating the error function on the coordinates visited by the particles. Agents are connected to each other according some topology describing the neighborhood of each agent.

The specific way the velocities of the agents are updated and the topology itself is not fixed: there are various approaches in the literature for specific applications and for more general problem solving. In our work we have followed the recommendations of Kennedy and Mendes [14], and found, that it performs well in finding coefficient configurations close to the global minimum.

We have used the Fully Informed Particle Swarm published in [14] with 9 agents in a von Neumann neighborhood[5]. The position and the velocity of the agents are updated according to the following equations:

$$\vec{v}_i \leftarrow \chi\left(\vec{v}_i + \sum_{n=1}^{N_i}\frac{U(0,\varphi)(\vec{b}_{nbr(n)} - \vec{x}_i)}{N_i}\right), \tag{2}$$

$$\vec{x}_i \leftarrow \vec{x}_i + \vec{v}_i, \tag{3}$$

where $\vec{x}_i$ and $\vec{v}_i$ denotes the coordinate and velocity of particle i, $U(min, max)$ is a uniform random number generator, $\vec{b}_i$ is the best location found so far by particle i, N_i is the number of neighbors i has and $nbr(n)$ is the nth neighbor of i. The formula has two parameters: χ is the constriction coefficient and φ is the acceleration constant. Again, we have used the recommendations of Kennedy et al., and set $\chi = 0.7298$ and $\varphi = 4.1$.

Algorithm 1 Particle Swarm Optimization

```
for all a_i do
    Initialize x_i for agent a_i within the boundaries of the search space
    Initialize v_i for agent a_i
    Set b_i ← x_i
    Select the neighbors of a_i according to the topology
end for
repeat
    for all a_i do
        Update v_i according to equation 2
        Update x_i according to equation 3
        Calculate the error function e(x_i) in position x_i
        if e(x_i) < e(b_i) then
            b_i ← x_i
        end if
    end for
until termination criterium is met
```

At the beginning of the search, the agents are initialized with zero velocities and random starting coordinates within some reasonable bounds of them. Then in each iteration these two vectors are updated according to equations 2 and 3 in a synchronized manner. The search is completed if the global minimum found considering all agents does not change for five consecutive iterations. We have experimented with other values and found, that increasing it does not improve the quality of the results, and decreasing it does not reduce the running time considerably.

IV. Evaluation

The most natural way to evaluate the stability of the optimization method is by counting the average and maximum number of iterations the method takes before it finishes. However, the quality of the solution of the Inverse Infection Problem depends on additional factors; we will discuss three of these. The first one is the choice of the attribute functions. Choosing an appropriate function is important, since depending on the available attributes this function either maps into the $[0, 1]$ interval directly or some additional form of normalization is required. The second one is the choice of heuristics applied for the GC model. These have a serious impact on both the accuracy and the running time of the learning method. The third factor is the number of learning patterns available. In case the exact a priori and a posteriori infection probabilities are not available, the only thing to do is to rely on counting the frequencies of infections. In real life we cannot hope to witness an infection process on any network in more than a handful of times. It is therefore necessary to investigate the sensitivity of our method to low-quality inputs.

As a basis of our analysis we have used graphs generated with the forest fire method of Leskovec et al. [18]. We have

[4]Again, the subject of the optimization is the coefficient values of the attribute function(s)

[5]Each agent has four neighbors in a grid, connected to the upper, lower, left and right, while wrapping around the edges.

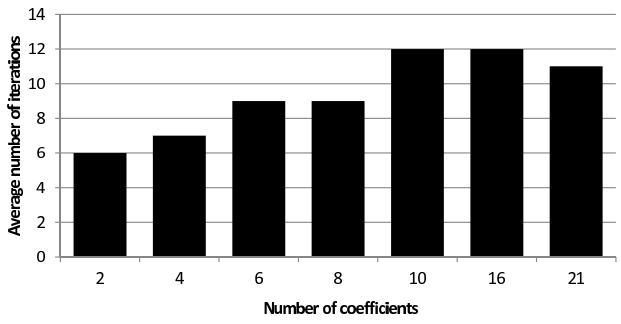


Fig. 2. The average number of iterations with different function configurations.

created a series of graphs of sizes $n = 1000, \ldots, 100000$, with parameters $p = 0.37$ and $p_b = 0.32$, for forward and backward burning probabilities, respectively. We have assigned a number of edge attributes $a_i, i = 2, \ldots, 10$ to these networks. These attributes are randomly generated: they were drawn independently from a uniform distribution between $[0, 0.5]$. We have also used randomly generated a priori infections. The expected size of the set of initial infectors is $0.3{*}n$. If v is selected, then an a priori infection probability was drawn from an uniform distribution between $[0, 0.5]$, otherwise $p_v = 0$. We have used various attribute functions, description of these will be given in section IV.B. Finally for each network and each attribute function we have created an a posteriori infection distribution as the reference dataset. For this purpose we have used Compete Simulation with sample size $k = 10000$, because this method gives the best approximation of the original IC model.

A. Stability of the optimization

The performance of the optimization method itself can be measured in two ways. The distance between the solution found by the method and the global minimum is conveniently measured by the error function itself. However, the precision of the algorithm also depends on the heuristics used to approximate the Generalized Cascade model. Consequently, we will discuss this in the following sections.

The time complexity of the method is the sum of two distinct parts of the algorithm: evaluating points on the error surface and the search method itself. The latter consists of the repeated evaluation of the formula above, after initializing the neighborhood and the starting coordinates. Since the number of agents is small, this part of the algorithm is very fast, and has negligible impact on the running time of the overall method.

In each iteration every agent evaluates the error function. This evaluation is the computation of the GC model using the coordinates - coefficients of the given agent. The time complexity of this step heavily depends on the used heuristic. Altogether, we can say, that the time complexity of a single run is $s * a * h$, where s is the number of iterations, a is the number of agents (a constant) and h is the time complexity of the infection heuristic. This also means, that we can describe the time complexity of the algorithm by measuring the average or maximum number of iterations and multiplying it with the time complexity of the infection method and the number of agents. Breaking the time complexity of the method into two different factors makes sense because of another reason: the individual runs of the GC heuristics may be run on multiple threads simultaneously, significantly improving the speed of the method.

On Figure 2 we can see the average number of iterations for different numbers of coefficients. We have used a small network with $|V(G)| = 1000$. The point of interest here is, starting from a simple problem with only two coefficients to more complex ones, the expected number of iterations grows slowly, and stabilizes around 12. The maximum number of iterations remains bounded as well, even in the experiment with 21 coefficients, it does not go beyond 30. The results shown on Figure 2 were computed with 9 agents. We have tried this problem with 16 agents as well and got similar results. If we compare the different infection heuristics, they perform similarly, with the non-Monte Carlo methods finishing slightly sooner, usually by 4-5 iterations.

We can conclude, that the Particle-Swarm Optimization method described in this section is able to solve the Inverse Infection Problem with satisfying results. The algorithm is very stable, and even in the worst case, it finishes within 30 iterations. We will evaluate the precision and running times of this method considering different heuristics of the Generalized Cascade model in section IV.C. We will also discuss choices for attribute functions, and the number of patterns required to get good estimations of the edge infection probabilities.

B. Choice of attribute functions

We have seen, that in our model, the edge infection probabilities are computed from some additional information on the edges in the form of edge attributes by so-called attribute functions. The choice of these attribute functions is an important part of our method. A natural requirement of this choice is, that it must result in infection probabilities: it must map into the $[0, 1]$ interval.

There are two approaches to this problem: the first one is to construct problem-specific functions, taking into account the structure of the network, the nature of the infection model and the number and domain of the attributes. This way it is possible to calculate the infection probabilities directly, without any form of additional normalization. This is the obvious choice if the above mentioned information is available.

If we do not have this information, we can try a more user-friendly approach. We can apply functions to the individual attributes, summarize them and finally normalize them. A variety of functions might be considered for this purpose. In our work, we have used low-degree polynomials for the individual attributes and simple addition or multiplication to join them. We have normalized the resulting edge infection probabilities according to

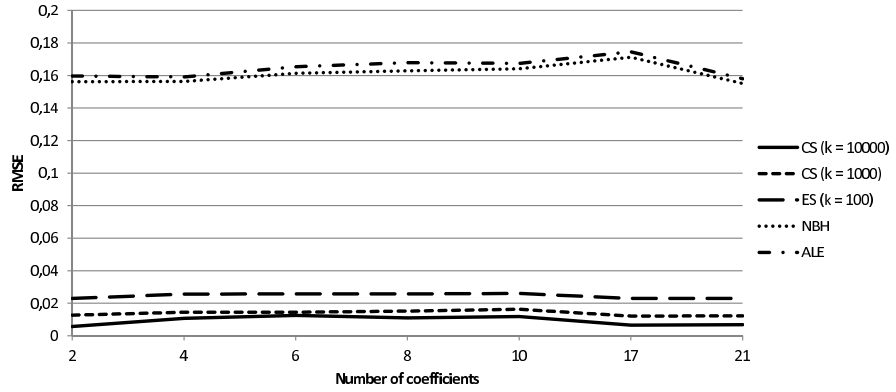


Fig. 3. The RMSE with different function configurations on a small network with $n = 1000$.

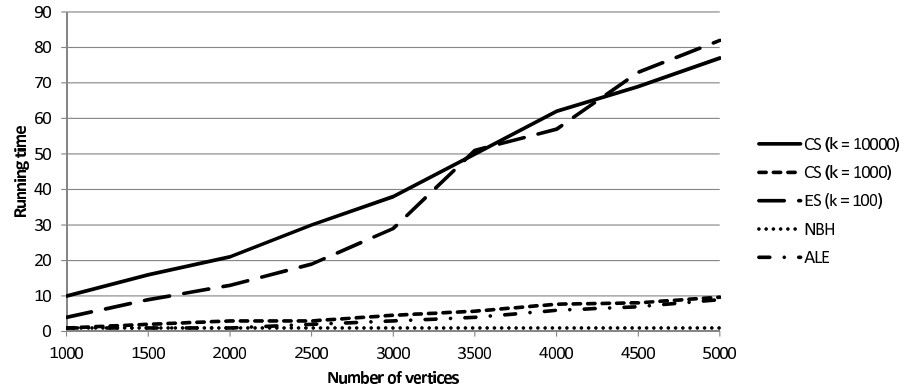


Fig. 4. The running time of the infection heuristics with different network sizes measured in seconds. Figure used with the permission of the authors [2].

$$\text{norm}(\vec{e}) = \frac{\vec{e} - \min(\vec{e})}{3(\max(\vec{e}) - \min(\vec{e}))}, \tag{4}$$

where $\vec{e}$ is a vector containing the infection probabilities for each individual edge. The reason why we have used the multiplier 3 in the denominator is, that according to our findings in the prediction of default events on banking data, the edge infection probabilities are low [7]. The normalizer function obviously distorts the shape of the individual attribute functions, but in real-life problems a simple weighted, normalized sum of attributes is sufficient to produce acceptable results.

In this paper, we have used seven attribute function configurations, a_i denotes attribute i and c_j denotes coefficient j:

- Weighted sum of two attributes: $c_1a_1 + c_2a_2$, two coefficients in total.
- Weighted sum of four attributes: $\sum_i c_ia_i, i = 1, 2, 3, 4$, four coefficients in total.
- Weighted sum of six attributes: $\sum_i c_ia_i, i = 1, \ldots, 6$, six coefficients in total.
- Weighted sum of eight attributes: $\sum_i c_ia_i, i = 1, \ldots, 8$, eight coefficients in total.
- Weighted sum of ten attributes: $\sum_i c_ia_i, i = 1, \ldots, 10$, ten coefficients in total.
- Sum of quadratic polynomials with eight attributes $c_1 + \sum_i (c_{2i}a_i^2 + c_{2i+1}a_i), i = 1, \ldots, 8$, 17 coefficients in total.
- Sum of quadratic polynomials with ten attributes $c_1 + \sum_i (c_{2i}a_i^2 + c_{2i+1}a_i),\ i = 1, \ldots, 10$, 21 coefficients in total.

In section IV, we have tested the effect of these function configurations on the stability and accuracy of the optimization method. Details of these can be found in the appropriate subsections.

C. Accuracy and the choice of heuristics

Previously, in section II.B, we have given short descriptions of some heuristics of the GC model [2]. In this section we will evaluate the performance of them in relation with the learning method described above. Complete Simulation is a direct adaptation of the idea of Kempe et al. [15], it can be considered as the best approximation of the original IC model. Therefore, we will use CS with sample size $k = 10000$ to create an a posteriori distribution as a reference set. Then, we

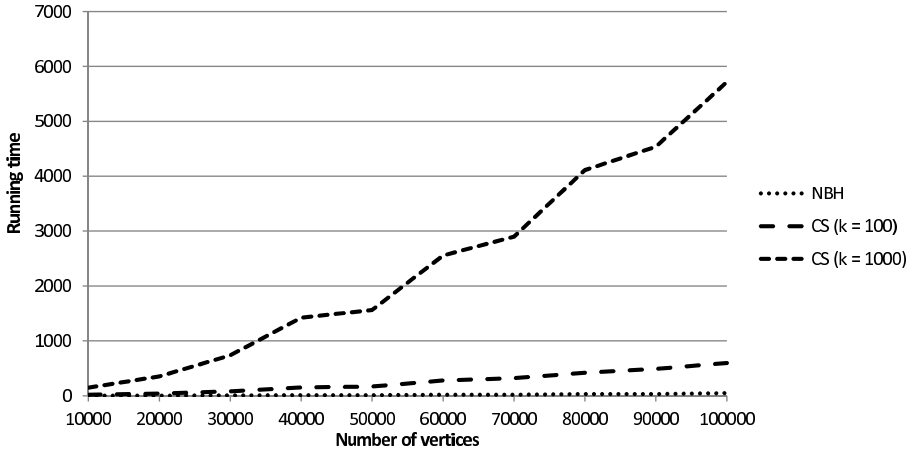


Fig. 5. The running time of the infection heuristics on large networks measured in seconds.

will use each heuristic together with the learning method to compute the edge infection probabilities:

- Complete Simulation (CS) with sample size $k = 10000$, a very accurate simulation.
- Complete Simulation (CS) with sample size $k = 1000$, a very fast simulation.
- Edge Simulation (ES) with sample size $k = 100$, a simulation based heuristic.
- Neighborhood Bound Heuristic (NBH), an extremely fast lower approximation.
- Aggregated Linear Effect (ALE) model, a de Groot [8] based simplification of the infection process.

Since the running time and the accuracy of the heuristics are different, we are going to use two different datasets to evaluate their performance. First, small networks with $|V(G)| \leq 5000$ will be used to make general observations, then we will test the more robust heuristics on large networks with $|V(G)| = 10000, \ldots, 100000$. Our largest network has 100000 vertices and 2.3 million edges.

As we can see on Figure 3, the Monte Carlo based simulations of the GC model (CS and ES) are able to estimate the reference distribution well, with the measured error between $0.01 - 0.03$. The other two heuristics (NBH and ALE) are tailored to small edge infection probabilities with rare infections, hence they do not perform so well on this dataset. Note, that in some cases even an error of this magnitude is acceptable, and the time complexity of these methods allows them to handle larger networks. If we compare the results computed by using different attribute functions, we can see that they have minimal effect on the accuracy of the methods.

Our results on the running times[6] of these heuristics on small networks correspond with our previous findings [2]. The speed of the simulations are governed by the sample size. Complete Simulation is considerably faster than ES[7] because the latter focuses on the fast computation of smaller infections. By decreasing the sample size CS can tackle larger networks as well. The Neighborhood Bound Heuristic is able to compute the a posteriori infections of the largest networks within a minute, enabling our method to scale upwards and handle real-life datasets and networks with possibly millions of nodes and edges.

We can conclude, that in general, the use of Complete Simulation is recommended. Both its precision and accuracy are good on large graphs. If the infection probabilities are lower than our current dataset, the use of Edge Simulation is also advisable. The Neighborhood Bound Heuristic and the Aggregated Linear Effect model are able to handle even larger networks, yet this comes at the cost of a significantly lower precision.

D. Number of patterns

In many real-life applications the a priori or a posteriori probabilities of infections are unavailable. In this section we are going to assume, that the initial infection probabilities are given, but we only have a small number of observations on the a posteriori infections. We are going to simulate this on a small network by generating an a posteriori distribution using CS with $k = 1, \ldots, 10$, corresponding to $1, \ldots, 10$ observations.

We can see, that the proposed method gives a rough estimate of the vertex infection probabilities in only a few iterations. If we consider a threshold of 0.15 as an acceptable estimation, our method only requires 6 observations to reach it. However, it is important to keep in mind, that the method tries to create a posteriori infections close to the reference. The problem is underdetermined even with exact possibilities of vertex infection, with only a handful number of observations many attribute function configurations (and edge weights) may result in the same infection. The results in this section only imply, that our method is able to give one of these.

Different infection heuristics are shown on Figure 6, one can see, that the simulations have identical performance regardless

[6]We have implemented the methods in JAVA, and we have used a computer with an Intel i7-2630QM processor, and 8 gigabytes of memory.

[7]Note the sample sizes.

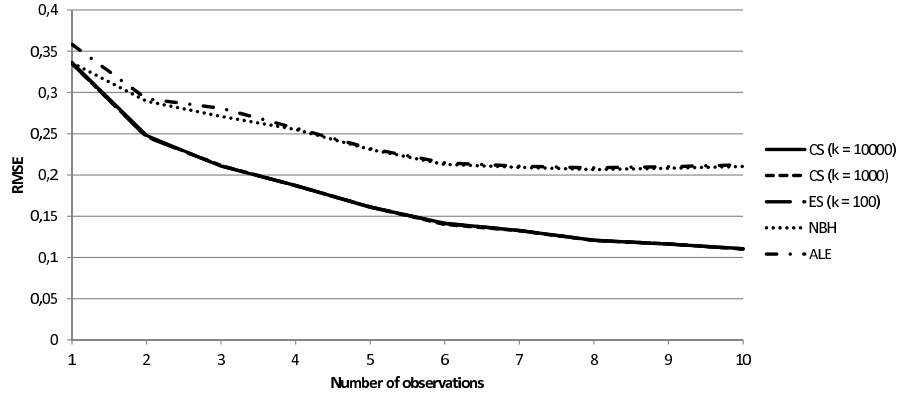


Fig. 6. The precision of the infection heuristics with a limited number of observations of the reference distribution.

of the sample size. As before, the non-Monte Carlo based methods perform poorly, their use is not recommended with low-quality inputs.

V. Conclusions

Our goal on this paper was to extend the previous results of the Inverse Infection Problem and its solution. We have given a detailed description and analysis of the Generalized Cascade Model, the Inverse Infection Problem and a Particle-Swarm Optimization algorithm capable of giving a good estimation of the latter. Several aspects of the method were investigated: We have tested the stability and accuracy of the optimization method, we have given a general approach to choose the correct attribute functions, we have examined the implications of choosing between the heuristics of the GC model and we have tested our method in low-quality inputs as well.

The given method is able to accurately predict the edge infection probabilities in a small number of iterations while the number of attributes and the shape of the attribute functions have only a small effect on this. Our method also handles low-quality inputs well. Of the infection heuristics we recommend the use of Complete Simulation, because it gives accurate results with acceptable standard deviation in reasonable time.

In our previous paper we have also given an application of this method in the prediction of credit default [7]. Our method was able to predict the default of the worst 5% of clients accurately.

Acknowledgment

The first and second authors were supported by the European Union and co-funded by the European Social Fund. Project title: "Telemedicine-focused research activities on the field of Mathematics, Informatics and Medical sciences. Project number": TÁMOP-4.2.2.A-11/1/KONV-2012-0073

The third author was supported by the European Union and the European Social Fund through project FuturICT.hu (grant no.: TÁMOP-4.2.2.C-11/1/KONV-2012-0013).

References

[1] A. Bóta, M. Krész and A. Pluhár, Applications of the Inverse Infection Problem on bank transaction networks. Submitted.

[2] A. Bóta, M. Krész and A. Pluhár, Approximations of the Generalized Cascade Model. *Acta Cybernetica* **21** (2013) 37–51.

[3] A. Bóta, M. Krész and A. Pluhár, Systematic learning of edge probabilities in the Domingos-Richardson model. *Int. J. Complex Systems in Science* Volume **1(2)** (2011) 115–118.

[4] Tianyu Cao, Xindong Wu, Tony Xiaohua Hu and Song Wang, Active Learning of Model Parameters for Influence Maximization. *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, eds. Gunopulos et al., Springer Berlin/Heidelberg, (2011) 280–295, http://dx.doi.org/10.1007/978-3-642-23780-5_28.

[5] Wei Chen, Yifei Yuan and Li Zhang, Scalable Influence Maximization in Social Networks under the Linear Threshold Model. *Proceeding ICDM '10 Proceedings of the 2010 IEEE International Conference on Data Mining*, IEEE Computer Society (2010) 88–97, http://dx.doi.org/10.1109/ICDM.2010.118.

[6] Wei Chen, Chi Wang and Yajun Wang, Scalable Influence Maximization for Prevalent Viral Marketing in Large-Scale Social Networks. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM (2010) 1029–1038, http://doi.acm.org/10.1145/1835804.1835934.

[7] A. Csernenszky, Gy. Kovács, M. Krész, A. Pluhár, T. Tóth, The use of infection models in accounting and crediting. *Challenges for Analysis of the Economy, the Businesses, and Social Progress* Szeged (2009) pp. 617–623.

[8] M. H. DeGroot Reaching a Consensus. *Journal of the American Statistical Association*, **69** (345): 118–21, http://www.tandfonline.com/doi/pdf/10.1080/01621459.1974.10480137.

[9] O. Diekmann, J. A. P. Heesterbeek, Mathematical epidemiology of infectious diseases. Model Building, Analysis and Interpretation. *John Wiley & Sons,* 2000.

[10] P. Domingos, M. Richardson, Mining the Network Value of Costumers. *Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining*, ACM (2001) 57–66, http://doi.acm.org/10.1145/502512.502525.

[11] A. Goyal, F. Bonchi, L.V.S. Lakshmanan Learning influence probabilities in social networks. *Proceedings of the third ACM International Conference on Web search and data mining.* ACM (2010) 241–250, http://doi.acm.org/10.1145/1718487.1718518.

[12] M. Granovetter, Threshold models of collective behavior. *American Journal of Sociology* **83**(6) (1978) 1420–1443, http://psycnet.apa.org/doi/10.1086/226707.

[13] J. Kennedy Particle Swarm Optimization. *Encyclopedia of Machine Learning*, Springer US (2010), 760–766, http://dx.doi.org/10.1007/978-0-387-30164-8_630.

[14] J. Kennedy, R. Mendes Neighborhood topologies in fully informed and best-of-neighborhood particle swarms. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews.* **36** (4) (2006) 515–519, http://dx.doi.org/10.1109/TSMCC.2006.875410.

[15] D. Kempe, J. Kleinberg, E. Tardos, Maximizing the Spread of Influence though a Social Network. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM (2003) 137–146, http://doi.acm.org/10.1145/956750.956769.

[16] D. Kempe, J. Kleinberg, E. Tardos, Influential Nodes in a Diffusion Model for Social Networks. *Proceedings of the 32nd International Colloquium on Automata, Languages and Programming (ICALP)*, Springer-Verlag (2005) 1127–1138, http://dx.doi.org/10.1007/11523468_91.

[17] M. Kimura, K. Saito, Tractable models for information diffusion in social networks. *Knowledge Discovery in Databases*, Lecture Notes in Computer Science Springer Berlin / Heidelberg, (2006), 259–271, http://dx.doi.org/10.1007/11871637_27.

[18] J. Leskovec, J. Kleinberg, C. Faloutsos, Graphs over time: densification laws, shrinking diameters and possible explanations. *Proceedings of the1th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM (2005) 177–187, http://doi.acm.org/10.1145/1081870.1081893.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 85–91
DOI: 10.15439/2014F495
ACSIS, Vol. 2

An unorthodox view on the problem of tracking facial expressions

Magdalena Błażek, Maria Kaźmierczak
University of Gdansk
Bażynskiego Str. 4, 80-952 Gdańsk, Poland
Email: psymb@ug.edu.pl, Email: psymk@ug.edu.pl

Artur Janowski
University of Warmia and Mazury
in Olsztyn
Michała Oczapowskiego Str. 2, 10-719 Olsztyn, Poland
Email: artur.janowski@geodezja.pl

Katarzyna Mokwa, Marek Przyborski, Jakub Szulwic
Gdansk University of Technology
Gdańsk, Narutowicza Str. 11/12, 80-233 Gdańsk, Poland
Email: katarzyna.a.mokwa@gmail.com, Email: marek.przyborski@pg.gda.pl, Email: jakub.szulwic@geodezja.pl

***Abstract*—Recent developments in imaging cameras has opened a new way of analyzing facial expression. We would like to take advantage from this new technology and present a method of imaging and processing images of human face as a response to the particular stimuli. The response in this case is represented by the facial expressions and the stimuli are still images representing six basic emotions according to Eckmann. Working hypothesis of presented research, states that the new method of tracking facial expressions is more precise and distinctive enough to give characteristic description of the analyzed human face. The biggest advantage of the presented method, in the opinion of research team, is the fact that it uses remote sensing techniques and presents dynamics of the changes happening on the human face. Therefore, FMRI might not be required, which decreases the costs of experiments, additionally, method is less stressful for the examined persons and provides more natural reactions.**

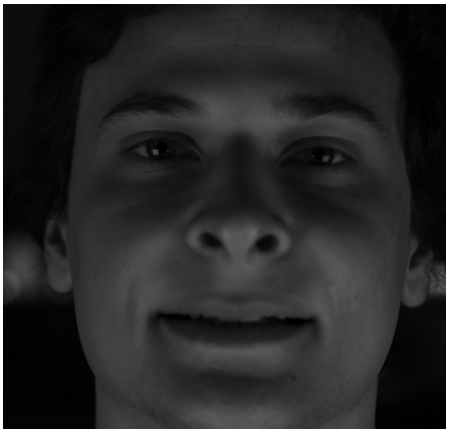
I. Introduction

SCIENTISTS all over the world are looking for new methods of tracking facial expressions to use in the field of psychology, neuroscience and affective computing. Emotion influence all modes of human activity, communication, interpersonal relations, family and business life. There are lots of empirical models of affect. Basically they are focused on analysis of speech, visual and biophysiological signals (see for example [1], [2], [3], [4], [5]. In this paper we would like to focus attention on visual methods. Typically, these methods are based on comprehensive description of the changes in facial expressions, discernment of action units for all visually distinguishable facial movements and mapping them with basic emotion (based on Facial Action Coding Systems [6], [7]). Characteristic points, located in specific segments of the human face (mouth, eyebrow, eyes etc.) are tracked and their movement is identified as a certain type of emotion [8], [9], [10]. However in our opinion, emotions represents extremely complicated states of the human's brain and they have rather fuzzy nature, thus tracking changes in the position of only few characteristic points might not be effective. Due to the nature of the observed phenomenon there is no simple method of distinguishing certain type of emotion among diversity of emotional states. Fig. 1 shows how complicated are representations of emotional states when they are expressed by human beings. Taking into account the movement of the lips or eyebrows we are not able to distinguish in what emotional state the tested person is. In addition, presenting on Fig. 1 images are taken when the emotion is the most intensive (for example the moment when the smile is the widest and brightest). We have decided to create our own database of facial expressions because we would like to investigate the process of creation facial expression from the beginning to the end (when the face comes to neutral state). Data gathered in the new database let us track changes on the face throughout the period of observation.

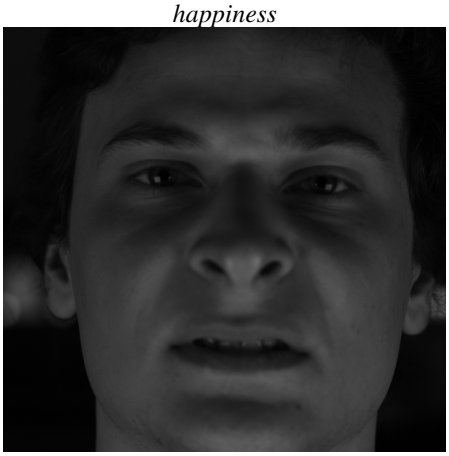
The caption of the Fig. 1 is a starting point for the further investigations. Humans ability to recognize different emotions is very difficult and complicated process to implement in the machine like computer. Recognizing emotional states is important for human-computer interaction or other form of coexisting in an artificial environment. In the opinion of the research team., dynamics of the facial expressions is the key to find the efficient way of recognizing emotional states. The main goal of presented experiment was to develop a new method of tracking changes in facial expressions, because, in the opinion of research group, it might be a promising direction to develop fully automatic method of recognition emotional states. Proposed method opens a long way to find the solution of the problem how emotions are expressed and how they arise.

Recent developments in imaging cameras has opened a new way of analyzing facial expression of emotions [11]. We would like to present a new method of imaging and processing images of human face in order to quantify the response to the particular stimuli [12]. This article presents a method of processing images of human face recorded by a fast camera Phantom MIRO 310. Method was applied on a collection of 65 persons (30 female students and 35 male students) expressing happiness, sadness, anger, fear and disgust, as well as surprise.

978-83-60810-58-3/$25.00



happiness



surprise

Fig. 1. 2 of 6 basic emotions expressed by research participant, it is noticeable that there are no clear limits between emotional states.

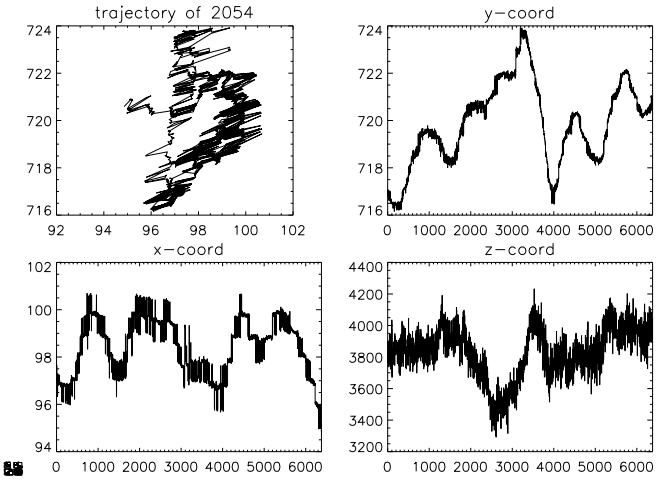


Fig. 2. Single object trajectory, $Z - coord$ - represents brightness

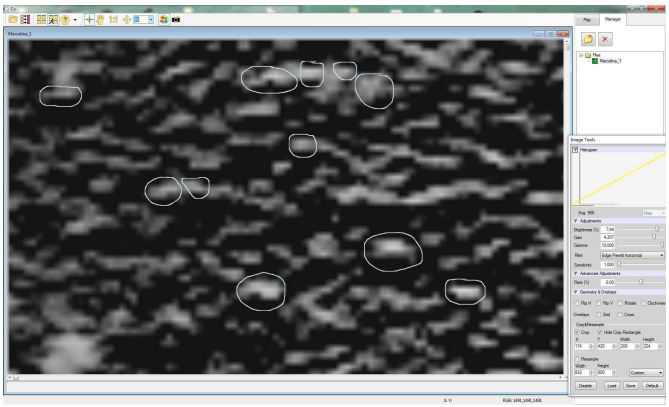


Fig. 3. Objects on the human face, characteristic for the certain structure of the skin

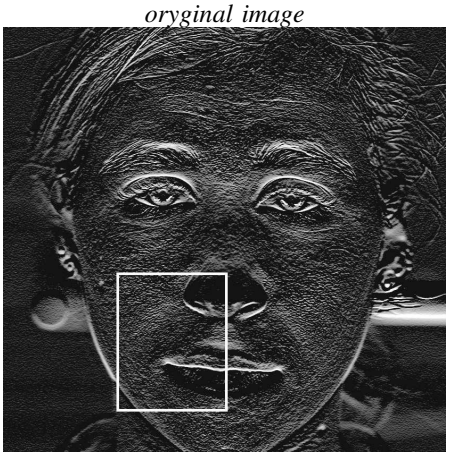
Each emotion was expressed be the students for about 2 seconds. Resulting images represent series of frames recorded at the speed of 1000 frames per second. Presented method is based on the assumptions coming from Particle Image Velocimetry (PIV) [13], [14]. The most important one is that there are some objects on the human face, characteristic to the certain type of skin (see Fig. 4) (as well as certain mechanical and chemical properties of the skin), that are determined and identified while processing series of images. Those objects are of non-spherical shape (as it is presented on the Fig. 3). Working hypothesis of presented research states that the facial expressions are determined by the mechanical and chemical properties of the human's skin and the tension of the muscles, however all those properties might influenced the motion of the identified objects only in very limited range, as it is presented on the Fig. 2.

The following sections are the attempt to the answer the question whether the working hypothesis is true or false.

The main idea of proposed research comes from the methods of studies colloidal particle's trajectories and particle image velocimetry. While both image processing, photographic, study of colloidal suspensions and particle image velocimetry are well developed fields, applying them to the field of affective computing and psychology gives new view on the problem of facial expressions what in consequence might constitute new approach to the problem of automatic recognition of human emotions [15], [2]. The methods described below are generalized to the objects or areas of non-spherical shape that can be identified on human face (see Fig. 3). Shape and position of non-spherical objects located on the face plays very important role in this method. Thanks to the speed of the recording (1000/frames per second) collected positions of the identified objects create trajectories of their movement.

In the following, section typical instrumentation required for collecting digital video images of human face has been described, and some details of the steps required to convert a digital movie into an ensemble of single-object trajectories. We stress those aspects of the analysis which allow us to track particular objects on the human face. High-resolution trajectory data makes possible a wide range of quantitative measurements of the changes occurring on human face.

oryginal image



processed image

Fig. 4. An example of recorded frames of human face

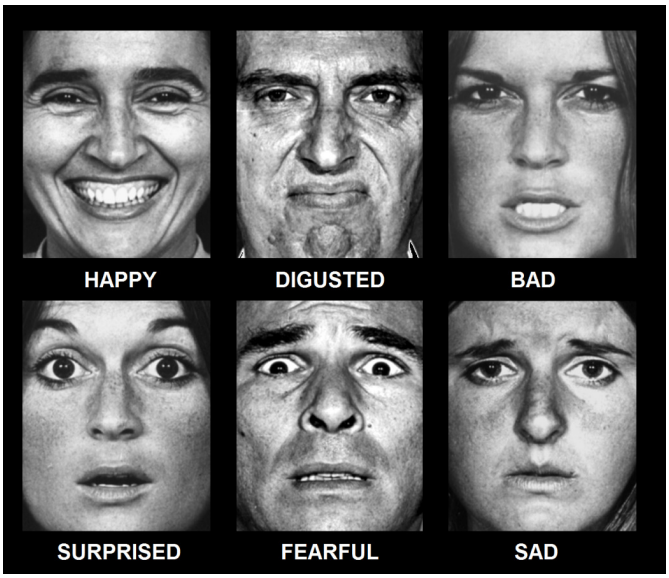


Fig. 5. Standard representations of the basic emotions

II. METHODS

Research was conducted in the Gdansk University of Technology, Remote Sensing Lab on 30 female and 35 male, students of the Psychology at the Gdansk University. The study was based on the assumption that students have to reproduce emotion which was showed them on the laptop's screen. Our team has prepared PowerPoint presentation based on the set of facial expression images proposed by Professor Eckman (see Fig. 5). Each emotion was earlier presented to the examined person and then on the mark he or she should reproduced it straight to the camera. It seems necessary to stress that the stimuli - picture of certain emotion wasn't active when examined person was asked to reproduced the image, because the eyes of the research participant were directed straight at the camera, not on the computer screen with the image of emotion pattern. In fact, only the imagination of the image is used to render the emotion.

Standard commercial video cameras produce 30 complete images per second but in order to track changes on the human face we need speed camera with ability to record at least 1000 complete images per second. Monochrome CCD cameras may be preferable to color models not only because they are less expensive but also because they tend to have superior noise figures and greater sensitivity to subtle brightness variations. Color information, furthermore, is not used in the techniques we describe below. We use an Phantom Miro 310 camera connected to the Dell m4700 mobile workstation. To apply Particle Images Velocimetry method we have to transform an image of human face in to map of small objects like it is presented on Fig. 6

map of small objects

Fig. 6. Zoom of the ROI from Fig. 4

At this stage of the processing the filtering methods are very helpful, in this particular case Prewitt filter has been used, which was implemented in the Phantom Miro 310 camera software. The following parameters has been used in this experiment:

- brightness of the image set to 7.44,
- gain factor set to 4.207,
- gamma factor set to 10.00.

By using camera software it was possible to collect series of frames from the recorded movie already transformed into a map of small objects, and represents them as a series of tiff images directly from the camera without loosing the time for additional computations and transformations, an example is presented on the Fig. 6. After this operation a series of images, about 2000 frames for each of 6 basic emotions, have been created. Every image represents face as a map of small objects that move while facial expression (typical for the certain type of emotion) come out on the face (see Fig. 6).

Digital video analysis and methods of tracking and identifying objects, enables extracting trajectories of individual micro-objects from a series of images. The time evolution of the distribution of interested objects (or areas) can be calculated according to the following equation

$$\theta(r,t) = \sum_{i=1}^{N} \delta(r - r_i(t)), \tag{1}$$

where:

- $r_i(t)$ is the location of the i-th object in a field of N objects at time t,
- δ - distance.

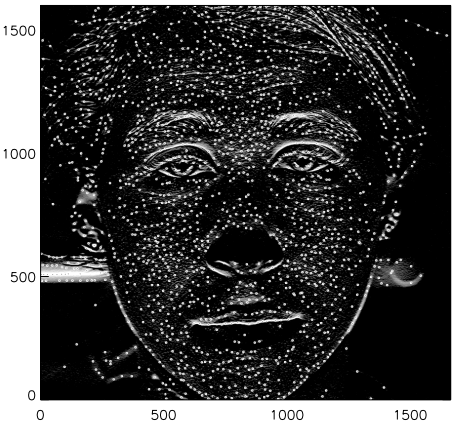
III. Results.

The process of extracting useful information from a sequence of digital images consists of five logical steps:

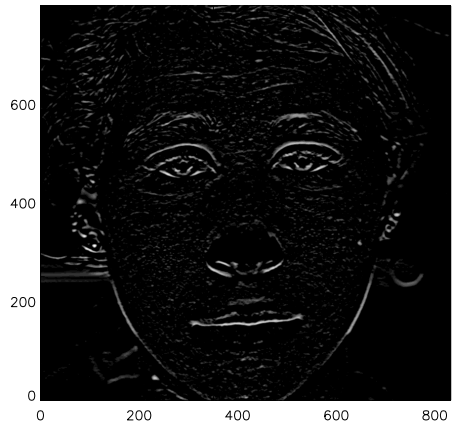
1) correcting imperfections in the individual images,
2) locating candidate particle positions,
3) refining these positions,
4) discriminating „false" objects,
5) and finally linking the time-resolved object locations into trajectories.

In this case, the tracking algorithm has to deal with the changing appearance of the interested areas as they move. Furthermore, it has to distinguish marginally focused particles from noise. In practice, the image $A(x, y)$ must be cast from an array of bytes to a higher precision data format, such as a floating point array, before convolution. We have used an object oriented method for identifying candidate object within an image as a local brightness maxima's. The pixel is assigned as a candidate if no other pixel within a certain distance w is brighter. Because only the brightest pixels correspond to the location of interested us objects on the human face, we further require candidates to be in the upper 50th percentile of brightness for the entire image, see Fig. 7.

This operation is called, the gray-scale dilation [16], [17] and provides an implementation of the regional maximum



Identified objects



Objects of the same brightness

Fig. 7. Non-spherical areas (objects or particles) on the human face after processing. Objects represents the same spectral feature - brightness

selection criterion. Gray-scale dilation is an elementary morphological operation which sets the value of pixel $A(x, y)$ to the maximum value within a distance w of coordinates (x, y). A pixel in the original image which has the same value in the dilated image is then a candidate particle location. We use the same value of w as was used in the filtering step.

Having already found a locally brightest pixel at (x, y), which presumably is near the geometric center of an interested us object's at the coordinates x_0, y_0 , we calculate the offset from (x, y) to the brightness-weighted centroid of the pixels in a region around (x, y) according to following equation:

$$\begin{pmatrix} \epsilon_x \\ \epsilon_y \end{pmatrix} = \frac{1}{m_0} \sum_{i^2+j^2 \le w^2} \begin{pmatrix} i \\ j \end{pmatrix} A(x+i, y+j), \tag{2}$$

where:

- moment $m_0 = \sum_{i^2+j^2 \le w^2} \binom{i}{j} A(x+i, y+j)$ is the integrated brightness of the sphere's image.

Having located objects or particles in a sequence of images, the next step is to match up locations in each image with corresponding locations in later images to produce the trajectories. This requires determining which particle in a given image most likely corresponds to one in the next image. Tracking more than one object is very difficult. Thus, it is necessary to seek the most probable set of N identifications between N locations in two consecutive images. Linking particle distributions into trajectories is only feasible if the typical single particle displacement in one time step is sufficiently smaller than the typical inter-particle spacing, otherwise, particle positions will become inextricably confused between snapshots. In this case the speed of recording (1000/frames per second) let us discriminate very small displacement between objects, and thus the trajectories are more accurate and precise. This process is repeated for the particle locations in each frame until is completely determined. The examples of calculated trajectories are presented on Fig. 8.

One of the most important aspects of particle dynamics in the fluid flow, which underlies the proposed method, is the mean squared particle displacement while the time of observation evolve. This quantity can be calculated according to the Eq. 3. On the Fig. 9 we would like to present results of calculating msd for the few members of the research group. The $Panel\ \ A$ represents $MSD\ \ curves$ calculated for 6 basic emotions reproduced by participant according to the images showing on the laptop's screen.

$$MSD(\tau) =< \triangle r(t)^2 >=< [r(t+\tau) - r(t)]^2 > \qquad (3)$$

where:

- $r(t)$ particle position at time t,
- τ lag time between two positions,
- operator $<>$ designates a time-average over t and/or an ensemble-average over several trajectories.

Figure 9 $Panel\ \ B$ presents the examples of calculating $MSD\ \ curves$ for selected number of representatives of the research group. Each curve for each person, is an average of 6 basic emotion presenting on the $Panel\ \ A$. In order to keep the figure as readable as possible only 3 MSD-curves are drawn. The most important conclusion from this figure states, that it is possible to see differences between particular curves (Fig. 9 $Panel\ \ B$). The second conclusion - shape of the MSD is similar for examined persons. However, in our opinion, it depends on the individual features of human skin. As it was mentioned in the Introduction in our opinion, mechanical and chemical properties of the skin plays an important role in this phenomenon. The next conclusion is, that noticeable differences between different emotional states or different person starts from about $t = 100$ of time of observation. While watching video of recorded persons it is noticeable that the facial expression appears after a few moments of inactivity. This behavior has its own mark in the results of calculating MSD. And the last conclusion which in opinion of the research group seems to be very important, especially in the field of neural science and psychology, there is some activity on the human face even before the facial expression appears (at the Fig. 9 segment from $t = 0$ till $t = 100$). We believe that human brain express the certain emotion before the face is able to showed it. This conclusion may acknowledge recent discoveries of Professor Laeng in the field of imagination and perception.

IV. Conclusion

The preceding sections describes image analysis methods we have developed to perform quantitative time-resolved imaging studies of non-spherical objects (or areas) on the human face which we have found useful in the process of tracking facial expressions.

We have shown that the trajectories of the identified areas on the human face are distinctive enough to find differences between persons taking part in the research.

The measurable effect of the presented in the article transformations of the collected series of images is the collection containing the identification number of the identified areas (same size and brightness) on the face and the trajectories of their movement throughout the observation period.

The study of that phenomenon is based on the assumption that there are objects on the human face and they movement depend on the muscles tension, mechanical and chemical properties of the skin. Presented method may open the way to a much larger and broader question of the automatic identification of emotional states. It gives the possibility to find quantitative description of the observed changes of emotional states happening on human face [18], [19] what may have big value for the security systems.

Recent research on the connections between imagination and perception suggest that unconditional reflex narrowing of the pupil and caused by changes in the amount of light reaching them, can be elicit also by only imagining the appropriate situation. Usually people treat ideas as private and subjective experience, which is not accompanied by significant physiological changes. The results achieved by the Professor Laeng team and presented in [20] challenge this view. They suggest that imagination and perception are based on similar sets of neural processes. Presented in this article experiment on tracking facial expressions also confirms their thesis.

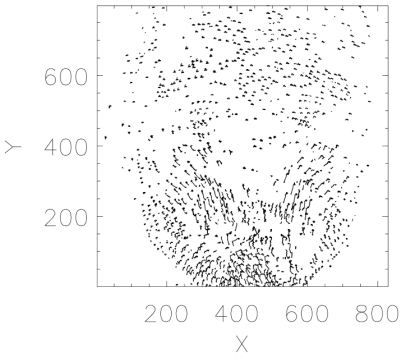
References

[1] R. W. Picard, S. Member, E. Vyzas, and J. Healey, "Toward Machine Emotional Intelligence : Analysis of Affective Physiological State," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

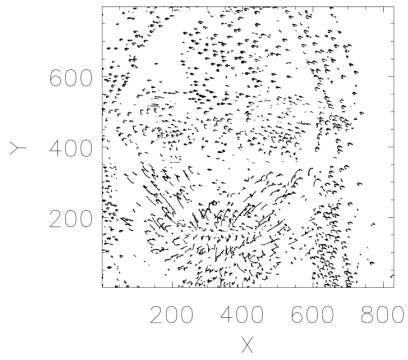
[2] D. Ghimire and J. Lee, "Geometric Feature-Based Facial Expression Recognition in Image Sequences Using Multi-Class AdaBoost and Support Vector Machines." *Sensors (Basel, Switzerland)*, vol. 13, no. 6, pp. 7714–34, Jan. 2013. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3715259\&tool=pmcentrez\&rendertype=abstract

[3] G. Guo, R. Guo, and X. Li, "Facial Expression Recognition Influenced by Human Aging," *IEEE Transactions on Affective Computing*, vol. 4, no. 3, pp. 291–298, Jul. 2013. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6517179

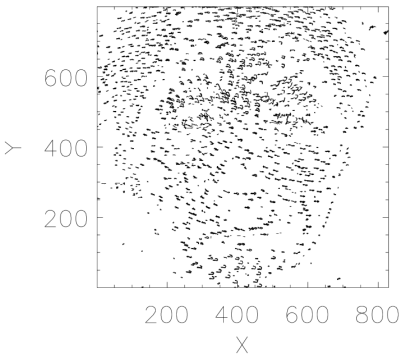
[4] D. Heylen, M. Ghijsen, A. Nijholt, and R. D. Akker, "Facial Signs of Affect During Tutoring Sessions," pp. 24–31, 2005.



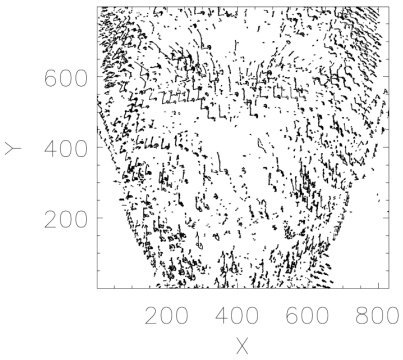
Person 1



Person 2

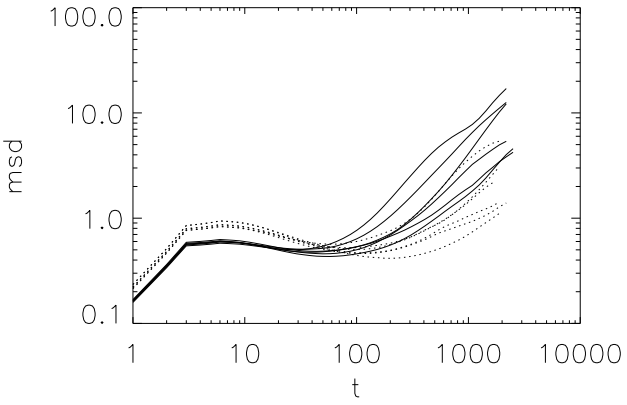


Person 3

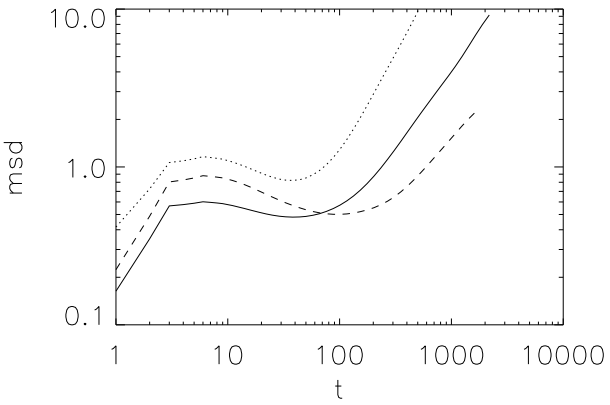


Person 4

Fig. 8. Trajectories of the identified objects on the faces of 4 representatives tested group. Each person reproduced the same emotion - happiness



Panel A



Panel B

Fig. 9. Mean squared displacement.

[5] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, "ECG Pattern Analysis for Emotion Detection," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 102–115, Jan. 2012. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5999653

[6] P. Ekman and J. C. Hager, "Computer Measurement of Sign Vehicles in Body Movement and Facial Expression." [Online]. Available: http://face-and-emotion.com/dataface/misctext/iwafgr.html

[7] P. Ekman, W. V. Friesen, and J. C. Hager, "Facial Action Coding System - Title Page," 666 Malibu Drive, Salt Lake City UT 84107, Tech. Rep., 2002. [Online]. Available: http://face-and-emotion.com/dataface/facs/manual/TitlePage.html

[8] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, Jan. 2003. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0031320302000523

[9] M. Pantic and M. S. Bartlett, *Machine Analysis of Facial Expressions*, 2007, no. June.

[10] T. Pfister, X. Li, G. Zhao, and M. Pietik, "Recognising Spontaneous Facial Micro-expressions."

[11] R. C. Gur, R. Sara, M. Hagendoorn, O. Marom, P. Hughett, L. Macy, T. Turner, R. Bajcsy, A. Posner, and R. E. Gur, "A method for obtaining 3-dimensional facial expressions and its standardization for use in neurocognitive studies." *Journal of neuroscience methods*, vol. 115, no. 2, pp. 137–43, Apr. 2002. [Online]. Available:

http://www.ncbi.nlm.nih.gov/pubmed/11992665

[12] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias, "Emotion Analysis in Man-Machine Interaction Systems," in *in Proc. MLMI, LNCS 3361*, 2005, pp. 318–328.

[13] M. Raffel, C. E. Willert, and J. Kompenhans, *Particle Image Velocimetry: A Practical Guide ; with 24 Tables*, 1998. [Online]. Available: http://www.google.pl/books?hl=pl\&lr=\&id=enOLTmfYVPQC\&pgis=1

[14] A. Schroeder, C. E. W. Eds, D. P. J. Barz, H. F. Zadeh, and P. Ehrhard, *Particle Image Velocimetry*, topics in ed. Springer - Verlag GmbH, 2008.

[15] Y. Cao, W. Zheng, L. Zhao, and C. Zhou, "LNCS 3784 - Expression Recognition Using Elastic Graph Matching," *Lecture Notes in Computer Science*, vol. 3784, pp. 8–15, 2005.

[16] A. K. Jain, *Fundamentals of digital image processing*, 1989. [Online]. Available: http://books.google.pl/books/about/Fundamentals_of_digital_image_processing.html?id=GANSAAAAMAAJ\&pgis=1

[17] W. K. Pratt, *Digital Image Processing*. New York, USA: John Wiley & Sons, Inc., 2001. [Online]. Available: http://doi.wiley.com/10.1002/0471221325

[18] R. W. Picard, "Emotion Research by the People, for the People," *Emotion Review*, vol. 2, no. 3, pp. 250–254, Jun. 2010. [Online]. Available: http://emr.sagepub.com/cgi/doi/10.1177/1754073910364256

[19] J. Tao, T. Tan, R. Picard, A. Choi, and W. Woo, *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, A. C. R. Paiva, R. Prada, and R. W. Picard, Eds. Springer Berlin Heidelberg, 2007, vol. 4738, no. September. [Online]. Available: http://www.springerlink.com/index/10.1007/978-3-540-74889-2

[20] B. Laeng and U. Sulutvedt, "The eye pupil adjusts to imaginary light." *Psychological science*, vol. 25, no. 1, pp. 188–97, Jan. 2014. [Online]. Available: http://pss.sagepub.com/content/25/1/188

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 93–100

DOI: 10.15439/2014F194
ACSIS, Vol. 2

Experimental evaluation of selected tree structures for exact and approximate *k*-nearest neighbor classification

Aleksander Cisłak
Technical University of Munich,
Department of Informatics,
Boltzmannstr. 3, D-85748 Garching, Germany
Email: a.cislak@tum.de

Szymon Grabowski
Lodz University of Technology,
Institute of Applied Computer Science,
Al. Politechniki 11, 90–924 Łódź, Poland
Email: sgrabow@kis.p.lodz.pl

***Abstract*—Spatial data structures, for vector or metric spaces, are a well-known means to speed-up proximity queries. One of the common uses of the found neighbors of the query object is in classification methods, e.g., the famous *k*-nearest neighbor algorithm. Still, most experimental works focus on providing attractive tradeoffs between neighbor search times and the neighborhood quality, but they ignore the impact of such tradeoffs on the classification accuracy.**

In this paper, we explore a few simple approximate and probabilistic variants of two popular spatial data structures, the k-d tree and the ball tree, with *k*-NN results on real data sets. The main difference between these two structures is the location of input data — in all nodes (k-d tree), or in the leaves (ball tree) — and for this reason they act as good representatives of other spatial structures. We show that in several cases significant speedups compared to the use of such structures in the exact *k*-NN classification are possible, with a moderate penalty in accuracy. We conclude that the usage of the k-d tree is a more promising approach.

I. Introduction

FINDING objects similar to a given one in a large database is a classic research topic, with applications in pattern recognition, multimedia processing, genomic analyses, and other fields. There exist many particular variants of the problem, but one of the most popular is: given object x, we wish to find its k nearest neighbors in a given database D of size n, according to the specified similarity measure. The parameter $k \geq 1$ is usually selected at query time. A naïve solution to this problem is to calculate the distances between the query x and all objects in the database and choose k nearest ones, but this approach requires computation of n distances. If database preprocessing is allowed, we can usually reduce the query time. One of major applications of the proximity search is *classification*, when the query sample is assigned a class label according to the known class labels of its neighbors, and the rest of this paper is focused on this application.

We assume a vector space, in which objects are identified with d real-valued vectors (tuples). The distance function in this space is usually a metric (i.e. it satisfies non-negativity, identity of indiscernibles, symmetry, and the triangle inequality), and the most common particular metrics used are the Euclidean or Manhattan (city-block) one. In vector spaces, the popular search structures include the k-d tree, R-tree, quadtree, X-tree, and their numerous variants. Their common trait is to cluster objects in space, to allow pruning the dataset during most queries. For example, the popular k-d trees partition the space along different coordinates while R-trees group objects in hyperrectangles.

As the (in)famous curse of dimensionality subdues the performance of practically any (however sophisticated) nearest-neighbor finding data structure in high dimensions, it is interesting to investigate how approximate or probabilistic variations of the true nearest neighborhood of the given query affect the classification accuracy. This question has met significant interest from both theoreticians and practitioners, see for instance Arya et al. [1], Indyk and Motwani [2], or Jones et al. [3].

In this work we introduce simple modifications to well-known spatial data structures: the k-d tree and the ball tree, in order to explore how approximate or probabilistic speedup idea (e.g., via more aggressive pruning than in the original method) affect the time-accuracy tradeoff. While our conclusions are hardly definite, we believe that experimentations with popular (and relatively easy to implement) data structures have their own, practice-oriented, value.

II. K-d tree

One of the oldest spatial data structures, the *k-d tree*, was introduced by Jon Louis Bentley in 1975 [4], and the name refers to k dimensions it operates on. To avoid confusion with the number of neighbors in the k-NN rule, from now on we will use the symbol d for the number of dimensions.

The k-d tree is a binary tree, where every instance of the indexed data corresponds to one node. The left child together with its descendants contain points whose values of the feature (coordinate) f are smaller than the f value of the splitting hyperplane H — analogously, the right child together with its descendants contain points with higher f values. As regards the selection of H, the most popular approach is to choose the point whose f is the median, and divide the points into

978-83-60810-58-3/$25.00

two parts of equal size (assuming that the number of points to divide is even).

A. Construction of the tree

During the construction, the current feature space is recursively divided into two subspaces, with half of the points lying in each subspace. This division is based on the current dimension, and the algorithm switches to the next dimension with each step as the recursion progresses. After all dimensions have been processed, it goes back to the first dimension in a circular manner. The recursion stops when there is only a single point left, and this point is stored in a leaf. The result is a binary tree, where inner nodes represent points situated on the splitting hyperplanes, and leaves represent the rest of the given data.

B. K-nearest neighbor search

When the *k*-NN search is performed, the tree is traversed from the root to the leaf. The algorithm goes left or right depending on feature values, and this can be represented by following relations, where Q is the queried point, N is the point corresponding to the current node, and d is this node's split dimension: $Q_d \leq N_d \rightarrow left$, $Q_d > N_d \rightarrow right$. Dimensions are switched in the same way as during the construction, so that the dimension which is checked at each level is always the one on which the space was split in halves.

After a leaf has been found, the search goes back towards the root, following the same path which was traversed downwards. From this moment, the algorithm maintains the list of k points with smallest distances to the queried point Q, and tries to update it every time a new node is visited.

At each step upwards, there is a possibility of inspecting a subtree whose root is the sibling of the current node N_{cur}. Such a subtree can be pruned if and only if k points have already been found, and all distances from Q to these k points are smaller than or equal to the distance between Q and the point P_{spl}. P_{spl} represents the point located on the splitting hyperplane, and it is associated with the node which is the parent of N_{cur}. This is demonstrated by the relation in Figure 1, where P represents the set of points found so far, and D refers to the distance.

$$prune \leftrightarrow |P| = k \wedge \bigvee_{p \in P} D(p, Q) \leq D(Q, P_{spl})$$

Fig. 1. Pruning condition in an exact k-d tree.

If the subtree S could not be pruned and it has been checked, the list of best points from S must be merged with the current list of best points. This is rather straightforward, because we simply select k points with lower distances from both lists, or if the size of the combined list would be smaller than k, all points are retained.

The whole *k*-NN search procedure can be summarized as follows:

1) Find the leaf.
2) Go to the parent and try to update the list of k best points.
3) Recursively check the subtree whose root is the sibling of a current node, unless the relation in Figure 1 is satisfied.
4) If checked the subtree, merge the lists of best points.
5) Repeat 2. until found the root of the whole tree.

C. Complexity

As regards search time complexity, the average case for the nearest neighbor lookup (1-NN), under favorable assumptions (discussed in the next sentences), is equal to $O(\log n)$ [4], and the worst case, where all points are checked, is clearly equal to $O(n)$. Performance degrades to linear time when, roughly speaking, the number of dimensions is large, and for this reason the number of visited nodes also tends to be large. In general, for optimal performance the relation $2^d \ll n$ should hold [5]. When it comes to *k*-NN, the average case expands into $O(\log n \cdot \log k)$, and the worst case expands into $O(n \log k)$, assuming a heap is used to maintain the list of best points. In practice, a k-d tree might turn out to be slower than a naïve method, due to the search procedure overhead.

The construction takes $O(n \log n)$ time, assuming the median required to split points in halves is found with a linear worst-case time algorithm [6, Ch. 9]. Since the number of nodes is proportional to the number of points, the space complexity is equal to $O(n)$.

III. Ball tree

The aim of the *ball tree* is akin to the one of the k-d tree, as it attempts to reduce time spent on a nearest-neighbor query by partitioning the feature space. Just as the name suggests, this is achieved by constructing closed balls, that is geometric objects containing a sphere S and the space inside S.

The ball tree is a binary tree, where each internal node N_I is associated with one ball, and this ball contains all balls of the descendants of N_I. Hence, the biggest ball is stored as a root and it contains all other balls in the tree. The training data are stored in the leaves, with one leaf corresponding to one training instance, and internal nodes act only as guidance during the search. Subspaces resulting from the partitioning are clearly overlapping, unlike in the case of other structures, such as the aforementioned k-d tree.

A. Construction of the tree

We opt for the bottom-up construction algorithm, which is the most efficient one with respect to the search time of *k*-NN queries performed on the resulting tree [7]. This efficiency results from the fact that we try to reduce the volume (the radius) of the balls.

At the beginning of the construction, we create a set of balls from the training data, with one ball corresponding to one instance. At each step, we search for a pair of balls, whose resulting ball R_B (one that contains the selected pair of balls) is the smallest. Subsequently, two selected balls are set as children of R_B, and R_B is inserted back into the set of

available balls. Thus, at each step we reduce the size of the set by one. When we are left with only one ball, this ball is associated with the root node and the algorithm terminates.

For more details on other construction algorithms, we refer the reader to the original article by Omohundro [7].

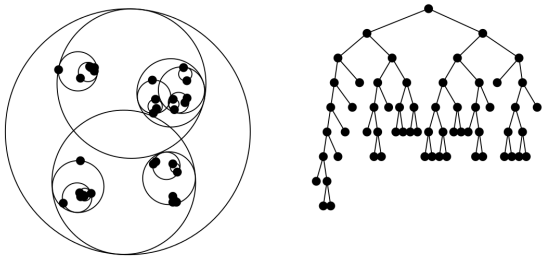


Fig. 2. Space partitioning with balls and the resulting binary tree using the bottom-up construction algorithm; reproduced from Omohundro [7].

B. K-nearest neighbor search

When the *k*-NN search is performed, the tree is traversed in the depth-first manner. At each step, there is a possibility of recursively inspecting two children of the current node, and each child can be pruned together with all of its descendants, if and only if the following condition is satisfied. For safe pruning, it is required that we have already found k points, and the ball that is centered at the query point and that contains all k points found so far does not intersect the ball of the child in question. This is demonstrated by the relation in Figure 3, where *P* represents the set of points found so far, and B_Q and B_C refer to the ball centered on the query and the ball centered on the child, respectively.

$$prune \leftrightarrow |P| = k \wedge intersects(B_Q, B_C) == false$$

Fig. 3. Pruning condition in an exact ball tree.

Whole *k*-NN search procedure can be summarized as follows:

1) If we are at the leaf, try to update the list of k nearest neighbors and terminate.
2) If k nearest neighbors have already been found, compute the ball that is centered at the query point and contains these k points.
3) Recursively inspect the left child, unless the relation in Figure 3 is satisfied.
4) If *k* nearest neighbors have already been found, recompute the ball from step 2).
5) Recursively inspect the right child, unless the relation in Figure 3 is satisfied.

C. Complexity

We construct a binary tree, whose bottom level contains n nodes, because all training data are stored as leaves. For this reason, the space complexity of the ball tree is equal to $O(n)$.

As regards the search time complexity, the worst case where all nodes are checked is clearly proportional to the size of the tree, that is $O(n)$. Because of the time overhead resulting from tree traversal, the ball tree can turn out to be slower than a brute-force algorithm. Assuming optimal space partition, where branches can be pruned, for a *k*-NN search it is possible to achieve the best case bound of $O(\log n + k)$.

Since our focus is solely on finding nearest neighbors, we ignore the preprocessing time complexity, however, it is worth noticing that it can be fairly expensive. For instance, a naïve bottom-up construction algorithm requires $O(n^3)$ time.

IV. Approximate algorithms

The objective of approximate search algorithms based on tree data structures introduced in previous sections is to decrease the time spent on classification, at the cost of an increase in the error rate. This is accomplished by limiting the visited area of a tree.

We utilize the notion of *bounds*, where after the specified bound has been crossed, current list of nearest neighbors is returned. This means that the bound should be chosen with care, since the list can actually contain less than k points when the algorithm terminates. For probabilistic pruning (Subsection IV-D), we ensure that branches are pruned only after k points have been found.

It is to be noted that time and space overheads presented in this section are relevant only to described modifications, and for a complete analysis, complexities of exact algorithms should be added.

A. CPU time bound

CPU time refers to the time spent by the processing unit on executing actual instructions, which means that it is not affected by context switches or time changes. The bound is checked for the first time after the leaf has been found, which is required for the ball tree, and allows us to concentrate the search in the bottom part of the k-d tree. Bounds checking introduces a small time overhead $O(v)$, where v refers to the number of nodes visited by the algorithm after the first leaf has been reached. The space overhead is constant.

B. Depth bound

The depth bound specifies a maximum depth of the tree that can be reached by the search algorithm. This is relevant only for the k-d tree, since in the ball tree all training data are situated in the leaf nodes. The depth is checked every time a new node is visited, and for this reason the time overhead is equal to $O(v_t)$, where v_t refers to the total number of nodes visited by the algorithm. The space overhead is equal to $O(n)$, because every node stores its depth.

C. Node bound

The idea of the node bound is simply to set a hard threshold t on the number of nodes, which can be checked by the algorithm. Analogically to the CPU time bound, this bound is checked for the first time after the leaf has been found, which

is required for the ball tree, and allows us to concentrate the search in the bottom part of the k-d tree. For this reason, the total number of nodes which have been traversed might turn out to be greater than t. Again, bounds checking introduces a small time overhead $O(v)$, where v refers to the number of nodes visited by the algorithm after the first leaf has been reached.

D. Probabilistic pruning

Similarly to approximate variants introduced in previous subsections, the aim of this algorithm is to limit the space that is inspected during the search procedure. This is achieved by introducing the pruning factor σ, which describes the probability that the subtree is pruned, even if the pruning condition presented in Figure 1 for the k-d tree or in Figure 3 for the ball tree is not satisfied. For instance, if $\sigma = 25\%$, every time a subtree should be inspected, there is a $1/4$ chance that it will be ignored instead. It should hold that $\sigma > 0 \wedge \sigma \leq 1$. It is worth noticing that in the case of $\sigma = 1$, the algorithm's behavior is in fact deterministic, as all possible branches are pruned.

Since pseudorandom number generation can be done in constant time, the time overhead is proportional to the number of pruning decisions which have to be taken. These decisions are made only when the subtree cannot be safely pruned, and the complexity is equal to $O(1)$ in the best case, since then all subtrees can be safely pruned. As regards the worst case, a decision has to be made every time a new subtree is encountered, and for this reason the time overhead expands into $O(n)$. The space overhead is constant.

E. Best bin first (BBF)

The *best bin first* (BBF) algorithm [8] is relevant only to the k-d tree and it aims to increase the accuracy of an approximate search. Since an inexact algorithm does not visit all nodes which would be required to provide an exact answer, the order in which the nodes are visited is crucial to the performance in terms of an error rate. After the leaf has been found, instead of following the path to the root from the bottom, going up one level per step, the algorithm selects an optimal node lying on this path. Subsequently, it continues to choose an optimal node from the remaining ones, until the path is exhausted, or some limit (such as the node bound) is exceeded. We choose a straightforward method to determine node's optimality, which selects the node whose splitting hyperplane is closest to the queried point [8].

The time overhead depends on the kind of priority queue that is used for selecting the smallest distance. We have $O(V_T)$ inserts and $O(t)$ delete-min operations, where V_T represents the total number of nodes traversed by the search procedure, and t is the number of nodes visited after the first leaf has been found. For instance, if the Fibonacci heap [9] were used, the complexity would be equal to $O(V_T + t \log V_T)$ amortized time. As regards the space overhead, it is possible to achieve a bound of $O(V_T)$. These complexities refer only to maintaining a priority queue and not to bounds checking.

V. Experimental results

The error rates presented in this section were calculated using the *leave-one-out* method for the k-d tree, and 5-fold *cross validation* for the ball tree, the reason for the second method being computational demands. We used the *Manhattan* metric for the similarity measure. Classification times presented in the diagrams refer to CPU time in milliseconds spent on classifying one sample, and the preprocessing time is not taken into account. CPU time values are arithmetic mean values obtained in the course of three runs, in order to minimize an influence of external factors such as cache utilization. Error rates presented for probabilistic variants are arithmetic mean values obtained in the course of five runs. The value $k = 5$ was selected arbitrarily. The machine used for experiments was equipped with Intel e2160 processor running at 2.9 GHz and 4 GB DDR2 memory. The code was compiled using the GCC suite and run on Ubuntu 12.04 64-bit operating system.

Only selected results are presented due to space constraints, nevertheless, results reported for different data sets (for a short description of the sets, see Appendix A) were consistent to a satisfactory degree. Unexpected or particularly unusual behavior was rare, and it was most probably caused by a unique structure of specific data set in question.

Selected diagrams were published in the Bachelor's Thesis of the first author [10].

A. CPU time bound

CPU time bound values are significantly smaller than the times spent on classification shown in other diagrams. This results from the fact that the execution time spent on classifying one sample is calculated as the total time used by all procedures in the *k*-NN algorithm, such as allocating the memory or comparing class counts. On the other hand, the CPU time bound refers only to the internal approximate search procedure.

Just as expected, we observed a growth in the error rate as the time bound decreased. Above certain higher bound value there was no change with respect to exact algorithms, and as the bound approached zero, there was a dramatic increase in the error rate. For the k-d tree with Ferrites data set, there was only a marginal increase in the error rate until the bound value of about 0.35 ms, as demonstrated in Figure 4. Other data sets and the ball tree behaved similarly, although relative increases in the error rate were more significant.

B. Depth bound

The depth bound introduced a rather moderate increase in the error rate, although associated time decreases were lower than in the case of other bounds. For instance, for Banknotes data set, there was almost no increase in the error rate up to the bound value of 7, with a roughly 3-fold decrease in classification time (see Figure 5). At the other extreme was the Isolet data set, for which both time decrease and error rate increase were more substantial than for other sets (see Figure 6).

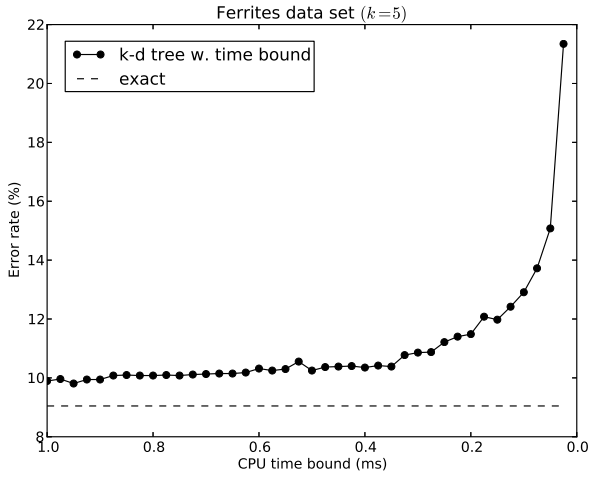


Fig. 4. Error rate vs CPU time bound for the k-d tree with Ferrites data set.

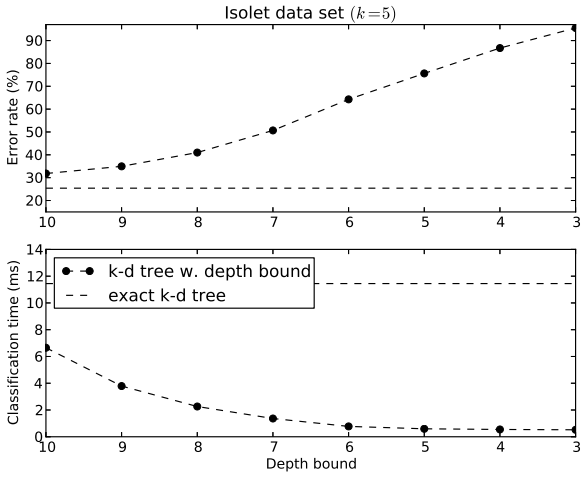


Fig. 6. Error rate and classification time vs depth bound for the k-d tree with Isolet data set.

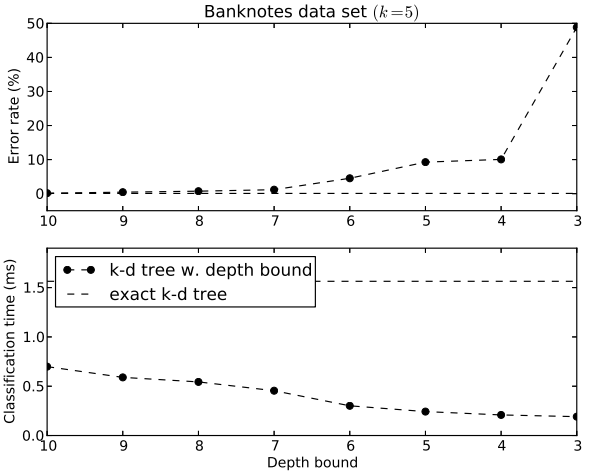


Fig. 5. Error rate and classification time vs depth bound for the k-d tree with Banknotes data set.

C. Node bound

Similarly to other algorithms with bounds, as the node bound value decreased, expected increase in the error rate and a decrease in classification time were observed. Lower limit for the node bound t was set so that the relation $t \geq k$ was satisfied, and above higher limits there was no change in the error rate with respect to an exact result.

For the k-d tree, all data sets demonstrated roughly similar behavior, showing that the node bound is indeed a promising approach. For instance, in the case of Ferrites data set, for $t = 9$ the time was reduced approximately 5-fold, with the absolute increase in the error rate of around 1.2% (see Figure 7).

The *best bin first* algorithm achieved mostly a slight improvement over the regular search procedure (e.g., for Banknotes data set demonstrated in Figure 8), and it turned out to be most effective for the Isolet data set with 617 dimensions (see Figure 9). Nonetheless, very optimistic results reported by Lowe [11] were not reproduced for data sets used in this article. The difference between classification time for BBF and the regular node bound approach was negligible, and thus the former is omitted.

As regards the ball tree, the error rate behaved rather strangely. For Banknotes and Iris data sets (presented in Figure 10 and Figure 11, respectively), we can see unexpected spikes in the error rate, although there remained a steady decrease in classification time.

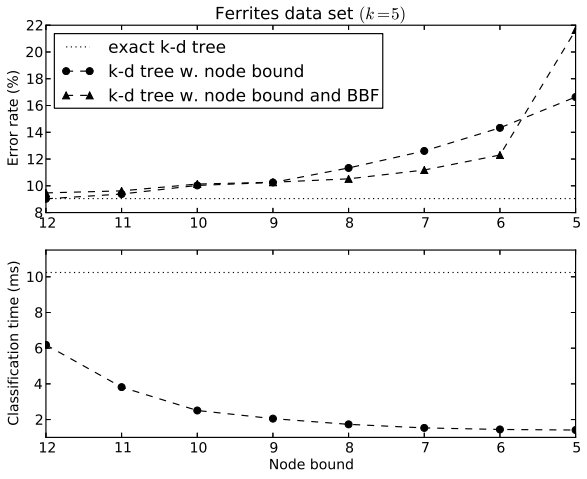


Fig. 7. Error rate and classification time vs node bound for the k-d tree with Ferrites data set, with and without BBF.

D. Probabilistic pruning

Since the pruning probability (σ) is equal for all subtrees, it might be the case that the pruned subtree contains one node, just as well as it might be the half of the entire tree. For this

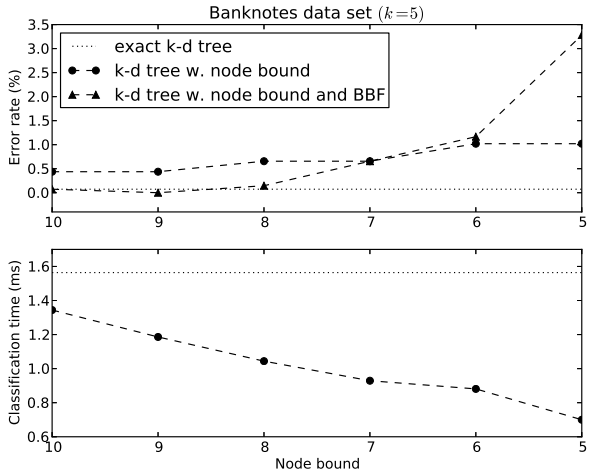


Fig. 8. Error rate and classification time vs node bound for the k-d tree with Banknotes data set, with and without BBF.

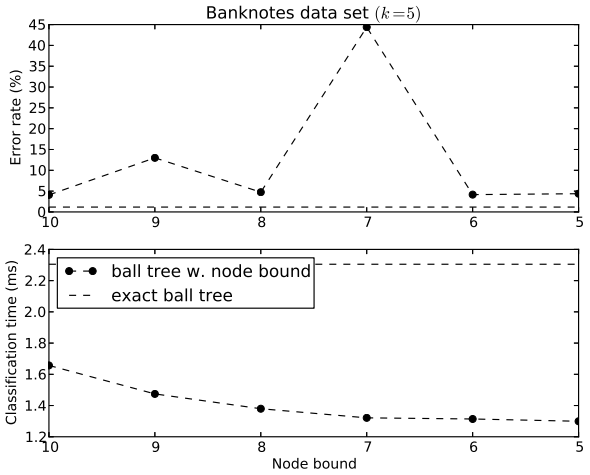


Fig. 10. Error rate and classification time vs node bound for the ball tree with Banknotes data set.

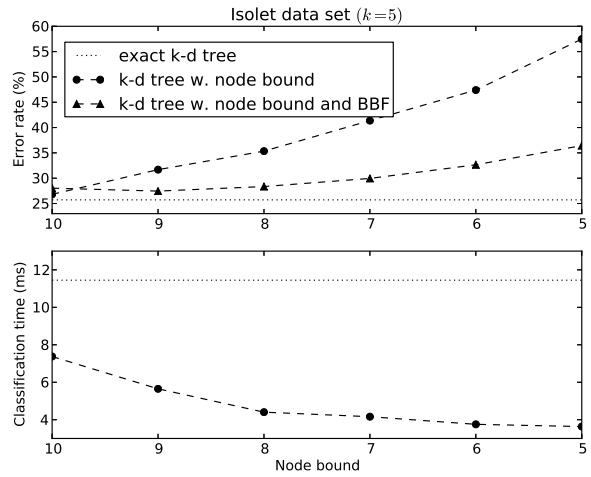


Fig. 9. Error rate and classification time vs node bound for the k-d tree with Isolet data set, with and without BBF.

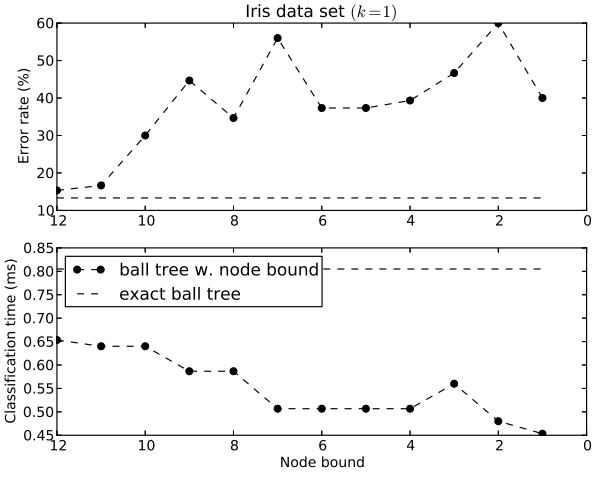


Fig. 11. Error rate and classification time vs node bound for the ball tree with Iris data set.

reason, effectiveness of this approach is clearly subject to a substantial amount of chance. Nevertheless, similar tendencies were observed for all data sets.

For the k-d tree, the ratio of error rate increase to classification time decrease tended to be less favorable than in the case of bounds presented in previous subsections — compare Figure 5 with Figure 13 to see how depth bound performed better for Banknotes data set. Still, for the Isolet data set (Figure 12), there was only a marginal increase in the error rate for $\sigma \leq 0.4$, and for these values the time was reduced by up to 35%.

Error rate increases for the ball tree were sharp (e.g., see Figure 14), and we observed once again unexpected results when the error rate for Banknotes data set actually decreased as more branches were pruned (see Figure 15). This can be most probably ascribed to simple luck resulting from the particular structure of this data set. Influence of the probabilistic nature of this algorithm was minimized by the fact that it was run five times.

VI. Conclusion

We have investigated two spatial data structures with identical applications, but different mechanics. The main difference between the k-d tree and the ball tree is the location of nodes associated with training data. The k-d tree does not consist of any redundant nodes, and each node corresponds to one instance from the training data. On the other hand, all training data in the ball tree are stored in the leaves, and internal nodes are utilized only in order to speed up the search procedure.

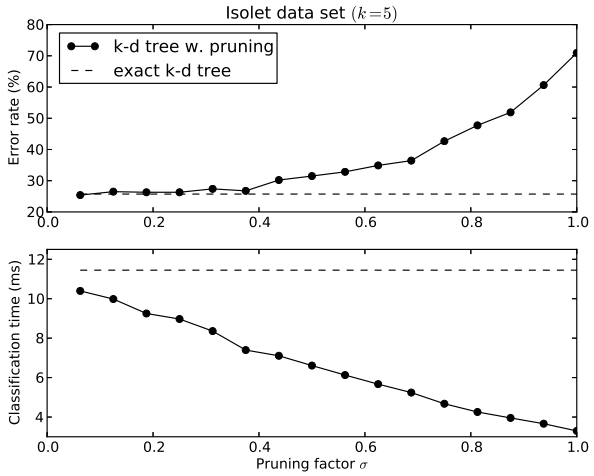


Fig. 12. Error rate and classification time vs pruning factor σ for the k-d tree with Isolet data set.

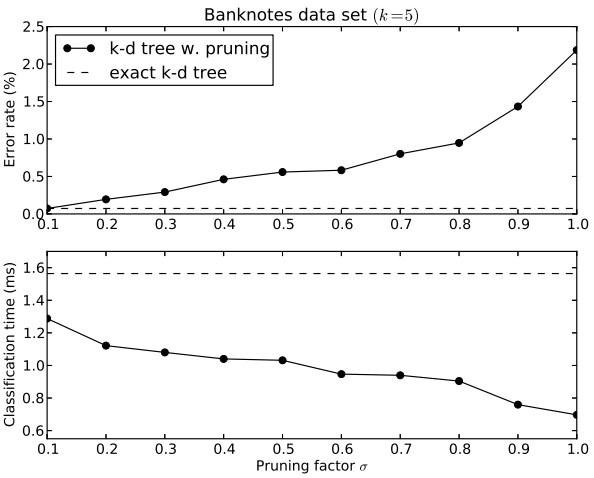


Fig. 13. Error rate and classification time vs pruning factor σ for the k-d tree with Banknotes data set.

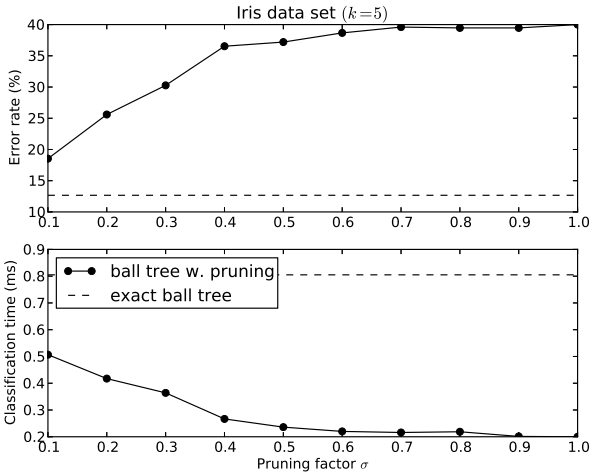


Fig. 14. Error rate and classification time vs pruning factor σ for the ball tree with Iris data set.

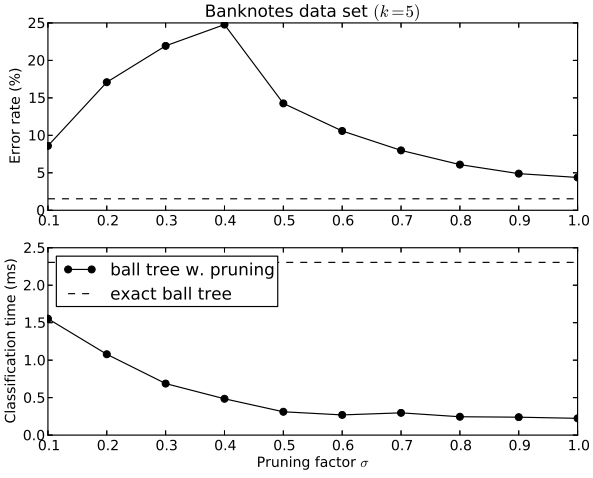


Fig. 15. Error rate and classification time vs pruning factor σ for the ball tree with Banknotes data set.

Approximate algorithms based on both trees turned out to perform rather well. The ratio of the increase in the error rate to the decrease in classification time was favorable, and the best results were reported for the k-d tree with node bound and best bin first (BBF) priority search. In the case of the ball tree, increases in the error rate were more rapid and unpredictable, however, this can be partially explained by the use of 5-fold cross validation instead of the leave-one-out method. Overall, the k-d tree was faster than the ball tree for both exact and approximate variants, which is consistent with exact performance measures presented by Munaga and Jarugumalli [12], and Kibriya and Frank [13].

Results depended chiefly on the data set that was used. No particular relation between the structure or size of the input data and results was observed, and it can be concluded that empirical findings remain the most valuable indicator, in spite of general tendencies.

We conclude that approximate variants of the *k*-nearest neighbor classification rule are indeed a very promising approach, and they are often indispensable when it comes to real-world massive data sets. Other data structures, which are based on the notion of partitioning the feature space, could also be adapted to use aforementioned bounds (CPU time, depth, node) and probabilistic pruning.

APPENDIX A

We list the data sets that were used for obtaining experimental results, along with their properties: class count, attribute count, and instance count.

- Banknotes — 2, 4, 1372
- Ferrites — 8, 30, 5903
- Iris — 3, 4, 150
- Isolet — 26, 617, 1559

REFERENCES

[1] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, "An optimal algorithm for approximate nearest neighbor searching fixed dimensions," *Journal of the ACM (JACM)*, vol. 45, no. 6, pp. 891–923, 1998. doi: 10.1145/293347.293348

[2] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. ACM, 1998. doi: 10.1145/276698.276876 pp. 604–613.

[3] P. W. Jones, A. Osipov, and V. Rokhlin, "Randomized approximate nearest neighbors algorithm," *Proceedings of the National Academy of Sciences*, vol. 108, no. 38, pp. 15 679–15 686, 2011. doi: 10.1073/pnas.1107769108

[4] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, no. 9, pp. 509–517, 1975. doi: 10.1145/361002.361007

[5] S. Arya, D. M. Mount, and O. Narayan, "Accounting for boundary effects in nearest-neighbor searching," *Discrete & Computational Geometry*, vol. 16, no. 2, pp. 155–176, 1996. doi: 10.1007/BF02716805

[6] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms (3. ed.)*. MIT Press, 2009. ISBN 9780262033848

[7] S. M. Omohundro, *Five balltree construction algorithms*. International Computer Science Institute Berkeley, 1989.

[8] J. S. Beis and D. G. Lowe, "Shape indexing using approximate nearest-neighbour search in high-dimensional spaces," in *CVPR*, 1997. doi: 10.1109/CVPR.1997.609451 pp. 1000–1006.

[9] M. L. Fredman and R. E. Tarjan, "Fibonacci heaps and their uses in improved network optimization algorithms," *J. ACM*, vol. 34, no. 3, pp. 596–615, 1987. doi: 10.1145/28869.28874

[10] A. Cisłak, "Approximate and probabilistic variants of the k-nearest neighbor classification rule," Bachelor's Thesis, Lodz University of Technology, 2014.

[11] D. G. Lowe, "Object recognition from local scale-invariant features," in *ICCV*, 1999. doi: 10.1109/ICCV.1999.790410 pp. 1150–1157.

[12] H. Munaga and V. Jarugumalli, "Performance evaluation: Ball-tree and kd-tree in the context of mst," *CoRR*, vol. abs/1210.6122, 2012. doi: 10.1007/978-3-642-32573-1_38

[13] A. M. Kibriya and E. Frank, "An empirical comparison of exact nearest neighbour algorithms," pp. 140–151, 2007. doi: 10.1007/978-3-540-74976-9_16

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 101–110
DOI: 10.15439/2014F265
ACSIS, Vol. 2

Identification of malware activities with rules

Bartosz Jasiul, Joanna Śliwa, Kamil Gleba
Military Communication Institute,
C4I Systems' Department,
ul. Warszawska 22a, 05-130 Zegrze, Poland
Email: {b.jasiul, j.sliwa, k.gleba}@wil.waw.pl

Marcin Szpyrka
AGH University of Science and Technology,
Department of Applied Computer Science,
al. Mickiewicza 30, 30-059 Kraków, Poland
Email: mszpyrka@agh.edu.pl

***Abstract*—The article describes the method of malware activities identification using ontology and rules. The method supports detection of malware at host level by observing its behavior. It sifts through hundred thousands of regular events and allows to identify suspicious ones. They are then passed on to the second building block responsible for malware tracking and matching stored models with observed malicious actions. The presented method was implemented and verified in the infected computer environment. As opposed to signature-based antivirus mechanisms it allows to detect malware the code of which has been obfuscated.**

I. Introduction

OVERWHELMING number of computer systems are connected to each other by global network – Internet, which allows to produce results beyond those achievable by the individual systems alone. Outcomes of cooperative work and accessibility of information are perceived and appreciated probably by all its users.

The advantages of this technology are available, unfortunately, also for hostile goals. The number of cyber threats arises rapidly from 23 680 646 in 2008 [1] to 1 595 587 670 in 2012 [2], and this is nowadays one of the most vexing problems in computer system security [3]. At the end of 2012 Kaspersky Lab, the Russian producer of antivirus software, reported that [4] *it currently detects and blocks more than 200 000 new malicious programs every day, a significant increase from the first half of 2012, when 125 000 malicious programs were detected and blocked each day on average*.

Although awareness about necessary security appliances seems to be common and the tools used for that purpose are getting more and more advanced, the number of successful attacks targeted on computer systems is growing [5]. They are mostly related to denial of offered services, gaining access or stealing private data, financial fraud, etc. Moreover, the evolution towards cloud computing, increasing use of social networks, mobile and peer-to-peer networking technologies that are intrinsic part of our life today, carrying many conveniences within our personal life, business and government, gives the possibility to use them as tools for cyber criminals and potential path of malware propagation [6]. Computer systems are prone to cyber attacks even though a number of security controls are already deployed [7], [8]. Cyber criminals are focused on finding a way to bypass security controls and gain access into the protected network. For that reason organizations, companies, governments and institutions as well as ordinary citizens all over the world are interested in detection of all attempts of malicious actions targeted on their computer networks and single machines [9].

Malicious activity detection starts with application of various techniques, the success rate of which depends on the reliability of the malware model. Usually they are based on code signatures. Security controls (e.g. antivirus tools) might be maladjusted because signatures of new threats are not identified yet. Hackers often use existing parts of code in order to implement new types of malware. This allows, in return, to quickly develop signatures of new dangerous software. Therefore, the more signatures are deployed the more malicious codes are identified. On the other hand, one of the methods for misleading signature-based detection systems is code obfuscation, the aim of which is generating – from already existing code – a new application that cannot be assessed yet as risky by security controls [10]. This technique is simple to be used and potentially successful. One of the countermeasures in this case is to follow behaviors of malicious software in order to identify them and eliminate from the protected system.

According to the study conducted in 2012 by the Verizon RISK Team [11] with cooperation from many national federal organizations, including e.g. Australian Federal Police, Irish Reporting and Information Security Service, and United States Secret Service *new techniques that speed up the process of malware detection to hours* are necessary. Authors of the report [12] indicate that *antivirus products should be supported by malware behavioral analysis tools in order to detect those of attacks for which signatures were not established*. An existing example of appliance that uses behavioral analysis for advanced persistent threats detection is Digital DNA by HBGary that extends the capabilities of McAfee Total Protection antivirus [13]. Detailed technical specifications of this solution have not been released for public. The product brochure explains that *multiple low level behaviors are identified for every running program or binary*. This leads to conclusion that each application is observed from behavioral perspective. McAfee is proud that the solution allowed to detect last year more 0-day

Work has been partially financed byby the National Centre for Research and Development project no. PBS1/A3/14/2012 "Sensor data correlation module for detection of unauthorized actions and support of decision process" and the European Regional Development Fund the Innovative Economy Operational Programme, INSIGMA project no. 01.01.02-00-062/09.

attacks than during the previous five years combined. This indicates the scale of new malware development and efficacy of the behavioral approach.

II. State of the art in malware detection

Currently there are two major techniques seen as prospective for malicious threats detection. One of them is *machine learning* [14] which allows to detect anomalies in the use of host machines by malicious software. This approach is only applicable in systems, for which a model of normal behavior can be established. It is only possible in such an environment where patterns for host machine activity can be identified, e.g. in production environment, SCADA systems, etc. Current usage of computer systems, mobility of users, enormous number of executed applications and visited internet sites cause that setting up a normal behavior model for malware detection is almost impossible. Such a method may also generate too many *false positive* alarms.

Other methods used for identification of malicious actions are based on different forms of specially prepared behavioral patterns prepared in the process of static code analysis or various host-based honeypots and sandboxes. Those patterns can be then applied in detection tools, e.g. rule based engines and complex event processing (CEP) tools. This article presents the method that uses rules [15] in order to identify malicious actions of the host machine and filters out from the number of system activities only these that are typical for malware.

III. Idea of the solution

In our work we proposed and developed behavior-oriented malware hunting tool, so called `PRONTO`, that could be used in parallel to existing signature-based tools.

The main assumption for the introduced method is that the malware was not recognized yet by the signature mechanisms. The aim therefore is to track its suspicious activities in order to find it while running in the system.

`PRONTO` hunting tool performs its activity in two stages (Fig. 1):

- **Filtering of the system events** registered by the system monitors (sensors) to discover the main features of the hostile activity. These features are related to particular objects and actions triggered on that objects – e.g. registry (add entry, modify entry, delete registry entry, etc.), process (start, stop process, etc.), file (copy, delete, run, open, close file, etc.), domain (connect to, etc.), IP address (connect to, etc.);
- **Tracking suspicious activity** in order to discover malicious exploits running in the system. Filtered events are correlated in order to find similarities with the stored malware activities modeled in the form of Colored Petri nets [16]. The result of malware tracking is the alarm that contains information vector about malicious activity, similarity to the known attacks and list of incidents that affected the system.

This article presents only the first stage which is related to capturing events from sensors and analyzing them with an expert system that uses – defined for the purpose of the method – comprehensive ontology, so called `PRONTOlogy`. Registered events in the form of XML objects are sent to the PRONTOntology engine and lifted to add entries to the Knowledge Base. `PRONTOlogy` describes events registered by system monitors and is able, on the basis of rule engine and inference, with the use of specially defined rules [17], to classify an event as potentially suspicious, malicious or regular. As a result, markings of the modeled malware in the form of CP-nets [18][19] are delivered for further analysis.

The second element of the threats' tracking component of the solution is `PRONTOnet` [20]. It provides formal model of malware behavior and allows to track suspicious activities potentially assigning them to the class of known malware types or identifying unknown ones. Known exploits can be invisible to signature-based malware detecting tools after their code has been obfuscated, although their activities can be easily observed. It also happens often that a new malware piece of software is composed of known components from other ones. This results in another behavior pattern that can be tracked as a new exploit, not identified yet. The result of threats tracking stage is an alert informing about identification of suspicious or malicious events with a certain similarity rate to the known malware types.

IV. Identification of malicious actions

Static analysis of malicious code or intelligent algorithms for malware behavior recognition provide patterns that can be defined in low level programming language (e.g. Assembler) or can be represented on the level of operating system activities. In case of our solution the second approach was selected, which means that identification of malicious actions is performed while monitoring actions of the host machine. It was mainly due to availability of tools for operating system monitoring and easiness of processing.

The two-stage malware hunting process presented in this article starts with sifting through a great number of actions that are generated by the up and running operating system. This aimes for identification of those events that should be perceived as suspicious and processed further on. This process should enable automatic filtering of events on the basis of their characteristic features. However, it is not trivial to assume an action is suspicious since the mechanism must catch the context of its invocation in order to assess if it is a regular operating system or user activity, or anomaly that should be investigated further on. For this reason, it was necessary to use a method that could provide the possibility to deduce from the gathered data and analyze possible correlation among events. These requirements were met by the semantic techniques based on ontology and rules that enable to create knowledge base and infer additional facts automatically.

According to [21] *An ontology is an explicit and formal specification of a conceptualization*.

In general, ontology describes a domain of discourse formally. Typically, ontology consists of a finite list of terms, and relationships between those terms. This set describes so

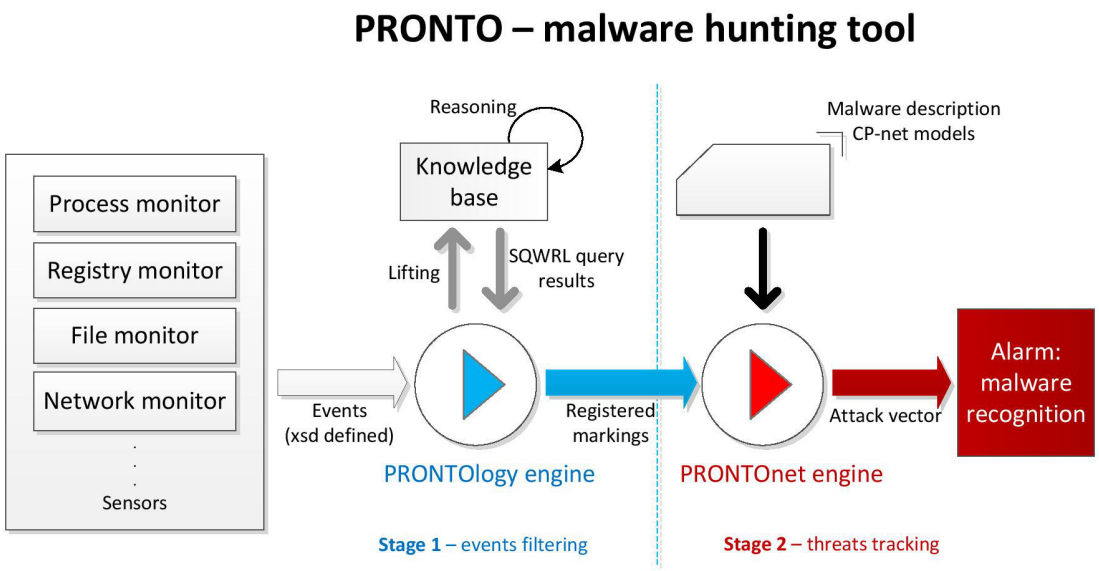


Fig. 1. PRONTO – malware hunting tool

called TBox statements, which are Terminological statements describing the domain in terms of controlled vocabularies. They describe important concepts (classes of objects) of the domain and their properties.

For the purpose of this solution there has been proposed an ontology modeled in the Web Ontology Language (OWL) titled `PRONTOlogy.owl` that describes basic classes and relationships among them. Since the investigated domain needs the description that would enable to reflect and represent facts that a resource executes an action on another resource particular object properties are used. They indicate actions executed on resources and enable to define appropriate triples (e.g. `run (x,y)`, where `x`, `y` are members of `Resource` class and `run` is *object property*, with *domain* and *range* equal to Resource).

Based on TBoxes there can be defined e.g. the following general statements:

```
Event(x)
Resource(y)
ResFile(z)
hasResource(x,y)
Resource(y) is a ResProcess
ResFile(z) is a Resource
run (y,z)
```

where:

- `Event, Resource, ResProcess, ResFile` – are classes,
- `hasResource, run` – are object properties,
- `is a` – is subclass relationship.

According to Fig. 2 the model ontology consists of the three main *classes*: `Event`, `Place`, `Resource`.

In order to differentiate types of resources that perform actions observed by system monitors, there have been defined

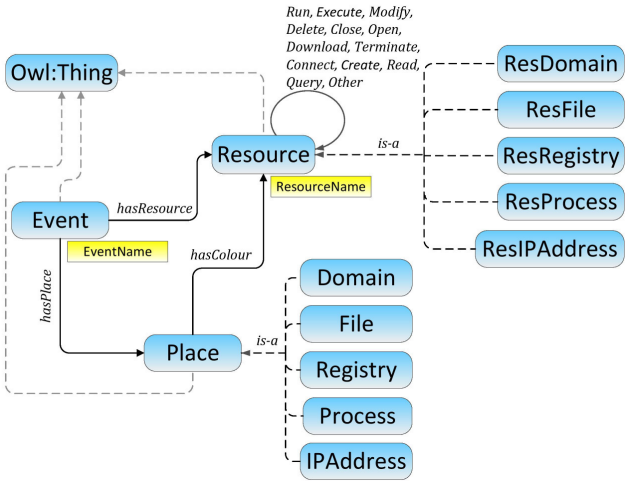


Fig. 2. PRONTOlogy model

the following *subclasses* of the `Resource` class:

- `ResFile` – where the resource is a file,
- `ResRegistry` – where the resource is a registry,
- `ResProcess` – where the resource is a process,
- `ResDomain` – where the resource is a domain the system is trying to connect to,
- `ResIPAddress` – where the resource is an IP address the system is trying to connect to.

In order to indicate particular registry entries, file names, etc. *datatype properties* have been proposed. They describe name of the Resource (`ResourceName`) and name of particular Event (`EventName`).

In order to reflect activities on system resources there have been modeled the following *object properties* that resemble

types of system activities:

- `run` – e.g. running a service, process;
- `open` – e.g. opening a file, registry;
- `close` – e.g. closing a file, application, process;
- `modify` – e.g. modification of registry entry, file, process;
- `execute` – e.g. executing an application;
- `terminate` – e.g. terminating a process;
- `connect` – e.g. connecting to the IP address or domain;
- `query` – e.g. querying the registry entry state;
- `download` – downloading data from remote location;
- `create` – creating a new object, e.g. registry entry, file;
- `delete` – deleting an object, e.g. file, registry entry.

For all *object properties* listed above `Domain` and `Range` are equal to `Resource`, which means that one resource can execute actions on other resources.

There are also additional *object properties* that can reflect the fact:

1) that particular event should be perceived as a `Place`: `hasPlace`, where `Domain = Event, Range = Place`,
2) that particular marking appears for particular `Place`: `hasColor`, where `Domain = Place, Range = Resource`, and
3) that particular event is related to a `Resource`: `hasResource`, where `Domain = Event, Range = Resource`.

The above defined types of *classes* and *object properties* enable to describe events that are registered by system monitors (sensors). Additionally, the model was constructed in such a way that it can reflect the fact that particular observed activity should be perceived as a token in the Petri net - used further on at the second stage of malware hunting tool operation. This process is performed automatically with the use of reasoning rules modelled with Semantic Web Rule Language (SWRL) [22], which offers appropriate expressiveness and tool support in order to use it for the assumed purpose [23].

A. PRONTOlogy engine

The ontology model presented above is used in the PRONTOlogy engine. System activities that are logged by different system sensors form a stream of hundreds to even hundred thousands of events per minute. They record activities of the user and related background activity of the system. In terms of presented solution sensors that cover the spectrum of incidents describing the behavior of different malware types are process, registry, file and network monitors, reflected in the ontology.

Sensors allow to log file system, registry and process/thread activities in real-time. After a proper configuration they enable sifting through incoming events and comprehensive event properties such as session ID numbers, user names, reliable process information, full thread stacks with integrated symbol support for each operation, simultaneous logging to a file, etc.

The first stage of the PRONTOlogy engine operation is devoted to the analysis of this events' stream and classification of single events as either suspicious or regular ones. This classification is assumed as a background activity for the monitoring of the system state realized by the second component of PRONTO, that is threats' tracking (PRONTOnet).

Particular types of malware perform distinctive activities. Each of them is different or, what is more, they can have their types. This entails various malware realization. If malware signature is unknown (the code has been obfuscated), identification of its activity can be done by analyzing system events. The first stage of this process is related to filtering of events and classifying them as neutral or suspicious. In the latter case information about the event is passed down to the next stage – threat tracking.

The stage of events filtering is based on the ontology engine that automatically, with the use of the knowledge base and a set of rules, defines if the registered event is suspicious and should be tracked further by the PRONTOnet module. Knowledge base is created with lifting the information about events registered by sensors to create assertions and facts (entering ABox statements into the knowledge base). Suspicious actions are modeled as instances of the `Place` class. As already mentioned, the rules will provide the possibility to infer facts that particular event indicates existence of a `Place` in the CP-net model (`hasPlace` object property) and therefore particular token (`hasColor` object property) exists and this fact should be passed further on to the PRONTOnet. For instance, the rules can infer that e.g. an event called *winlogon.exe_run_VRT7.tmp* which means that the `winlogon.exe` process has run `VRT7.tmp` file is suspicious and sends on this information to the threat tracing module for further investigation.

As events from sensors are delivered to PRONTOlogy engine, new facts are inserted into the knowledge base in the form of ABox statements. In order to provide additional facts to the knowledge base in terms of appearance of a new token in the CP-net model of particular attack, the rules are proposed. The following listing shows the rule which head defines a condition: an event where a process named *csrss* opens a file named *open.exe*, which in fact is an infected file. When this condition is met, it results in identification of a new `Place` in the CP-net model, which is a `File` with token `open.exe`.

```
Place(?c) ^ Resource(?y) ^ resourceName(?y,
"csrss.exe") ^ open(?y, ?z)   ^ ResFile(?z)
^ resourceName(?z, "open.exe") ->  File(?c)
^ hasColour(?c, ?z)
```

A new set of rules will be prepared whenever new threats appear in the process of system vulnerabilities analysis. For the purpose of the solution verification there has been shown an exemplary set that will provide the possibility to discover markings of places defined in the CP-net model and used in `PRONTOnet`.

V. PRONTOLOGY.OWL EVALUATION

This section is devoted to ontology evaluation which, according to [24], should consist in *validation and verification*

of an ontology in terms of its scope, consistency and expressiveness.

Semantic model defined in PRONTOlogy is devoted to reflect events that occur in the monitored system and enable to identify these suspicious ones. This model has direct relationship with the CP-net through the use of the `Place` class the instances of which are passed over to the PRONTOnet. PRONTOlogy defines:

- The event instance modeled with the use of `Event` class. Data about occurred events is received from the sensors, then lifted to the ontology model as instances of the `Event` class. Each instance has `eventName` (*data property*) defined by the sensor. The description of an event is modeled with the use of `hasResource` *object property* which indicates initiator of the event which is some system resource.
- System resources that are under monitoring by sensors – File, Registry, Process, Domain, IPaddress. They are modeled by the following classes: `ResFile`, `ResRegistry`, `ResProcess`, `ResDomain`, `ResIPAddress`, which are subclasses of the `Resource` class.
- The event description modeled with the use of *object properties* (`run`, `create`, `modify`, `delete`, `download`, `open`, `close`, `read`, `execute`, `terminate`, `connect`, `query`). These *object properties* `domain` and `range` is the `Resource` class, which means that resources perform actions on other resources.
- An abstract `Place` class that defines the fact that particular event is suspicious and should be handled over by the CP-net model for further investigation. This class has five subclasses that define the type of a Place, which in turn results from event originator and reflects Places in the CP-net model.

With the use of the proposed ontology it is possible to describe the event of running a file by particular process. For instance occurrence of `winlogon.exe_run_VRT7.tmp` event would cause inserting of the following instances into the knowledge base:

```
http://wil.waw.pl/secor/PRONTOlogy.owl#
Event_1
http://wil.waw.pl/secor/PRONTOlogy.owl#
eventName(http://wil.waw.pl/secor/
PRONTOlogy.owl#Event_1,
"winlogon.exe_run_VRT7.tmp")
http://wil.waw.pl/secor/PRONTOlogy.owl#
ResProcess_8
http://wil.waw.pl/secor/PRONTOlogy.owl#
resourceName(http://wil.waw.pl/secor/
PRONTOlogy.owl#ResProcess_8,
"winlogon.exe")
http://wil.waw.pl/secor/PRONTOlogy.owl#
ResFile_9
http://wil.waw.pl/secor/PRONTOlogy.owl#
resourceName(http://wil.waw.pl/secor/
PRONTOlogy.owl#ResFile_9, "vrt7.tmp")
http://wil.waw.pl/secor/PRONTOlogy.owl#
run(http://wil.waw.pl/secor/
PRONTOlogy.owl#ResProcess_8,
http://wil.waw.pl/secor/
PRONTOlogy.owl#ResFile_9)
http://wil.waw.pl/secor/PRONTOlogy.owl#
hasResource(http://wil.waw.pl/secor
/PRONTOlogy.owl#Event_1,
http://wil.waw.pl/secor/PRONTOlogy.owl#
ResProcess_8)
```

The stage of events filtering is based on the ontology engine that automatically, with the use of the knowledge base and a set of rules defines if the registered event is suspicious and should be tracked further by the threat tracing module. As already mentioned knowledge base is created with lifting the information about events registered by sensors to create assertions and facts. Suspicious actions are modeled as instances of the `Place` class. The knowledge about the object which is also an instance of the `Place` class is derived by the set of rules proposed for the purpose of PRONTOlogy.

If the following rule is applied:

```
Place(?c)^Resource(?y)^resourceName
(?y, "winlogon.exe")^run(?y, ?z)
^ResFile(?z)^resourceName(?z,
"vrt7.tmp") -> File(?c)^hasColour(?c, ?z)
```

the following instances are added to the knowledge base:

```
http://wil.waw.pl/secor/PRONTOlogy.owl#
Place_1-is a member of File class
(inferred knowledge)
http://wil.waw.pl/secor/PRONTOlogy.owl#
hasColour(http://wil.waw.pl/secor/
PRONTOlogy.owl#Place_1,
http://wil.waw.pl/secor/PRONTOlogy.owl#
ResFile_9).
```

If the following rule is applied:

```
Event(?e)^Place(?c)^hasResource(?e,?y)
^Resource(?y)^resourceName(?y,
"winlogon.exe")^run(?y, ?z)^ResFile(?z)
^resourceName(?z,  "vrt7.tmp")
-> hasPlace(?e,?c)^File(?c)^
hasColour(?c, ?z)
```

additionally the relation

```
http://wil.waw.pl/secor/PRONTOlogy.owl#
hasPlace(http://wil.waw.pl/secor/
PRONTOlogy.owl# Event_1,
http://wil.waw.pl/secor/PRONTOlogy.owl#
Place_1).
```

is added.

If an event has `hasPlace` relation to any `Place` instance, it is a suspicious event.

The PRONTOlogy defines all entities that are necessary to describe events monitoring system behavior and identify suspicious events. Moreover, the direct relation between ontology and CP-nets has been modeled with the use of the `Place` class. That is why, they satisfy the required scope and expressiveness of ontology.

The second ontology evaluation step consists in checking the ontology consistency. According to [25] *ontology is consistent (also called satisfiable) when it does not contain a contradiction*. The lack of contradiction can be defined in either semantic or syntactic terms. The syntactic definition states that a theory is consistent if there is no such P formula that both P and its negation are provable from the axioms of the theory under its associated deductive system. The ontology model that contains formal definitions of classes, properties and individuals allows inferring new knowledge from knowledge that is already present. The fact that it is based on formal description logic makes it prone to logical reasoning and enables to infer knowledge from existing *facts* and *axioms* [26].

The consistency of `PRONTOlogy.owl` has been verified in the Protégé ontology editing tool (version 3.4.6) [27] using the Pellet 1.5.2 [28] reasoner on a machine with the following configuration:

- Processor: Intel Core i7 (2 cores 2,8 GHz each);
- RAM: 6 GB;
- Operating System: Windows 7 (64 bit).

The consistency check on this machine was successful and `PRONTOlogy.owl` has been proven consistent in 0,022 seconds.

To satisfy verification of events filtering with ontology and reasoning a web service `PRONTOlogyInterface` was implemented in Java programming environment. The service was developed with utilization of Protégé, Pellet, SWRL Jess bridge, and Jess71p2 programming libraries. Web Service was run on the GlassFish Server 3.1.2.

`PRONTOlogyInterface` consists of two programming packages:

- `wil.waw.pl.protegeclass.prontology`
- `wil.waw.pl.prontology`.

Java classes of `wil.waw.pl.protegclass.prontology` package were developed with *Generate Protégé-OWL Java Code* plug-in of Protégé editor. The generator allowed to define Java classes on the basis of `PRONTOlogy.owl` automatically.

Package `wil.waw.pl.prontology` consists of the following classes:

- `InferenceResult.java`,
- `OperationType.java`,
- `ResourceType.java`,
- `PRONTOlogy.java`.

`InferenceResult.java` class defines result code of `PRONTOlogyInterface` service. `OperationType.java` defines types of operations on resources:

```
public enum OperationType {RUN,EXECUTE,
CREATE,MODIFY,DELETE,CLOSE,OPEN,
DOWNLOAD,CONNECT ACCESS,TERMINATE,
QUERY,READ,OTHER}.
```

Enumerate class `ResourceType.java` defines types of resource:

```
public enum ResourceType{RESDOMAIN,
RESIPADDRES,RESUSER,RESREGISTRY,
RESFILE,RESPROCESS}.
```

`PRONTOlogy.java` is the main class of `PRONTOlogyInterface` and it implements the following web methods:

- `readOntologyFromFile()` that reads the ontology from the file;
- `inferKnowledge()` that realizes ontology reasoning;
- `queryForPetriPlace()` that identifies the *Place* in CP-net on the basis of defined knowledge base;
- `queryForPetriToken()` that identifies, on the basis of the knowledge base, a token assigned to a particular *Place*.

VI. Cyber attacks detection – an experiment

A. Data acquisition

The verification based on experimental data was made with the use of the most popular target of cyber infections – Microsoft Windows operating system. The authors do not claim that this is the most vulnerable system. In the authors' opinion the reason of cyber attacks on Windows operating system is the popularity of the system and potentially high gain from conducted attacks. Microsoft products are very popular which makes them attractive for cyber criminals.

For the observation of activities, applications, services and network connections in the native Microsoft Windows 7 operating system environment Sysinternals Suite utility package [29] was used. The Sysinternals Suite is a set of over 70 advanced diagnostic and troubleshooting programs for the Windows platform. These programs are available for free download from Microsoft's Technet web page [30].

Majority of events were observed with Process Monitor utility [31] – part of the Sysinternals Suite. Process Monitor is an advanced monitoring application for Windows that registers events which relate to file system, registry, and process activity in real-time. It enables monitoring event properties such as session IDs, user names, process information, thread stacks, simultaneous logging to a file, etc. It is a powerful utility that supports PRONTOlogy module with detailed information on activities in the protected system. An example of a single event acquired with Process Monitor is presented in following listing:

```
<event>
<ProcessIndex>14340</ProcessIndex>
<Time_of_Day>17:22:25,1104786</Time_of_Day>
<Process_Name>ThreatProc.exe</Process_Name>
<PID>2728</PID>
```

```
<Operation>RegQueryValue</Operation>
<Path>HKLMS\Microsoft\Windows NT\...\
SurrogateFallback\Plane2</Path>
<Result>SUCCESS</Result>
<Detail>Type: REG_SZ, Length: 24, Data:
SimSun-ExtB</Detail>
</event>
```

Process Monitor allows to report system events for further analysis and reasoning to the PRONTOlogy module. Detailed report on system activities includes, but is not limited to:

- **process name** – the name of the process performing the operation;
- **operation** – the name of the operation being logged;
- **path** (if applicable) – the path of the object that the operation is performed on (e.g. a registry path, a file system path);
- **result** – the result of the operation (e.g. Success, EoF, Buffer Overflow);
- **detail** – additional operation-specific information about the event.

For the purpose of data acquisition it is also possible to use API hooking tools [32], [33], however, they inject themselves (like viruses) to the processes, thus they can affect results of the verification. In case of utilization of PRONTO malware hunting tool for detection of network attacks various network utilities, e.g. SNORT [34], ARAKIS [35], iptables [36], should also be used [37], [38].

Having stored CP-net models of cyber attacks in the database, it is possible to go further with the experiment to malware detection phase. As mentioned above, the aim of the experiment is not only to identify existing malware that was obfuscated, but also 0-day attacks that have, to some degree, similar behavior to the already identified one.

B. Malware detection scenario

Within one minute operation of Windows 7 OS thousands or even hundreds of thousands single events may be observed. Report from the Process Monitor includes everything that took place in the system. It includes both regular and suspicious activities.

For the purpose of verification and, in particular, generation of these *unwanted* activities, three different machines were infected by Virut, VBMania@MM, and 0-day attack that was simulated with events typical to different parts of malicious codes.

At the same time, various programs were executed on these three machines in order to simulate legitimate user activity. This allowed us to generate background regular events.

In the article we show only the first example and provide the reader with information on steps of PRONTO operation in terms of malware detection with emphasis on events filtering phase.

Let us assume that data acquisition phase allowed to gather information about events collected by the Process Monitor. Obviously, the whole file with captured events will not be presented in this chapter although an exemplary excerpt from it is presented in following listing:

```
<event>
   <ProcessIndex>14340</ProcessIndex>
   <Time_of_Day>17:20:21,1001813
</Time_of_Day>
   <Process_Name>WINLOGON.EXE
</Process_Name>
   <PID>2728</PID>
   <Operation>ReadFile</Operation>
   <Path>C:\Windows\Temp\vrt7.tmp</Path>
   <Result>SUCCESS</Result>
   <Detail>Offset: 734 720, Length: 16 384,
 Priority: Normal</Detail>
</event>
<event>
   <ProcessIndex>14560</ProcessIndex>
   <Time_of_Day>17:22:25,1104786
</Time_of_Day>
   <Process_Name>ThreatProc.exe
</Process_Name>
   <PID>6043</PID>
   <Operation>RegSetValueEx</Operation>
   <Path>HKLMS\Microsoft\Windows\
CurrentVersion\Run\
   Windows System Monitor:
"C:\Windows\system\winrsc.exe"
   </Path>
   <Result>SUCCESS</Result>
   <Detail>Type: REG_SZ, Length: 24, Data:
SimSun-ExtB</Detail>
</event>
<event>
   <ProcessIndex>16640</ProcessIndex>
   <Time_of_Day>17:22:36,2548113
</Time_of_Day>
   <Process_Name>WINWORD.EXE
</Process_Name>
   <PID>6733</PID>
   <Operation>RegQueryKey</Operation>
   <Path>HKLM</Path>
   <Result>SUCCESS</Result>
   <Detail>Query: HandleTags, HandleTags:
0x0</Detail>
</event>
<event>
   <ProcessIndex>19240</ProcessIndex>
   <Time_of_Day>17:47:02,1294174
</Time_of_Day>
   <Process_Name>mmirc.exe
</Process_Name>
   <PID>12188</PID>
   <Operation>TCP Connect</Operation>
   <Path>MalwareTest1-VAIO:55052 ->
 irc.zief.pl:6667</Path>
```

```
   <Result>SUCCESS</Result>
   <Event_Class>Network</Event_Class>
   <Image_Path>C:\Windows\Temp\
mmirc.exe</Image_Path>
   <Session>1</Session>
</event>
```

The events presented in above listing are processed and XML data is lifted to the semantic metadata. Based on this example the following instances are inserted into the ontology model (as ABox entries):
for the first event:

```
http://wil.waw.pl/secor/PRONTOlogy.owl#
Event_1 - an instance of the  Event class
http://wil.waw.pl/secor/PRONTOlogy.owl#
eventName(http://wil.waw.pl/
secor/PRONTOlogy.owl#Event_1,
"winlogon_read_vrt.7")
http://wil.waw.pl/secor/PRONTOlogy.owl#
ResProcess_2728
http://wil.waw.pl/secor/PRONTOlogy.owl#
resourceName(http://wil.waw.pl/
secor/PRONTOlogy.owl#ResProcess_2728,
"winlogon.exe")
http://wil.waw.pl/secor/PRONTOlogy.owl#
ResFile_1
http://wil.waw.pl/secor/PRONTOlogy.owl#
resourceName(http://wil.waw.pl/
secor/PRONTOlogy.owl#ResFile_1,
"vrt7.tmp")
http://wil.waw.pl/secor/PRONTOlogy.owl#
read(http://wil.waw.pl/secor/
PRONTOlogy.owl#ResProcess_2728,
http://wil.waw.pl/secor/  PRONTOlogy.owl#
ResFile_1)
http://wil.waw.pl/secor/PRONTOlogy.owl#
hasResource(http://wil.waw.pl/
secor/PRONTOlogy.owl#Event_1,
http://wil.waw.pl/secor/PRONTOlogy.owl#
ResProcess_2728)
```

for the second event:

```
http://wil.waw.pl/secor/PRONTOlogy.owl#
Event_2 - an instance of the  Event class
http://wil.waw.pl/secor/PRONTOlogy.owl#
eventName(http://wil.waw.pl/secor/
PRONTOlogy.owl#Event_2,
"ThreadProc_modify_Windows_System_Monitor")
http://wil.waw.pl/secor/PRONTOlogy.owl#
ResProcess_6043
http://wil.waw.pl/secor/PRONTOlogy.owl#
resourceName(http://wil.waw.pl/
secor/PRONTOlogy.owl#ResProcess_6043,
"ThreatProc.exe")
http://wil.waw.pl/secor/PRONTOlogy.owl#
ResRegistry_1
http://wil.waw.pl/secor/PRONTOlogy.owl#
resourceName(http://wil.waw.pl/
secor/PRONTOlogy.owl#ResRegistry_1,
"HKEY_LOCAL_MACHINE\SOFTWARE\
 Microsoft\Windows\CurrentVersion\Run
\Windows System Monitor:
C:\Windows\system\winrsc.exe")
http://wil.waw.pl/secor/PRONTOlogy.owl#
modify(http://wil.waw.pl/secor/
PRONTOlogy.owl#ResProcess_6043,
http://wil.waw.pl/secor/  PRONTOlogy.owl#
ResRegistry_1)
http://wil.waw.pl/secor/PRONTOlogy.owl#
hasResource(http://wil.waw.pl/
secor/PRONTOlogy.owl#Event_2,
http://wil.waw.pl/secor/PRONTOlogy.owl#
ResProcess_6043)
```

for the third event:

```
http://wil.waw.pl/secor/PRONTOlogy.owl#
Event_3 - an instance of the  Event class
http://wil.waw.pl/secor/PRONTOlogy.owl#
eventName(http://wil.waw.pl/secor/
PRONTOlogy.owl#Event_3,
"Winword_read_HKLM")
http://wil.waw.pl/secor/PRONTOlogy.owl#
ResProcess_6733
http://wil.waw.pl/secor/PRONTOlogy.owl#
resourceName(http://wil.waw.pl/
secor/PRONTOlogy.owl#ResProcess_6733,
"Winword.exe")
http://wil.waw.pl/secor/PRONTOlogy.owl#
ResRegistry_2
http://wil.waw.pl/secor/PRONTOlogy.owl#
resourceName(http://wil.waw.pl/
secor/PRONTOlogy.owl#ResRegistry_2,
 "HKLM")
http://wil.waw.pl/secor/PRONTOlogy.owl#
read(http://wil.waw.pl/secor/
PRONTOlogy.owl#ResProcess_6733,
http://wil.waw.pl/secor/  PRONTOlogy.owl#
ResRegistry_2)
http://wil.waw.pl/secor/PRONTOlogy.owl#
hasResource(http://wil.waw.pl/
secor/PRONTOlogy.owl#Event_3,
http://wil.waw.pl/secor/PRONTOlogy.owl#
ResProcess_6733)
```

for the fourth event:

```
http://wil.waw.pl/secor/PRONTOlogy.owl#
Event_4 - an instance of the  Event class
http://wil.waw.pl/secor/PRONTOlogy.owl#
eventName(http://wil.waw.pl/secor/
PRONTOlogy.owl#Event_4,
"mmirc_connect_irc_zief_pl")
http://wil.waw.pl/secor/PRONTOlogy.owl#
```

```
ResProcess_12188
http://wil.waw.pl/secor/PRONTOlogy.owl#
resourceName(http://wil.waw.pl/
secor/PRONTOlogy.owl#ResProcess_12188,
 "mmirc.exe")
http://wil.waw.pl/secor/PRONTOlogy.owl#
ResDomain_1
http://wil.waw.pl/secor/PRONTOlogy.owl#
resourceName(http://wil.waw.pl/
secor/PRONTOlogy.owl#ResDomain_1,
"irc.zief.pl")
http://wil.waw.pl/secor/PRONTOlogy.owl#
connect(http://wil.waw.pl/secor/
PRONTOlogy.owl#ResProcess_12188,
http://wil.waw.pl/secor/  PRONTOlogy.owl#
ResDomain_1)
http://wil.waw.pl/secor/PRONTOlogy.owl#
hasResource(http://wil.waw.pl/
secor/PRONTOlogy.owl#
Event_4, http://wil.waw.pl/secor/
PRONTOlogy.owl#ResProcess_12188)
```

The rules that are valid in the presented scenario allow to infer that three of the above events are suspicious. These are the following rules:

```
Event(?e)^Place(?c) hasResource(?e,?y)^
resourceName(?y, "winlogon.exe")^
read(?y, ?z)^ ResFile(?z) ^
resourceName(?z,  "vrt7.tmp")->
hasPlace (?e,?c)^File(?c)^hasColour(?c, ?z)
```

```
Event(?e)^Place(?c)^hasResource(?e,?y)
^modify(?y, ?z)^ResRegistry(?z)^
resourceName(?z,"HKLMS\Microsoft\...\Run\
Windows System Monitor: C:\Windows\
system\winrsc.exe")->hasPlace(?e,?c)
^ Registry(?c)^hasColour(?c, ?z)
```

```
Event(?e)^Place(?c)^hasResource(?e,?y)
^connect(?y, ?z)^ResDomain(?z)
^resourceName(?z,  "irc.zief.pl")
-> hasPlace (?e,?c)^Domain(?c)^
hasColour(?c, ?z)
```

On the basis of these rules the following facts are inferred:

```
http://wil.waw.pl/secor/PRONTOlogy.owl#
Place_1 - member of the File class
http://wil.waw.pl/secor/PRONTOlogy.owl#
hasPlace(http://wil.waw.pl/secor/
PRONTOlogy.owl#Event_1, http://wil.waw.pl/
secor/PRONTOlogy.owl#Place_1).
http://wil.waw.pl/secor/PRONTOlogy.owl#
hasColour(http://wil.waw.pl/secor/
PRONTOlogy.owl#Place_1,
http://wil.waw.pl/secor/PRONTOlogy.owl#
ResFile_1).

http://wil.waw.pl/secor/PRONTOlogy.owl#
Place_2 - member of the Registry class
http://wil.waw.pl/secor/PRONTOlogy.owl#
hasPlace(http://wil.waw.pl/secor/
PRONTOlogy.owl#Event_2,
http://wil.waw.pl/secor/
PRONTOlogy.owl#Place_2).
http://wil.waw.pl/secor/PRONTOlogy.owl#
hasColour(http://wil.waw.pl/secor/
PRONTOlogy.owl#Place_2,
http://wil.waw.pl/secor/PRONTOlogy.owl#
ResRegistry_1).

http://wil.waw.pl/secor/PRONTOlogy.owl#
Place_3 - member of the Domain class
http://wil.waw.pl/secor/PRONTOlogy.owl#
hasPlace(http://wil.waw.pl/secor/
PRONTOlogy.owl# Event_2,
http://wil.waw.pl/secor/
PRONTOlogy.owl#Place_2).
http://wil.waw.pl/secor/PRONTOlogy.owl#
hasColour(http://wil.waw.pl/secor/
PRONTOlogy.owl#Place_2,
http://wil.waw.pl/secor/PRONTOlogy.owl#
ResDomain_1).
```

Events 1, 2 and 4 have been identified as suspicious, whereas event 3 – as a regular system activity. The SQWRL query that allowed to select this knowledge from the ontology had the following structure:

```
tbox:isSubClassOf(?subClass, Place)^
abox:hasIndividual(?subClass, x)->
sqwrl:select(?subClass)
```

```
Place(?p)^hasColour(?p, ?c)^
resourceName(?c, ?n)->sqwrl:select(?n)
```

The rules applied in the PRONTOlogy module allowed to pass forward to the PRONTOnet module only information about suspicious events in the form of *Places* and appropriate *tokens* assigned to them (with the use of hasColour *object property*). It takes place in the *acquisition module* as presented in the architecture of solution. Then, in the PRONTOnet, these tokens are passed to *verification module* where marking M_a of *Places* is:

$M_a = M_{File} \cup M_{Domain} \cup M_{Registry}$, where:

- M_{File}={vrt7.tmp},
- M_{Domain}={irc.zief.pl},
- $M_{Registry}$ = {HKLMS\Microsoft\...\Run\Windows System Monitor:"C:\Windows\system\winrsc.exe"}.

At the machine described in this scenario the detection realized with the use of CPN MM and marking M_a allowed to identify Virut attack. The result vector is as follows:

```
1’ 1|Virut|vrt7.tmp,irc.zief.pl,
Windows System Monitor:
"C:\Windows\system\winrsc.exe"
```

VII. Summary

Realization of this scenario allowed to prove that the proposed ontology model as well as applied reasoning rules were successfully adapted to detection of single malicious incidents. Then, these incidents were collected and compared with the CP-net models. As a result, Virut malware has been detected.

Ontology presented for malware activities identification together with rules allows to filter out suspicious system activities and strongly supports malware detection mechanism implemented in PRONTO. The effectivenes of signature-based antivirus software is rapidly decreasing. Behavior based methods give promissing effects and should be investigated further on in modern security controls such as one presented here.

References

[1] A. Gostev, *Kaspersky Security Bulletin: Statistics 2008*, http://www.securelist.com/en/analysis/204792052/

[2] D. Maslennikov and Y. Namestnikov, *Kaspersky Security Bulletin. The overall statistics for 2012*, http://www.securelist.com/en/analysis/204792255/

[3] M. Conti, R. Di Pietro, L. Mancini, and A. Mei, “Mobility and cooperation to thwart node capture attacks in MANETs,” *EURASIP J. Wirel. Commun. Netw.*, vol. 2009, no. 1, pp. 8:1–8:13, 2009. http://dx.doi.org/10.1155/2009/945943

[4] C. Raiu, *Virus News: 2012 by the numbers*, http://www.kaspersky.com/

[5] H. Tibbs, S. Ambler-Edwards, and M. Corcoran, *The Global Cyber Game: Achieving strategic resilience in the global knowledge society*, 2013, Defence Academy of The United Kingdom.

[6] S. Adair, R. Deibert, R. Rohozinski, N. Villeneuve, and G. Walton, *Shadows in the cloud: Investigating Cyber Espionage 2.0*, 2010, Information Warfare Monitor Shadowserver Foundation, http://shadows-in-the-cloud.net

[7] M. Szpyrka, B. Jasiul, K. Wrona, and F. Dziedzic, “Telecommunications networks risk assessment with Bayesian networks,” in *Computer Information Systems and Industrial Management*, LNCS, Springer, 2013, vol. 8104, pp. 277–288. http://dx.doi.org/10.1007/978-3-642-40925-7_26

[8] P. Bereziński, M. Szpyrka, B. Jasiul, and M. Mazur, “Network anomaly detection using parameterized entropy,” in *CISIM 2014*, ser. LNCS, K. Saeed and V. Snášel, Eds. Springer, 2014, vol. 8838, pp. 473–486.

[9] Z. Tarapata, M. Chmielewski, and R. Kasprzyk, “An algorithmic approach to social knowledge processing and reasoning based on graph representation – a case study,” in *Intelligent Information and Database Systems*, LNCS, Springer, 2010, vol. 5991, pp. 93-104. http://dx.doi.org/10.1007/978-3-642-12101-2_11

[10] P. Szwed and P. Skrzyński, “A new lightweight method for security risk assessment based on fuzzy cognitive maps,” *Applied Mathematics and Computer Science*, vol. 24, no. 1, pp. 213–225, 2014. http://dx.doi.org/10.2478/amcs-2014-0016

[11] *Verizone. 2012 Data Breach Investigations Report*, http://www.verizonenterprise.com/DBIR/2012/.

[12] A. Takeshi, K. Masaki, and T. Murakami, *Cyber Security Trend – Annual Review 2012*, http://www.nri-secure.co.jp/news/2012/pdf/cyber _security _trend_report_en.pdf

[13] *McAfee and HB Garry Solution Brief. Extend McAfee Total Protection for Endpoint with HBGary Digital DNA and Responder*, http://www.mcafee.com/us/resources/solution-briefs/sb- hbgary.pdf

[14] S. Bobek, K. Porzycki, and G. Nalepa, “Learning sensors usage patterns in mobile context-aware systems,” in *Proceedings of the Federated Conference on Computer Science and Information Systems – FedCSIS*, IEEE, 2013, pp. 993–998.

[15] M. Szpyrka, “Exclusion rule-based systems – case study,” in *Computer Science and Information Technology, IMCSIT*, 2008, pp. 237–242. http://dx.doi.org/10.1109/IMCSIT.2008.4747245

[16] M. Szpyrka, “Analysis of VME-Bus communication protocol – RTCP-net approach,” *Real-Time Systems*, vol. 35, no. 1, pp. 91– 108, 2007. http://dx.doi.org/10.1007/s11241-006-9003-0

[17] G. Nalepa and S. Bobek, “Rule-based solution for context-aware reasoning on mobile devices,” *Computer Science and Information Systems*, vol. 11, no. 1, pp. 171–193, 2014.

[18] K. Jensen and L. Kristensen, *Coloured Petri Nets: Modelling and Validation of Concurrent Systems*, 1st ed. Springer, 2009.

[19] M. Szpyrka and T. Szmuc, “Decision tables in Petri net models,” in *Rough Sets and Intelligent Systems Paradigms*, LNCS, Springer, 2007, vol. 4585, pp. 648–657. http://dx.doi.org/10.1007/978-3-540-73451-2_68

[20] B. Jasiul, M. Szpyrka, and J. Śliwa, “Malware behavior modeling with Colored Petri nets,” in *CISIM 2014*, ser. LNCS, K. Saeed and V. Snášel, Eds. Springer, 2014, vol. 8838, pp. 667–679.

[21] G. Antoniou and F. van Harmelen, *A Semantic Web P rimer*. Cambridge, England: The MIT Press, 2008.

[22] I. Horrocks, P. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, and M. Dean, *SWRL: A Semantic Web Rule Language. Combining OWL and RuleML*, http://www.w3.org/Submission/SWRL/

[23] J. Śliwa and B. Jasiul, “Efficiency of dynamic content adaptation based on semantic description of web service call context,” in *Proceedings - IEEE Military Communications Conference MILCOM 2012, Orlando, USA*, 2012, pp. 1–6. http://dx.doi.org/10.1109/MILCOM.2012.6415810

[24] J. Śliwa, K. Gleba, W. Chmiel, P. Szwed, and A. Głowacz, “ IOEM – Ontology engineering methodology for large systems,” in *Computational Collective Intelligence. Technologies and Applications*, LNCS, Springer, 2011, vol. 6922, pp. 602–611. http://dx.doi.org/10.1007/978-3-642-23935-9_59

[25] A. Tarski, *Introduction to Logic and to the Methodology of Deductive Sciences, Second Edition*. New York: Dover Publications, Inc., 1946.

[26] J. Śliwa and M. Amanowicz, “A mediation service for web services provision in tactical disadvatnaged environment,” in *IEEE Military Communications Conference, MILCOM*, 2008, pp. 1–7. http://dx.doi.org/10.1109/MILCOM.2008.4753323

[27] *Protégé – ontology editor and knowledge-base framework*, http://protege.stanford.edu/

[28] E. Sirin, B. Parsia, B. Cuenca Grau, A. Kalyanpur, and Y. Katz, “Pellet: A practical OWL-DL reasoner,” in *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 5, 2007, pp. 51 – 53. http://dx.doi.org/10.1016/j.websem.2007.03.004

[29] M. Russinovich and A. Margosis, *Windows Sysinternals Administrator’s Reference*. Redmond, Washington, USA: Microsoft Press, 2011.

[30] *Microsoft Technet – Sysinternals*, http://technet.microsoft.com/en-us/sysinternals/

[31] M. Russinovich and B. Cogswell, *Process Monitor v3.05*, http://technet.microsoft.com/pl-pl/sysinternals/bb896645.aspx

[32] *API hooking revealed*, http://www.codeproject.com/Articles/2082/API-hooking-revealed

[33] *EasyHook*, http://easyhook.codeplex.com/

[34] *SNORT*, http://www.snort.org/

[35] *ARAKIS*, http://www.arakis.pl

[36] *Netfilter*, http://www.netfilter.org/

[37] B. Jasiul, R. Piotrowski, P. Bereziński, M. Choraś, R. Kozik, and J. Brzostek, “Federated Cyber Defence System – applied methods and techniques,” in *2012 Military Communications and Information Systems Conference, MCC 2012*, 2012, pp. 145–150.

[38] M. Choraś, R. Kozik, R. Piotrowski, J. Brzostek, and W. Hołubowicz, “Network events correlation for federated networks protection system,” in *Towards a Service-Based Internet*, LNCS, Springer, 2011, vol. 6994, pp. 100–111. http://dx.doi.org/10.1007/978-3-642-24755-2_9

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 111–118
DOI: 10.15439/2014F358
ACSIS, Vol. 2

The influence of using fractal analysis in hybrid MLP model for short-term forecast of close prices on Warsaw Stock Exchange

Michał Paluch
Institute of Applied Computer Science,
Lodz University of Technology
90-924 Łódź, ul. Stefanowskiego 18/22
Tel. +48-533-538-113,
mpaluch@kis.p.lodz.pl

Lidia Jackowska-Strumiłło
Institute of Applied Computer Science,
Lodz University of Technology
90-924 Łódź, ul. Stefanowskiego 18/22
Tel. (+48) 42 631 27 50
lidia_js@kis.p.lodz.pl

***Abstract*—The paper describes a new method of combining Artificial Neural Networks (ANN), technical analysis and fractal analysis for predicting share prices on the Warsaw Stock Exchange. The proposed hybrid model consists of two consecutive modules. In the first step share prices are preprocessed and calculated into moving averages and oscillators. Then, in the next step, they are given to the ANN inputs, which provides the closing values of the asset for the next day. ANN of Multi-Layer Perceptron (MLP) type, and fractal analysis are applied. The hybrid model combining ANN with technical and fractal analysis is compared with hybrid model combining ANN with technical analysis. The obtained results indicate that hybrid model combined with fractal analysis is more accurate and stable in the long run than the hybrid model.**

I. Introduction

AS TRADING systems are becoming more complex, there is also a growing interest in applying artificial intelligence methods, i.e. artificial neural networks [28; 24; 27], fuzzy logic [31] or increasingly popular fractal analysis to support stockbrokers and investors in their decisions aimed at maximizing profits. The financial market, which uses the most advanced IT solutions, provides a variety of products to meet this goal. From all of them, the most popular are financial instruments offered by the Stock Exchange, which may be very profitable, but with a big profit there is also a risk of losing all assets [5]. Recently, artificial neural network (ANN) are gaining in importance for stock quotes time series prediction. The most commonly used artificial neural networks to predict trading signals are the feed-forward neural networks (FNN) [10, 24, 28, 32] of Multi-Layer Perceptron (MLP) type, but also new approaches and ANN structures, like for e.g.: dynamic artificial neural network [11], Probabilistic Neural Networks (PNN) [27], State Space Wavelet Network (SSWN) [3] or a neural-wavelet analysis [14] are still subject of scientific studies. The most common use of ANN on Stock Exchange is: prediction of future stock market indices [3, 24, 26], exchange rates [27], share prices, etc.

Nowadays, hybrid modelling approach is used more often by many researchers. The aim of using hybrid models for Stock Exchange shares forecasting is to reduce risk of failure and obtain the results which are more accurate. Typically, this is done because the underlying process cannot easily be determined. The motivation for combining models comes from the assumption that either one cannot identify the true data generating process or that a single model may not be sufficient to identify all the characteristics of the time series. Different hybrid models were used for this purpose. Khashei and Bijari [18] proposed a new combination of ARIMA and ANN approaches, in which a time series predicted by ANN is considered as nonlinear function of several past observations and random errors. This model was more accurate than ARIMA, ANN and Zhang models [33].

Güresen and Kayakutlu [13] investigated hybrid neural networks which used generalized autoregressive conditional heteroscedasticity (GARCH) and exponential generalized autoregressive conditional heteroscedasticity (EGARCH) to extract new ANN input variables. They also tested combinations of statistical GARCH and EGARCH models with different neural networks [12], i.e. MLP and DAN2 dynamic artificial neural network developed by Ghiassi and Saidane [11]. The lowest error for the testing data in prediction of NASDAQ index was achieved by the use of DAN2 network and next by MLP network. Hybrid GARCH-ANN and EGARCH-ANN models ensured worse results, contrary to expectations.

Majhi, Panda, and Sahoo [19] compared functional link artificial neural network (FLANN), cascaded functional link artificial neural network (CFLANN), and LMS model and also observed that the CFLANN model performs the best prediction of exchange rates followed by the FLANN and the LMS models. Interesting hybrid approach combining technical analysis and ANN for trading systems development was proposed by Witkowska and Marcinkiewicz [29]. 15 trading systems were designed for the Warsaw Stock Exchange future contracts and compared. Five strategies of investment decisions were investigated, including four based on

978-83-60810-58-3/$25.00

technical analysis indicators, which were combined with three methods of the WIG20 index future closing prices forecasting. The final conclusion was that the combination of the technical analysis and artificial intelligence in order to gain profit from trading on the Stock Exchange can bring much better investment results than trade in the traditional way. The best results for the WIG20 index time series forecasting were obtained by the use of ANN of MLP type with a set of about 30 input variables, which were divided in 3 subsets: variables related to the WIG20 index Close prices, variables related to the technical analysis indicators, variables related to the external factors.

In this paper, a new hybrid analytical and ANN model is proposed, which combines ANN with technical and fractal analysis. Previously, the hybrid models combining technical analysis with ANN without fractal analysis were compared with purely ANN based approach [22].

It will be shown that hybrid ANN model which uses technical and fractal analysis is more stable in the long run than hybrid ANN model using only technical analysis and that fractal analysis reduces the error of shares forecasting.

II. TECHNICAL ANALYSIS INDICATORS

Technical analysis indicators are used to determine trend of the market, the strength of the market, and the direction of the market. Some technical analysis indicators can be quantified in the form of an equation or algorithm. Others can show up as patterns (e.g., head and shoulders, trend lines, support, and resistance levels). At some point, the technical analyst will receive a signal. This signal is the result of one technical analysis indicator or a combination of two or more indicators. The signal indicates to the technical analyst a course of action whether to buy, sell, or hold [29].

The most commonly used technical analysis indicators are moving averages and oscillators [20], which were selected for the proposed approach. These include the following:

- Moving averages:

a. Expotential (5-, 10-, 20-days) – EMA (Expotential Moving Average)

$$EMA_{N,C}(k) = \frac{C(k) + aC(k-1) + a^2C(k-2) + ... + a^{N-1}C(k-N+1)}{1 + a + a^2 + ... + a^{N-1}} \quad (1)$$

where:

a- coefficient

b. Envelopes (3% error with 20-days average)

- Oscillators

a. ROC - Rate of Change (5-, 10-, 20-days) – determines the rate of price changes in a given period (usually 10 days)

$$ROC_N(k) = C(k) / C(k-N) \quad (4)$$

b. RSI - Relative Strength Index – i.e. the measure of overbought / oversold market. It assumes values in the range of 0-100. For values greater than 70 it is considered that the market is buyout. When oscillator values are below 30, it signifies that market is sold out. In the case of periods of strong trends it is assumed that the market is buyout when RSI> 80 (at the time of a bull market) and sold out for RSI <20 (during a bear market).

For:

C(k) > C(k-1), U(k) = C(k) – C(k-1)

C(k) < C(k-1), D(k) = |C(k) – C(k-1)|

$$RSI(k) = 100 - \left[\frac{100}{1 + \frac{EMA_{N,U}(k)}{EMA_{N,D}(k)}} \right] \quad (5)$$

where

U(k) – average increase in the k-th day

D(k) – average decrease in the k-th day

c. Stochastic oscillator (K%D) – determines the relation between the last closing price and the range of price fluctuations in the given period. The result belongs to the range of 0-100. K% D> 70 is interpreted as the closing price near the top of the range of its fluctuations, and K% D <30 points to the fact that prices are shaping near the lower limit of that range.

$$K\%D(k) = 100 * \left[\frac{C(k) - L(14)}{H(14) - L(14)} \right] \quad (6)$$

where:

L(14) – the lowest price from last fourteen days

H(14) - the highest price from last fourteen days

d. Moving Average Convergence/Divergence (MACD) is the difference between two moving averages. On the graphs, it usually occurs with 10- day, exponential moving average (called the signal line). The intersection of the signal line (SL) with the MACD line

coming from the bottom is a buying signal, while with the line from the top - a selling signal.

$$MACD(k) = EMA_{12,C}(k) - EMA_{26,C}(k) \quad (7)$$

$$SL(k) = EMA_{9,MACD}(k) \quad (8)$$

e. Accumulation/Distribution (AD) indicator presents whether price changes are accompanied by increased accumulation and distribution movements.

$$AD(k) = V(k) * \frac{C(k) - L(k) - [H(k) - C(k)]}{H(k) - L(k)} \quad (9)$$

where:

V(k) - total number of shares which were rotated on k day

f. Bollinger Oscillator

Its construction is based on Bollinger bands. Bollinger oscillator informs when market is overbought or oversold.

$$BOS_k = \frac{C_{k+(N-1)} - SMA_N(C(k))}{StandardDev(k)} \quad (10)$$

For the purpose of counting highest errors between predicted value and real CLOSE value, the following formulas have been used:

a. The highest prediction error per month

E_{max} = Highest difference between real CLOSE value and predicted by ANN value per month (11)

b. Arithmetical mean of E_{max} value per tested period of time

$$\overline{E_{max}} = \frac{1}{N}\sum_{i=1}^{N} E_{max_i} \quad (12)$$

III. FRACTAL ANALYSIS

Recently it can be seen that fractal market hypothesis is constantly expanding. It was presented for the first time by Peters [7] in 1994, and is based on chaos theory [8]. Fractal shapes can be formed in many ways. The simplest is a multiple iteration of generating rule (e.g. the Koch curve or Sierpinski triangle). They are generated in deterministic way and all have fractal dimension. There are also random fractals, like stock prices, which are generated with the use of probability rules.

Performing a fractal analysis is based on identification of fractal dimension. To do this, chart has to be divided into N small elements with S surface. The relationship between the number of objects N_1 and N_2, which are used to cover the first and second graph with objects of surface size, respectively S_1 and S_2, describes the relationship [9]:

$$\frac{N_2}{N_1} = \left(\frac{S_1}{S_2}\right)^D \quad (13)$$

$$D = \frac{\log\left(\frac{N_1}{N_2}\right)}{\log\left(\frac{S_1}{S_2}\right)} \quad (14)$$

where:

D – fractal dimension

In order to measure fractal dimension on stock exchange, we need to divide the given period of time by two. For each period, share prices curve have to be divided into N pieces. It can be done by dividing the subtraction result of highest and lowest value on graph in given period of time, by this period:

$$N_{1T}(k) = \frac{H_T(k) - L_T(k)}{T} \quad (15)$$

$$N_{2T}(k) = \frac{H_{2T}(k) - L_{2T}(k)}{T} \quad (16)$$

$$N_{0-2T}(k) = \frac{H_{0-2T}(k) - L_{0-2T}(k)}{2T} \quad (17)$$

$$D = \frac{\log\left(\frac{N_{1T} + N_{2T}}{N_{(0-2)T}}\right)}{\log\left(\frac{2T}{T}\right)} = \frac{\log(N_{1T} + N_{2T}) - \log(N_{(0-2)T})}{\log(2)} \quad (18)$$

where:

$H_T(k)$ – the highest share price in the first period T

$H_{2T}(k)$ – the highest share price in the second period (from T till 2T)

$H_{0\text{-}2T}(k)$ – the highest share price in 2T period

$L_T(k)$ – the lowest share price in the first period T

$L_{2T}(k)$ – the lowest share price in the period from T till 2T

$L_{0\text{-}2T}(k)$ – the lowest share price in 2T period

Fractal dimension is used in this paper in Fractal Moving Average (FRAMA). This moving average is based on Expotential Moving Average (eq. 1) where a coefficient is constructed with the use of fractal dimension:

$$a = \exp(-4.6*(D-1)) \quad (19)$$

IV. APPLICATION OF TECHNICAL ANALYSIS AND ANN FOR PREDICTION OF CLOSING PRICES

Closing price of the asset for the next day is one of the most important parameters for investors, who plan to make transactions at the Stock Exchange. In this work a hybrid approach combining technical analysis with ANN and a hybrid approach combining technical and fractal analysis with ANN is being compared. The main idea of the proposed method is shown in Fig. 1. Technical analysis methods are used to calculate moving averages and oscillators, which are important market indicators. These are the inputs of ANN, which predicts the CLOSE value of the next day. The aim of this paper was to investigate if the proposed data pre-processing with market indicators calculation connected with the use of fractal dimension would improve the ANN effectiveness in the CLOSE value prediction.

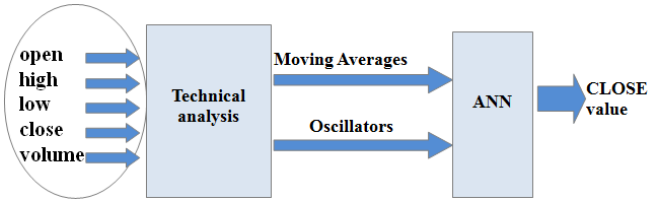


Fig. 1. Processing scheme for predicting course of a CLOSE value for the next day [19]

The programming application was designed and implemented for the data collecting and pre-processing. The calculated moving averages and oscillators were used for neural network training and testing. Feedforward networks of Multi-Layer Perceptron (MLP) type [25] trained with the Levenberg-Marquardt algorithm were used for the CLOSE value prediction. The choice of the ANN input variables presented in this paper was made based on the experience and knowledge of a stock market expert.

The model structure was overtaken from the Authors' previous experience with hybrid analytical-neural approach in engineering applications, which ensured better modelling results as pure ANN solutions [16, 17]. Thanks to the data preprocessing the designed model is efficient and ANN has a simple structure.

V. EXPERIMENTAL RESEARCH

Research was conducted for 2 exemplary companies (Vistula SA and Budimex SA) appearing on the Warsaw stock market since 2.01.1996 until 14.04.2014. The aim of the research was to compare the results of short-term prediction of two hybrid models. Their hybridity differs in terms of the technical analysis indicators and fractal moving averages used as ANN inputs (Table II). The research was performed with the use of Java and Encog 3.2 library, creating ANN of MLP type. Each network consists of an input, hidden and output layer. A common feature of all of the tested network architectures is a small number of input nodes and neurons in the hidden layer, and only one neuron in the output layer. Too many neurons would increase the network training error and could cause learning time extension [5]. The relations between the number of input nodes and the number of neurons in the hidden layer were tested for the combinations shown in Table 1.

Table I

COMBINATIONS OF THE TESTED MLP ARCHITECTURES

Input layer	Hidden layer	Output layer
n	n+1	1
	1.5n	
	2n-1	
	2n+1	

where n – number of neurons (n = 4, 5, 6 neurons)

Market indicators for the input data were selected based on the literature [1, 4, 10, 20, 32] and an advice of a stock market expert. ANN training was performed according to the following rules:

1. All entered data were normalized using the following heuristic formula:
$$(\text{Value}/\text{Value}_{max})*0.8 + 0.1 \quad (10)$$
2. The results of each company were divided into two groups: learning data and testing data in the proportion 70:30 [19]
3. Neural networks were taught with the Levenberg-Marquardt algorithm [16, 23].
4. For each ANN architecture and each set of input data, eight neural networks were trained, and the ANN with the smallest medium square error (MSE) for the testing data has been selected as the best one.

The hybrid model structures combining MLP with technical analysis with or without using the fractal dimension, and also the obtained results are gathered in Table II. Technical analysis indicators, which are ANN inputs are listed in the second column. The MLP (7-15-1) structure means, that it consists of seven input nodes, fifteen neurons in a hidden layer and one neuron in an output layer.

The results of short-term forecast of CLOSE value of Vistula SA shares predicted with the use of Hybrid MLP (7-15-1) model (no. 1 in Table 2) and Hybrid MLP (7-15-1) model with fractal dimension (no. 2 in Table 2) are shown in

figures 2 and 3. In the first case (Fig. 2) share prices of Vistula SA are in horizontal trend and in the second period of time (Fig. 3) in downward trend.

Table II

HYBRID MODELS WITH SELECTED ANN STUCTURES, FOR WHICH THE BEST RESULTS WERE ACHIEVED

No.	ANN inputs	Model structure	Transfer function	Periods	Training error	Testing error
1.	RSI MACD AD BO EMA_{k-4} EMA_{k-9} ROC	Hybrid MLP(7-15-1)	sigmoidal	7000	0.0327	0.023
2.	RSI MACD AD BO $FRAMA_{k-4}$ $FRAMA_{k-9}$ ROC	Hybrid MLP(7-15-1)	sigmoidal	7000	0.0312	0.021

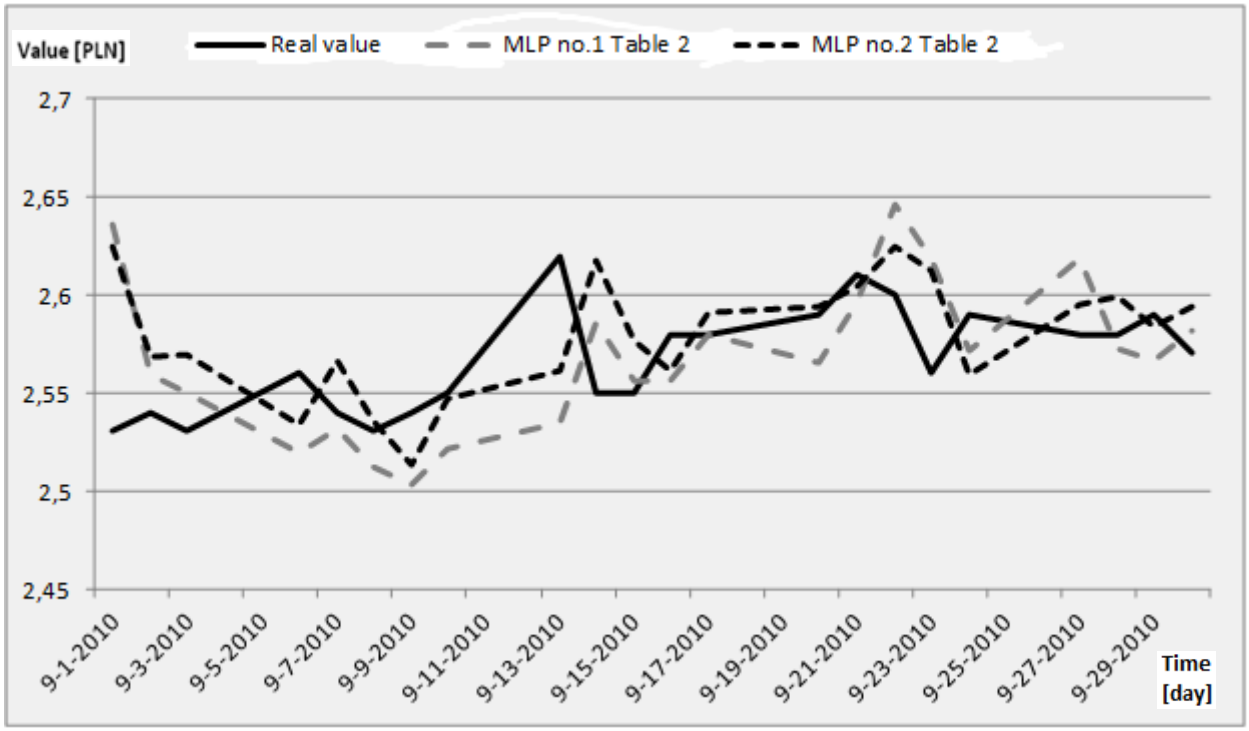


Fig. 2. Short-term forecast of two hybrid models with MLP (7-15-1) network: no. 1 in Table 2 and no. 2 in Table 2 (with fractal dimension) and real CLOSE value of Vistula SA shares in September 2010.

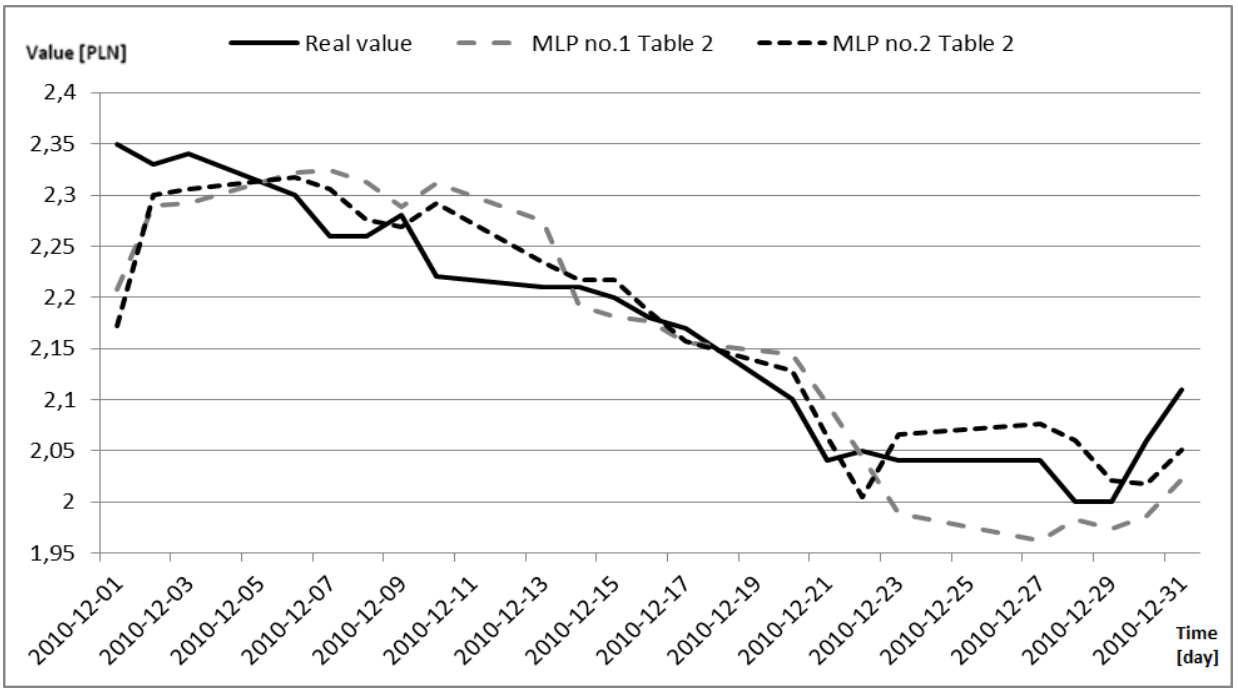


Fig. 3 Short-term forecast of two hybrid models with MLP (7-15-1) network: no. 1 in Table 2 and no. 2 in Table 2 (with fractal dimension) and real CLOSE value of Vistula SA shares in December 2010.

Comparison of the obtained results shows that prediction of hybrid model with FRAMA (no. 2 in Table 2) was more accurate while share prices were in distinct trend. Example is being shown in figure 3, where Vistula SA shares have been in downward trend on December 2010. On the other hand, both hybrid models provide similar results when share prices are in horizontal trend, which can be seen in figure 2.

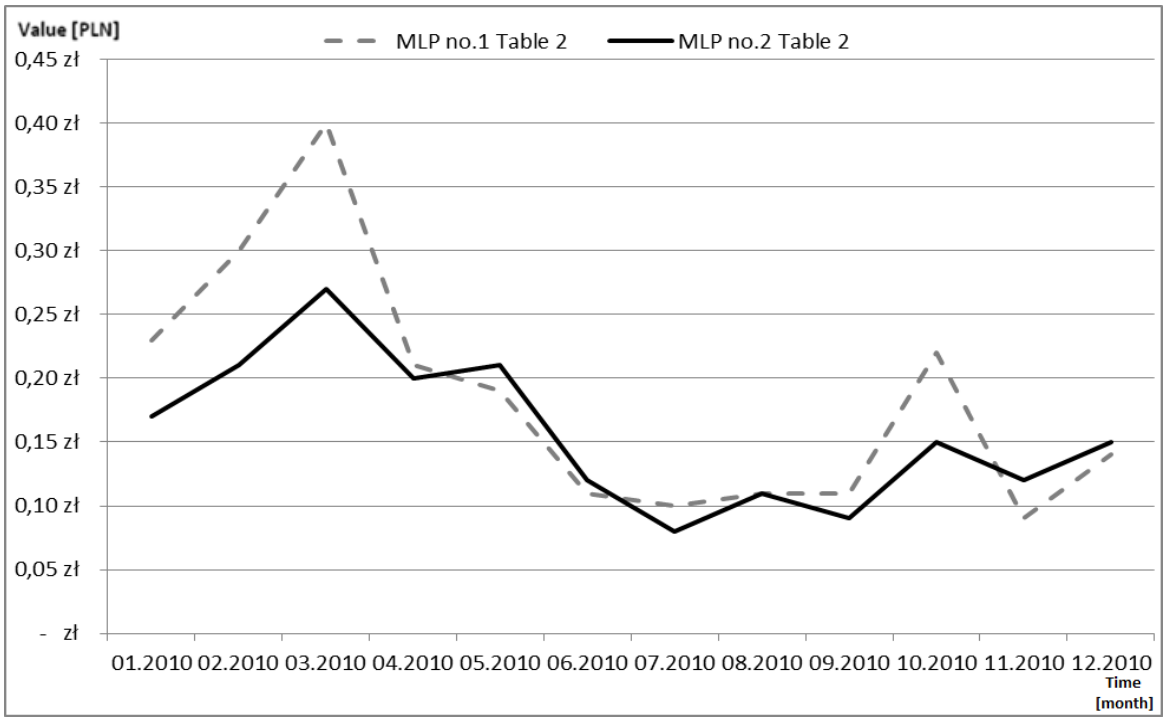


Fig. 4. Comparison of E_{max} values between hybrid model with hybrid MLP (7-15-1) (no. 1 in Table 2) and hybrid MLP (7-15-1) (no. 2 in Table 2) for Vistula SA

To assess which model is more accurate and stable, the maximum absolute errors E_{max} of prediction in the period of one year were compared. Maximum absolute error was calculated as the maximum absolute difference between true CLOSE value and the model prediction for each month. Summary of the results for hybrid models with and without FRAMA are presented in figures 4 and 5. Similar results have been achieved for Vistula SA (Fig. 4) and for Budimex SA (Fig. 5). For a comparison results are shown for the same period of time.

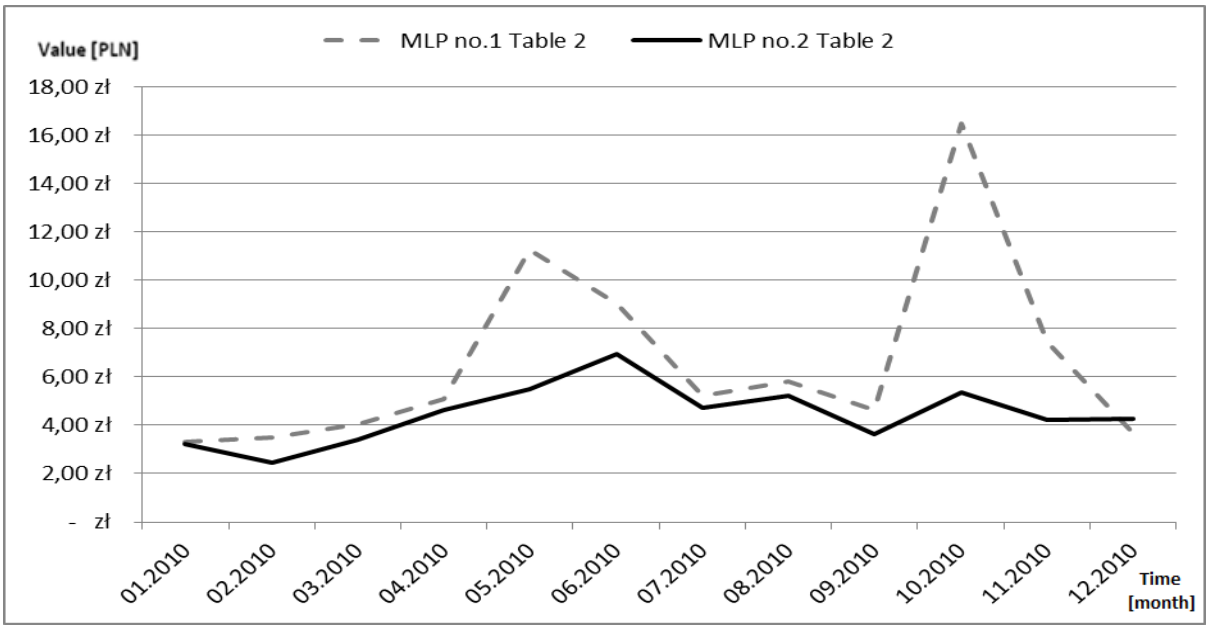


Fig. 5. Comparison of E_{max} values between hybrid model with hybrid MLP (7-15-1) (no. 1 in Table 2) and hybrid MLP (7-15-1) (no. 2 in Table 2) for Budimex SA

VI. CONCLUSIONS

The obtained experimental results for Vistula SA and for Budimex SA suggest that the hybrid analytical-neural model combined with the fractal analysis yields better approximation of the real shares values than the hybrid model without fractal analysis. Comparison of maximum error E_{max} values for two hybrid models: with and without fractal analysis (no. 2 and 1 in table 2), during the audited period from January 2010 to December 2010, allows to conclude that the proposed prediction hybrid model of hybrid ANN with fractal moving average is more stable and accurate. Thus, the more stable hybrid model reduces the probability of influence of false signals coming from the market in decision-making process, simultaneously increasing potential profits.

Finally, the results lead also to the conclusion that the proposed hybrid models forecast correct direction of CLOSE price changes. Therefore, they can be used as a basis for the decision-making system, which would be used to support investor decisions on the Warsaw Stock Exchange.

VII. REFERENCES

[1] Bensignor R.: New Concepts in Technical Analysis. Wig-Press, Warsaw 2004 (in Polish).

[2] Box G. E. P., & Jenkins G. M. (1976). Time Series Analysis. Forecasting and control. Holden-Day Inc., San Francisco, California, USA.

[3] Brdyś M. A., Borowa A., Idźkowiak P., Brdyś M. T.: Adaptive Prediction of Stock Exchange Indices by State Space Wavelet Networks. Int. J. Appl. Math. Comput. Sci., 2009 , Vol. 19 , No. 2 , 337-348. DOI: 10.1.1.390.8001

[4] Bulkowski Thomas N., Formation Analysis on Stock Charts. Linia, Warsaw 2011 (in Polish)

[5] Dębski W.: Financial Market and it mechanisms. PWN, Warsaw 2010 (in Polish)

[6] Dourraa H., Siyb P. (2002). Investment using technical analysis and fuzzy logic. Fuzzy Sets and Systems, 127, 221-240.

[7] Drabik E.: Applications of game theory to invest in securities, Wydawnictwo Uniwersytetu w Białymstoku, Bialystok 2000. (in Polish)

[8] Ehlers J.: Fractal Adaptive Moving Average", Technical Analysis of Stock & Commodities" October 2005.

[9] Ehlers J.: "Cybernetics Analysis For Stocks And Futures", John Wiley & Sons, New York 2004.

[10] Gately E. (1995). Neural Networks for Financial Forecasting, New York: Wiley.

[11] Ghiassi M., Saidane H., Zimbra D. K.: Dynamic artificial neural network model for forecasting time series events. *International Journal of Forecasting,* 2005, Vol. 21, pp. 341-362. DOI: DOI:10.1016/j.ijforecast.2004.10.008

[12] Güresen E., Kayakutlu G.: Forecasting Stock Exchange Movements Using Artificial Neural Network Models and Hybrid Models. In *IFIP International Federation for Information Processing,* 2008, Volume 288; Intelligent Information Processing IV; Zhongzhi Shi, E. Mercier-Laurent, D. Leake; (Boston: Springer), pp. 129–137.

[13] Güresen E., Kayakutlu G., Daim T. U.: Using artificial neural network models in stock market index prediction. *Expert Systems with Applications,* 2011, Vol. 38, pp. 10389-10397. DOI: 10.1016/j.eswa.2011.02.068

[14] Hajto P. (2012). A Neural Economic Time Series Prediction with the Use of a Wavelet Analysis. *Schedae Informaticae,* 11, 115-132.

[15] Hamzacebi, C., Akay, D., & Kutay, F. (2009). Comparison of direct and iterative artificial neural network forecast approaches in multi-periodic time series forecasting. *Expert Systems with Applications,* 36, 3839-3844. DOI: 10.1016/j.eswa.2008.02.042

[16] Jackowska-Strumiłło L.: Hybrid Analytical and ANN-based Modelling of Temperature Sensors Nonlinear Dynamic Properties, *The 6th International Conference on Hybrid Artificial Intelligence Systems, HAIS 2011,* Wroclaw, Poland, 23-25 May, Lecture Notes in Artificial Intelligence, LNAI 6678, 2011, Springer-Verlag, Part I, pp. 356-363. DOI: 10.1007/978-3-642-21219-2_45

[17] Jackowska-Strumiłło L., Jackowski T., Chylewska B., Cyniak D.: *Application of hybrid neural model to determination of selected yarn parameters. Fibres & Textiles in Eastern Europe,* ISSN 1230-3666, 1998, Vol. 6, Nr 4 (23), pp. 27-32.

[18] Khashei, M., & Bijari, M. (2010). An artificial neural network (p, d, q) model for timeseries forecasting. *Expert Systems with Applications,* 37(1), 479–489. DOI: 10.1016/j.eswa.2009.05.044

[19] Majhi, R., Panda, G., & Sahoo, G. (2009). Efficient prediction of exchange rates with low complexity artificial neural network models. *Expert Systems with Applications,* 36, 181-189. DOI: 10.1016/j.eswa.2007.09.005

[20] Murphy J. J.: *Technical Analysis of Financial Markets.* Wig-Press, Warsaw, 2008 (in Polish).

[21] Narendra K. S., Parthasarathy K.: Identification and control of dynamics systems using neural networks, *IEEE Transactions on Neural Networks,* 1990, vol.1, no. 1, pp. 4-27 DOI: 10.4236/ica.2011.23021

[22] Paluch M., Jackowska-Strumiłło L.: Prediction of closing prices on the Stock Exchange with the use of artificial neural networks. *Image Processing & Communication,* 2012, Vol. 17, No. 4, pp. 275-282.

[23] Rutkowski L.,: *Methods and Techniques of Artificial Intelligence.* PWN, Warsaw 2009 (in Polish)

[24] Sutheebanjard, P., Premchaiswadi, W.: Stock Exchange of Thailand Index Prediction Using Back Propagation Neural Networks. In: *Proc. of the Second International Conference on Computer and Network Technology (ICCNT),* 2010, Bangkok, pp. 377-380. DOI: 10.1109/ICCNT.2010.21

[25] Tadeusiewicz R.: *Artificial Neural Networks.* Warsaw 1993 (in Polish).

[26] Tadeusiewicz R.: *Discovering Neural Networks.* Cracow 2007 (in Polish).

[27] Tilakaratne C. D., Morris S. A., Mammadov M. A., Hurst C. P. (2007). Predicting Stock Market Index Trading Signals Using Neural Networks. In: *Proc. of the 14th Annual Global Finance Conference (GFC 2007),* Melbourne, Australia, pp. 171-179 (Sep. 2007)

[28] Witkowska D.: *Artificial Neural Networks and statistical methods. Selected financial issues,* C. H. Beck, Warsaw 2002, (in Polish)

[29] Witkowska D., & Marcinkiewicz E. (2005). Construction and Evaluation of Trading Systems: Warsaw Index Futures. *International Advances in Economic Research,* 11, 83-92. DOI: 10.1007/s11294-004-7496-7

[30] Zaremba A.: *Stock Exachange,* 2010 (in Polish)

[31] Zhou X. S., Dong M. (2004). Can fuzzy logic make technical analysis 20/20? *Financial Analyst Journal,* 60, 54-75. DOI: 10.2469/faj.v60.n4.2637

[32] Zieliński J.: *Intelligent management systems – theory and practice.* Warsaw 2000 (in Polish).

[33] Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing,* 50, 159-175.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 119–128
DOI: 10.15439/2014F387
ACSIS, Vol. 2

Fuzzy Logic Rules Modeling Similarity-based Strict Equality

Ginés Moreno, Jaime Penabad and Carlos Vázquez
Faculty of Computer Science Engineering
University of Castilla-La Mancha
Albacete (02071), Spain
Email: {Gines.Moreno,Jaime.Penabad,Carlos.Vazquez}@uclm.es

***Abstract*—A classical, but even nowadays challenging research topic in declarative programming, consists in the design of powerful notions of "equality", as occurs with the flexible (fuzzy) and efficient (lazy) model that we have recently proposed for hybrid declarative languages amalgamating functional-fuzzy-logic features. The crucial idea is that, by extending at a very low cost the notion of "strict equality" typically used in lazy functional (HASKELL) and functional-logic (CURRY) languages, and by relaxing it to the more flexible one of similarity-based equality used in modern fuzzy-logic programming languages (such as LIKELOG and BOUSI~PROLOG), similarity relations can be successfully treated while mathematical functions are lazily evaluated at execution time. Now, we are concerned with the so-called *Multi-Adjoint Logic Programming approach*, MALP in brief, which can be seen as an enrichment of PROLOG based on *weighted rules* with a wide range of fuzzy connectives. In this work, we revisit our initial notion of SSE (*Similarity-based Strict Equality*) in order to re-model it at a very high abstraction level by means of a simple set of MALP rules. The resulting technique (which can be tested on-line in `dectau.uclm.es/sse`) not only simulates, but also surpass in our target framework, the effects obtained in other fuzzy logic languages based on similarity relations (with much more complex/reinforced unification algorithms in the core of their procedural principles), even when the current operational semantics of MALP relies on the simpler, purely syntactic unification method of PROLOG.**

***Index Terms*—Equality, Similarity, Fuzzy Logic Programming**

I. INTRODUCTION

THANKS to the high expressive power and the rule-based nature of declarative languages, their influences are growing in the design of intelligent systems and techniques related with artificial/computational intelligence, expert systems, soft-computing and so on. In particular, *Logic Programming* (LP) [1] has been widely used for problem solving and knowledge representation in the past. Nevertheless, traditional logic programming languages are not able to treat with partial truth. *Fuzzy Logic Programming* is an interesting and still growing research area that agglutinates the efforts for introducing Fuzzy Logic into Logic Programming, in order to provide these traditional languages with techniques or constructs (coming up from the mathematical background of fuzzy logic [2]) to deal with uncertainty in a natural way. In the last two decades, several fuzzy logic programming languages have been developed where, in essence, the classical SLD resolution principle of PROLOG [3] (based on syntactic unification) has been replaced by a fuzzy variant of itself, with the aim of dealing with partial truth and reasoning with uncertainty in a natural way. Most of these languages implement (extended versions of) the resolution principle introduced by Lee [4], such as Elf-Prolog [5], Fril [6], F-Prolog [7] and MALP [8]. There exists also a family of fuzzy languages based on sophisticated unification methods [9] to cope with similarity/proximity relations, as occurs with LIKELOG [10], SQLP [11] and BOUSI~PROLOG [12], [13] (some related approaches based on *probabilistic logic programming* can be found in [14], [15]).

On the other hand, during the last three decades of investigation in the field of the integration of declarative programming paradigms (functional, fuzzy and logic), the scientific community of the area has produced important and advanced contributions related to both theoretical and practical aspects. However, whereas the functional and logic programming styles have been successfully integrated in the past and, as said before, more recently fuzzy logic has also been introduced into the logic programming paradigm, there is not precedent for a total integration of all these frameworks, apart from our preliminary approach presented in [16].

In [17], we gave a new step in this last sense, by proposing a method combining different equality models traditionally supported by each one of these declarative paradigms. It is important to take into account that an appropriate notion of equality has a crucial importance when designing the repertoire of expressive resources for a particular declarative language. In general, when we use the term "equality" in declarative programming, there are several different meanings depending of the concrete paradigm being considered. A representative (not exhaustive) list of some cases could be:

- **Syntactic equality.** It is the simplest equality model used in the context of classical pure logic programming (as occurs with PROLOG, but also in the fuzzy logic language MALP) which is simply concerned with syntactic identity. In this sense, two element are considered "equal" if they have exactly the same syntax. For instance, $f(a)$ is equal to $f(a)$ but not to $g(b)$.

This work was supported by the EU (FEDER), and the Spanish MINECO Ministry (Ministerio de Economía y Competitividad) under grant TIN2013-45732-C4-2-P.

- **Strict equality.** When considering lazy languages, both pure functional (HASKELL [18]) and integrated functional-logic (CURRY [19]) languages, this new equality notion is the only applicable one in a lazy setting, mainly due to the possible presence of non terminating functions. For instance, if the evaluation of $f(a)$ does not finish then we can not say that $f(a)$ is strictly equal to itself. And, on the contrary, two terms with different syntax, such as $g(b)$ and $h(c)$, could be proved equal if they produce the same final value (for example 0) after being evaluated by rewriting or narrowing.
- **Similarity-based equality.** As we will see in Section II, this model emerges as a direct consequence of several attempts for fuzzifying the original notion of syntactic equality, which are appreciable in the design of fuzzy logic languages such as LIKELOG, SQLP and BOUSI~PROLOG. In this case, the idea is to allow the presence of a set of the so-called "similarity/proximity equations" between symbols of a given program. So, if we have a program with the equations $eq(a, b) = 0.5$ and $eq(f, g) = 0.3$ then, it could be proved that expressions $f(a)$ and $g(b)$ are similar with a concrete truth degree.

Here, we recall from [17] our original definition of SSE (*Similarity-based Strict Equality*), initially modeled by means of a set of rewriting rules and which fuses the last two equality versions above. The crucial idea of our method is to simply add to a given functional-logic program (written in CURRY, for instance) a set of rewriting rules defining the new symbol $\approx$:$\approx$ which captures similarities and thus, is implemented at a very low cost by simply performing a syntactic pre-process on programs.

The main goal of this paper is to adapt such definition to the MALP framework. In Section III we will see that SSE admits a much more natural formulation by means of a set of MALP rules instead of using rewriting rules. Moreover, although this fuzzy programming style is based on pure syntactic unification, our method introduces a similarity-based equality model without altering its core, which is useful not only for testing if two ground data terms are comparable (as occurs too with more complex languages -LIKELOG, BOUSI~PROLOG- with extended unification algorithms), but also for producing complete lists of similar terms (not achievable by LIKELOG and BOUSI~PROLOG). Although the technique is recasted from [20], the main contribution of the present paper consists in proving some interesting formal properties for it. Moreover, before concluding in Section V, we describe in Section IV some implementation details regarding the two main processes needed for effectively embedding SSE into MALP: after performing the reflexive-symmetric-transitive closure of a set of similarity equations for obtaining a similarity relation, then it is easily translated into a set of MALP rules modeling SSE.

II. SIMILARITY RELATIONS AND FUZZY LOGIC PROGRAMMING

As we have just said, although in principle it is not the case of MALP (whose operational semantics uses syntactic unification on its core), some fuzzy languages such as LIKELOG, SQLP and BOUSI~PROLOG are able to treat with the mathematical notions of similarity (and proximity), by incorporating a flexible variant of unification -beyond the simpler case of PROLOG- on their procedural principles.

A similarity relation is a mathematical notion able to manipulate alternative instances of a given entity that can be considered equals with concrete truth degrees. Similarity relations are closely related with equivalence relations (and, then, to closure operators) [21]. Let us recall that a T-norm $\wedge$ in $[0, 1]$ is a binary operation $\wedge : [0, 1] \times [0, 1] \to [0, 1]$ associative, commutative, non-decreasing in both the variables, and such that $x \wedge 1 = 1 \wedge x = x$ for any $x \in [0, 1]$. Formally, a *similarity relation* $\Re$ on a domain $\mathcal{U}$ is a fuzzy subset $\Re : \mathcal{U} \times \mathcal{U} \to [0, 1]$ of $\mathcal{U} \times \mathcal{U}$ such that, $\forall x, y, z \in \mathcal{U}$, the following properties hold: reflexivity $\Re(x, x) = 1$, symmetry $\Re(x, y) = \Re(y, x)$ and transitivity $\Re(x, z) \geq \Re(x, y) \wedge \Re(y, z)$. It is important to note that this last property is not required when considering *proximity relations*. In order to simplify our developments, as in [9], we assume that $x \wedge y$ is the minimum between the two elements $x, y \in [0, 1]$.

A very simple, but effective way, to introduce similarity relations into pure logic programming, generating one of the most promising ways for the integrated paradigm of fuzzy logic programming, consists of modeling them by a set of the so-called *similarity equations* of the form $eq(s1, s2) = \alpha$, with the intended meaning that $s1$ and $s2$ are predicate/function symbols of the same arity with a similarity degree α. As in [16], we assume here that the intended similarity relation $\Re$ associated to a given program $\mathcal{P}$, is induced from the (safe) set of similarity equations of $\mathcal{P}$, verifying that the similarity degree of two symbols s_1 and s_2 is 1 if $s_1 = s_2$ or, otherwise, it is recursively defined as the transitive closure of the similarity equations.

This approach is followed, for instance, in the fuzzy logic languages LIKELOG [10] and BOUSI~PROLOG [12], where a set of usual PROLOG clauses are accompanied by a set of similarity equations playing an important role at (fuzzy) unification time. Instead of classical *syntactic unification*, we speak now about *weak unification* [12]. Of course, the set of similarity equations is assumed to be safe in the sense that each equation connects two symbols of the same arity and nature (both predicates or both functions) and the properties of the definition of similarity relation are not violated, as occurs, for instance, with the wrong set $\{eq(a, b) = 0.5,\ eq(b, a) = 0.9\}$ which, in particular, it does not satisfy the symmetric property.

Example 2.1: Following [10], if we consider a database of books containing the fact "`book(horror,drakula)`", then goal "`?-book(adventures,Title)`" should not have classical solution in the case that there were no rule in the database unifying with atom "`book(adventures,Title)`". Nevertheless, it seems reasonable that the user considers the words "`adventures`" and "`horror`" to be *similar* with a certain degree. More precisely, if the user introduces a similarity equation like "`eq(adventures, horror) = 0.9`" into a

LIKELOG or BOUSI~PROLOG interpreter, the system should successfully respond with a computed answer incorporating the corresponding truth degree "`0.9`" (i.e, something like the 90 % of credibility) to substitution "`Title/ drakula`", as obviously expected.

III. SSE FOR/WITH MULTI-ADJOINT LOGIC PROGRAMMING

In this section we firstly summarize the main features of the MALP language[1], next we introduce the "*Fuzzy LOgic Programming Environment for Research*", $\mathcal{FLOPER}$ in brief, developed in our research group (see [26], [27] and visit `http://dectau.uclm.es/floper/`) and finally, we illustrate and formally prove the properties of our new MALP-based model of SSE according to Figure 2.

A. MALP

We work with a first order language containing variables, function symbols, predicate symbols, constants, quantifiers ($\forall$ and $\exists$), and several arbitrary connectives such as implications ($\leftarrow_1, \leftarrow_2, \ldots, \leftarrow_m$), conjunctions ($\&_1, \&_2, \ldots, \&_k$), disjunctions ($\vee_1, \vee_2, \ldots, \vee_l$), and general hybrid operators ("aggregators" $@_1, @_2, \ldots, @_n$), used for combining/propagating truth values through the rules, and thus increasing the language expressiveness. Additionally, our language contains the values of a multi-adjoint lattice $\mathcal{L} = \langle L, \preceq, \leftarrow_1, \&_1, \ldots, \leftarrow_n, \&_n \rangle$, equipped with a collection of adjoint pairs $\langle \leftarrow_i, \&_i \rangle$ (where each $\&_i$ is a conjunctor intended to the evaluation of *modus ponens*) verifying the so-called *adjoint property*: $\forall x, y, z \in L, \quad x \preceq (y \leftarrow_i z)$ if and only if $(x \,\&_i\, z) \preceq y$. The set of truth values L may be the carrier of any complete bounded lattice, as for instance occurs with the set of real numbers in the interval $[0, 1]$ with their corresponding ordering $\leq$. A *rule* is a formula $[A \leftarrow_i \mathcal{B}\ with\ \alpha]$, where A is an atomic formula (usually called the *head*), $\mathcal{B}$ (which is called the *body*) is a formula built from atomic formulas $B_1, \ldots, B_n$ ($n \geq 0$), truth values of L and conjunctions, disjunctions and aggregators, and finally $\alpha \in L$ is the "weight" or *truth degree* of the rule. A rule with empty body, written $[A\ with\ \alpha]$, is called *fact*. Consider, for instance, the following program $\mathcal{P}$ composed by three rules with associated multi-adjoint lattice $\langle [0,1], \leq, \leftarrow_{\mathtt{P}}, \&_{\mathtt{P}}, \leftarrow_{\mathtt{G}}, \&_{\mathtt{G}} \rangle$ (where labels `P` and `G` mean for *Product logic* and *Gödel intuitionistic logic*, respectively, with the following connective definitions: "$\leftarrow_{\mathtt{P}} \ (x, y) = \min(1, x/y)$", "$\&_{\mathtt{P}}(x, y) = x * y$", "$\leftarrow_{\mathtt{G}} \ (x, y) = 1$ if $y \leq x$ or x otherwise" and "$\&_{\mathtt{G}}(x, y) = min(x, y)$"):

$$\begin{array}{lllll} \mathcal{R}_1: & p(X) & \leftarrow_{\mathtt{P}} \quad q(X,Y) \ \&_{\mathtt{G}}\ r(Y) & with & 0.8 \\ \mathcal{R}_2: & q(a,Y) & & with & 0.9 \\ \mathcal{R}_3: & r(b) & & with & 0.7 \end{array}$$

[1]As said before, this fuzzy language uses a syntax near to PROLOG and enjoys high level of flexibility, for which we give some theoretical/practical reinforcements in our precedent works [22], [23], [24], [25].

In order to describe the procedural semantics of the multi–adjoint logic language, in the following we denote by $\mathcal{C}[A]$ a formula where A is a sub-expression (usually an atom) which occurs in the –possibly empty– one hole context $\mathcal{C}[]$ whereas $\mathcal{C}[A/A']$ means the replacement of A by A' in context $\mathcal{C}[]$, and $mgu(E)$ is the *most general unifier* of an equation set E. The pair $\langle \mathcal{Q}; \sigma \rangle$ composed by a goal and a substitution is called a *state*. So, given a program $\mathcal{P}$, an *admissible computation* is formalized as a state transition system, whose transition relation $\overset{\text{AS}}{\rightsquigarrow}$ is the smallest relation satisfying the following *admissible rules*:

1) $\langle \mathcal{Q}[A]; \sigma \rangle \overset{\text{AS}}{\rightsquigarrow} \langle (\mathcal{Q}[A/v \&_i \mathcal{B}])\theta; \sigma\theta \rangle$ if A is the selected atom in goal $\mathcal{Q}$, $[A' \leftarrow_i \mathcal{B}\ with\ v] \in \mathcal{P}$, where $\mathcal{B}$ is not empty, and $\theta = mgu(\{A' = A\})$.

2) $\langle \mathcal{Q}[A]; \sigma \rangle \overset{\text{AS}}{\rightsquigarrow} \langle (\mathcal{Q}[A/v])\theta; \sigma\theta \rangle$ if $[A'\ with\ v] \in \mathcal{P}$ and $\theta = mgu(\{A' = A\})$.

The following derivation illustrates our definition (note that the exact program rule used -after being renamed- in the corresponding step is annotated as a super–index of the $\overset{\text{AS}}{\rightsquigarrow}$ symbol, whereas exploited atoms appear underlined and id represents the empty substitution):

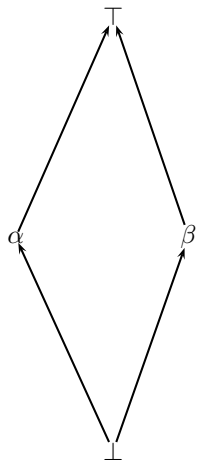
$$\begin{array}{ll} \langle \underline{p(X)}; id \rangle & \overset{\text{AS}}{\rightsquigarrow}{}^{\mathcal{R}_1} \\ \langle 0.8 \ \&_{\mathtt{P}}\ (\underline{q(X_1, Y_1)} \ \&_{\mathtt{G}}\ r(Y_1)); \{X/X_1\} \rangle & \overset{\text{AS}}{\rightsquigarrow}{}^{\mathcal{R}_2} \\ \langle 0.8 \ \&_{\mathtt{P}}\ (0.9 \ \&_{\mathtt{G}}\ \underline{r(Y_2)}); \{X/a, X_1/a, Y_1/Y_2\} \rangle & \overset{\text{AS}}{\rightsquigarrow}{}^{\mathcal{R}_3} \\ \langle 0.8 \ \&_{\mathtt{P}}\ (0.9 \ \&_{\mathtt{G}}\ \overline{0.7}); \{X/a, X_1/a, Y_1/b, Y_2/b\} \rangle & \end{array}$$

The final formula without atoms can be directly interpreted in lattice $\mathcal{L}$ to obtain the desired *fuzzy computed answer* (or *f.c.a.*, in brief), where the substitution only contains bindings associated to variables of the initial goal. So, since $0.8 \ \&_{\mathtt{P}}\ (0.9 \ \&_{\mathtt{G}}\ 0.7) = 0.8 * \min(0.9, 0.7) = 0.56$, in our case the fuzzy computed answer is $\langle 0.56, \{X/a\} \rangle$ indicating that goal $p(X)$ is true at 56 % when X is a.

B. $\mathcal{FLOPER}$

As detailed in [28], [26], our parser has been implemented by using the classical DCG's (*Definite Clause Grammars*) resource of the PROLOG language, since it is a convenient notation for expressing grammar rules. Once the application is loaded inside a PROLOG interpreter (such as Sicstus or SWI), it shows a menu which includes options for loading/compiling, parsing, listing and saving fuzzy programs, as well as for executing/debugging goals and managing multi-adjoint lattices.

All these actions are based in the compilation of the fuzzy code into standard PROLOG code. The key point is to extend each atom with an extra argument, called *truth variable* of the form "`_TV`$_i$", which is intended to contain the truth degree obtained after the subsequent evaluation of the atom. For instance, the first rule in our target program is translated into




```
member(bottom).   member(alpha).
member(beta).     member(top).

bot(bottom).      top(top).

leq(bottom,X).     leq(alpha,alpha).
leq(alpha,top).   leq(beta,beta).
leq(beta,top).    leq(X,top).

and_godel(X,Y,Z) :- pri_inf(X,Y,Z).

pri_inf(bottom,X,bottom) :- !.
pri_inf(alpha,X,alpha) :- leq(alpha,X),!.
pri_inf(beta,X,beta) :- leq(beta,X),!.
pri_inf(top,X,X) :- !.
pri_inf(X,Y,bottom).
```

Figure 1. A finite, partially ordered multi-adjoint lattice modeled in PROLOG

$$\texttt{sse}(c,d) \qquad\qquad \texttt{with}\ \ \Re(c,d)$$
$$\texttt{sse}(c(x_1,..,x_n),d(y_1,..,y_n)) \leftarrow_G \texttt{sse}(x_1,y_1)\ \&_G \ldots \&_G\ \texttt{sse}(x_n,y_n) \qquad \texttt{with}\ \ \Re(c,d)$$

Figure 2. MALP Rules defining "Similarity-based Strict Equality"

the clause:

```
p(X,_TV0) :- q(X,Y,_TV1), r(Y,_TV2),
             and_godel(_TV1,_TV2,_TV3),
             and_prod(0.8,_TV3, _TV0).
```

Moreover, the remaining rules in our fuzzy program, becomes the pure PROLOG facts "`q(a,Y,0.9)`" and "`r(b,0.7)`", whereas the corresponding lattice is expressed by these clauses (the meaning of the mandatory predicates `member`, `top`, `bot` and `leq` is obvious):

```
member(X) :- number(X),0=<X,X=<1.
bot(0).                    top(1).
leq(X,Y) :- X=<Y.
and_godel(X,Y,Z):- pri_min(X,Y,Z).
pri_min(X,Y,Z) :- (X=<Y,Z=X;X>Y,Z=Y).
and_prod(X,Y,Z) :- pri_prod(X,Y,Z).
pri_prod(X,Y,Z) :- Z is X * Y
```

Finally, a fuzzy goal like "`p(X)`", is obviously translated into the pure PROLOG goal: "`p(X, Truth_degree)`" (note that the last truth degree variable is not anonymous now) for which, after choosing option "`run`", the PROLOG interpreter returns the desired fuzzy computed answer $[\texttt{Truth\_degree} = 0.56, \texttt{X} = \texttt{a}]$. Note that all internal computations (including compiling and executing) are pure PROLOG derivations, whereas inputs (fuzzy programs and goals) and outputs (fuzzy computed answers) have always a fuzzy taste, thus producing the illusion on the final user of being working with a purely fuzzy logic programming tool.

Moreover, it is also possible to select into the menu of $\mathcal{FLOPER}$, options "`tree`" and "`depth`", which are useful for tracing execution trees and fixing the maximum length allowed for their branches (initially 3), respectively. To finish this block, in Figure 1 we show the PROLOG clauses modeling a lattice which will be used afterwards in Section IV. Here, apart for dealing with a partially ordered lattice, we use the conjunction of the *Gödel* logic described in this non numeric case as: $\&_{\texttt{G}}(x,y) \triangleq inf(x,y)$. From `http://dectau.uclm.es/floper/` it is possible can download our last version of the $\mathcal{FLOPER}$ tool, which incorporates a graphical interface as shown in Figures 3 and 4.

C. SSE, MALP *and* $\mathcal{FLOPER}$

Now, we are ready to illustrate and prove the properties of our MALP-based model of SSE which is defined according to [20] in Figure 2, where we assume that both c and d are constants (i.e., constructor symbols with arity 0) in the first rule, or both are functions with the same arity n in the second rule and then, $\Re(c,d)$ represents the similarity degree between such pair of symbols with the same arity. In order to illustrate our technique, assume that we plan to compare data terms built with constants "mary" and "maria", which have a similarity degree of 80% and function symbols (with arity one) "brother" and "sibling" which are similar at 90%. According to our

MALP-based definition of SSE we generate a set of MALP rules using the "min" operator (based on *Gödel logic*, as usual in LIKELOG and BOUSI~PROLOG) to propagate similarity degrees. Instead, in the following MALP program loaded into $\mathcal{FLOPER}$ we have used a version inspired on "product logic' (in the following section we describe an application which allows to select the desired conjunction operator or t-norm for composing similarity degrees):

```
sse(maria,maria)                    with 1.
sse(mary,mary)                      with 1.
sse(mary,maria)                     with 0.8.
sse(maria,mary)                     with 0.8.
sse(sibling(X),sibling(Y)) <prod
                        sse(X,Y) with 1.
sse(brother(X),brother(Y)) <prod
                        sse(X,Y) with 1.
sse(sibling(X),brother(Y)) <prod
                        sse(X,Y) with 0.9.
sse(brother(X),sibling(Y)) <prod
                        sse(X,Y) with 0.9.
```

Now, for a goal like "`sse(brother(mary), sibling(maria))`", our technique tests that both parameters are similar terms (with degree $0.9 * 0.8 = 0.72$) in the same way than LIKELOG and BOUSI~PROLOG. Anyway, these last languages only would report just one solution for goals "`sse(brother(mary),X)`" and "`sse(X,Y)`" (the answers computed by LIKELOG and BOUSI~PROLOG for those queries would include the bindings "`{X/ brother(mary)}`" and "`{ X/ Y }`", respectively, both ones with similarity degree 1), whereas our system is able to provide the corresponding four answer for the first query shown in Figure 3, as well as infinite solutions for the second goal (some of them displayed in Figure 4), including the following ones:

```
[Truth_degree=1,X=mary,Y=mary]
[Truth_degree=0.8,X=mary,Y=maria]
[Truth_degree=0.9, X=brother(maria),
                   Y=sibling(maria)]
[Truth_degree=0.72,X=brother(mary),
                   Y=sibling(maria)]
```

In order to formally prove the properties we have just illustrated, it is mandatory to introduce the following auxiliary definition:

Definition 3.1 (Similar terms): Let t and t' be two ground terms, $\Re$ a similarity relation and $\mathcal{L} = \langle L, \preceq, \leftarrow, \& \rangle$ a multi-adjoint lattice. We say that t and t' are similar terms according $\Re$ and $\&$ with similarity degree $s \in L$, if the evaluation of function $\Phi(t, t')$ returns $s \neq \bot$, where function Φ is recursively defined as follows:

$$\Phi(t,t') = \begin{cases} \Re(t,t'), & \text{if } t \text{ and } t' \text{ are constants} \\ \Re(c,c')\&\Phi(t_1,t'_1)\& \ldots \&\Phi(t_n,t'_n) & \text{if } t = c(t_1,\ldots,t_n) \text{ and } t' = c'(t'_1,\ldots,t'_n) \end{cases}$$

The following result reveals the ability of our technique for testing similar terms.

Theorem 3.2: Let t and t' be two ground terms, $\mathcal{L} = \langle L, \preceq, \leftarrow, \& \rangle$ a multi-adjoint lattice, $\Re$ a similarity relation and $\mathcal{P}^{\Re}_{sse}$ the set of MALP rules defining predicate sse w.r.t. $\Re$. Then, t and t' are similar terms according $\Re$ and $\&$ with similarity degree $s \in L$, iff $\langle s, id \rangle$ is a fuzzy computed answer for goal $sse(t, t')$ in $\mathcal{P}^{\Re}_{sse}$.

Proof: We prove this claim by structural induction on the shape of t and t'.

• Base case. We assume here that t and t' are similar constants, and then, $\Re(t, t') = s \neq \bot$ whereas rule $[\mathcal{R} : sse(t, t')\ with\ s]$ belongs to $\mathcal{P}^{\Re}_{sse}$. Then, it is easy to see that $\Phi(t, t') = \Re(t, t') = s$ as well as to perform with rule $\mathcal{R}$ the following admissible step $\langle sse(t, t'), id \rangle \overset{\text{AS}}{\rightsquigarrow}^{\mathcal{R}} \langle s, id \rangle$.

• Induction step. Now we have that $t = c(t_1, \ldots, t_n)$ and $t' = c'(t'_1, \ldots, t'_n)$. Assuming that $\Re(c, c') = s_0 \neq \bot$ and $\Phi(t_i, t'_i) = s_i \neq \bot$, $1 \leq i \leq n$, then $\Phi(t, t') = s_0 \& s_1 \& \ldots \& s_n \neq \bot$. Moreover, since our technique generates the rule (which belongs to $\mathcal{P}^{\Re}_{sse}$):

$$\mathcal{R} : sse(c(x_1,\ldots,x_n), c(x'_1,\ldots,x'_n)) \leftarrow sse(x_1,x'_1)\& \ldots \& sse(x_n,x'_n)\ \ with\ s_0$$

and by the inductive hypothesis we can assume that $\langle s_i, id \rangle$ is a fuzzy computed answer for goal $sse(t_i, t'_i)$, $1 \leq i \leq n$, then it is possible to generate the following sequence of admissible steps (for readability reasons, we omit in the substitution component of each state the bindings associated to variables not belonging to the initial goal):

$$\begin{array}{ll} \langle sse(c(t_1,\ldots,t_n), c'(t'_1,\ldots,t'_n)); id \rangle & \overset{\text{AS}}{\rightsquigarrow}^{\mathcal{R}} \\ \langle s_0 \ \&\ sse(t_1,t'_1)\ \& \ldots \&\ sse(t_n,t'_n); id \rangle & \overset{\text{AS}}{\rightsquigarrow} \ldots \overset{\text{AS}}{\rightsquigarrow} \\ \langle s_0 \ \&\ s_1\ \& \ldots \&\ sse(t_n,t'_n); id \rangle & \overset{\text{AS}}{\rightsquigarrow} \ldots \overset{\text{AS}}{\rightsquigarrow} \\ \langle s_0 \ \&\ s_1\ \& \ldots \&\ s_n; id \rangle & \end{array}$$

which concludes our proof. ■

Next theorem reinforces the previous one by establishing the capability of our technique for generating (not only for testing) pairs of similar ground terms.

Theorem 3.3: Let t and t' be two ground terms, x a variable, $\mathcal{L} = \langle L, \preceq, \leftarrow, \& \rangle$ a multi-adjoint lattice, $\Re$ a similarity relation and $\mathcal{P}^{\Re}_{sse}$ the set of MALP rules defining predicate sse w.r.t. $\Re$. Then, t and t' are similar terms according $\Re$ and $\&$ with similarity degree $s \in L$, iff $\langle s, \{x/t'\} \rangle$ is a fuzzy computed answer for goal $sse(t, x)$ in $\mathcal{P}^{\Re}_{sse}$.

Proof: Our proof is based again on structural induction on the shape of t, and it clearly resembles the one built for Theorem 3.2 but pointing out now the effects on the variables of the original goal.

• Base case. We assume here that t and t' are similar constants, and then, $\Re(t, t') = s \neq \bot$ whereas rule $[\mathcal{R} : sse(t, t')\ with\ s]$ belongs to $\mathcal{P}^{\Re}_{sse}$. Then, obviously $\phi(t, t') = \Re(t, t') = s$ whereas it is possible too to perform with rule $\mathcal{R}$ the following admissible step $\langle sse(t, x), id \rangle \overset{\text{AS}}{\rightsquigarrow}^{\mathcal{R}} \langle s, \{x/t'\} \rangle$.

• Induction step. Now we have that $t = c(t_1, \ldots, t_n)$ and $t' = c'(t'_1, \ldots, t'_n)$. Assuming that $\Re(c, c') = s_0 \neq \bot$

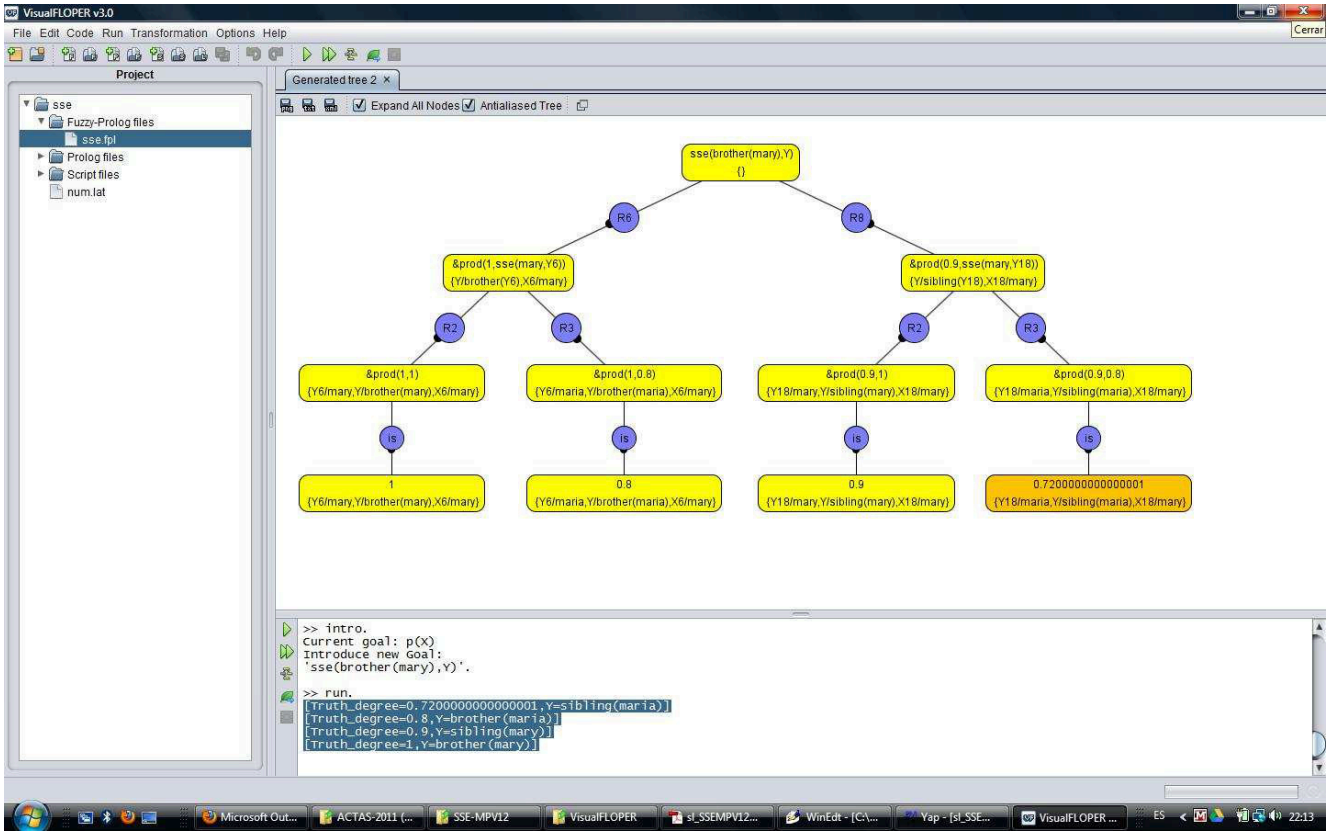


Figure 3. Screen-shot of a work session with $\mathcal{FLOPER}$

and $\Phi(t_i, t'_i) = s_i \neq \bot$, $1 \leq i \leq n$, then $\Phi(t, t') = s_0 \& s_1 \& \ldots \& s_n \neq \bot$. Moreover, since our technique generates the rule (which belongs to $\mathcal{P}^{\Re}_{sse}$):

$$\mathcal{R} : sse(c(x_1, \ldots, x_n), c(x'_1, \ldots, x'_n)) \leftarrow sse(x_1, x'_1) \& \ldots \& sse(x_n, x'_n) \quad with\ s_0$$

and by the inductive hypothesis we can assume that $\langle s_i, \{x'_i / t'_i\}\rangle$ is a fuzzy computed answer for goal $sse(t_i, x'_i)$, $1 \leq i \leq n$, then it is possible to generate the derivation shown in Figure 5 (for simplifying, we only include in the substitution component of each state those bindings which are relevant for our purposes) which concludes our proof. ■

The repeated application of the previous theorem implies the following result which, in essence, confirms the power of our method for producing all pairs of similar data terms.

Corollary 3.1: Let t and t' be two ground terms, x and x' two variables, $\mathcal{L} = \langle L, \preceq, \leftarrow, \& \rangle$ a multi-adjoint lattice, $\Re$ a similarity relation and $\mathcal{P}^{\Re}_{sse}$ the set of MALP rules defining predicate sse w.r.t. $\Re$. Then, t and t' are similar terms according $\Re$ and $\&$ with similarity degree $s \in L$, iff $\langle s, \{x/t, x'/t'\}\rangle$ is a fuzzy computed answer for goal $sse(x, x')$ in $\mathcal{P}^{\Re}_{sse}$.

IV. Implementation Issues

We start this section by firstly describing in sub-section IV-A how users can introduce into the new SSE tool (written in Prolog and freely accessible from `http://dectau.uclm.es/sse/`) a small set of similarity equations with a natural and very easy syntax. After that, the tool performs the reflexive-symmetric-transitive closure of that specification in order to obtain a similarity relation $\Re$ which is translated into a Prolog program, as explained in sub-section IV-B. Finally, the application uses $\Re$ to generate a MALP program defining SSE, as described in sub-section IV-C.

A. Syntax for Similarity Files

To specify a similarity relation, it is mandatory to load a file with extension '.sim' into the tool. This file is intended to contain a set of similarity equations, where each equation is expressed by separating two literals (the ones to be considered similar) with the '$\sim$' symbol, and adding a *truth value* to the similarity (usually, a number of the real interval [0,1], but our tool also admits an element from any multi-adjoint lattice, in contrast with other fuzzy languages such as Bousi~Prolog or Likelog) after the '=' symbol. So, for instance, `brother ~ sibling = 0.9`, is a valid similarity equation. Our syntax also allows to specify the arity of each symbol after a suffixed slash (i.e. '`brother/1`'). Thus, it is possible to discriminate between functors with the same name but different arities. When the user does not include arity information, it is simple assumed to be zero. To relate literals without arity specification (i.e., with no arity

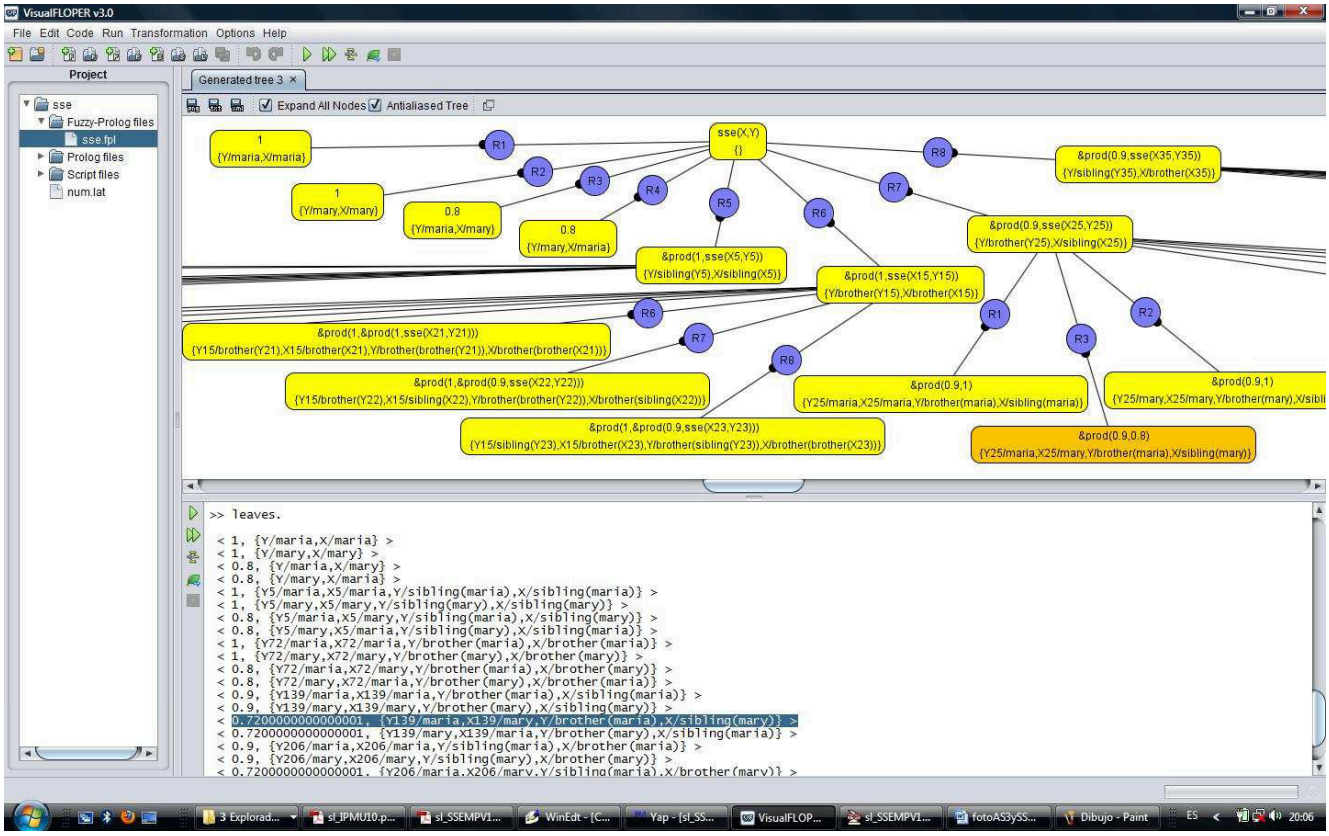


Figure 4. $\mathcal{FLOPER}$ showing three levels of an infinite evaluation tree

$$\langle sse(c(t_1,\ldots,t_n),x)); id\rangle \overset{\text{AS}}{\rightsquigarrow}^{\mathcal{R}}$$
$$\langle s_0 \;\&\; sse(t_1,x'_1) \;\&\ldots\&\; sse(t_n,x'_n); \{x/c'(x'_1,\ldots,x'_n), x_1/t_1,\ldots,x_n/t_n\}\rangle \overset{\text{AS}}{\rightsquigarrow}$$
$$\ldots \overset{\text{AS}}{\rightsquigarrow} \ldots \overset{\text{AS}}{\rightsquigarrow}$$
$$\langle s_0 \;\&\; s_1 \;\&\ldots\&\; sse(t_n,x'_n); \{x/c'(t'_1,\ldots,x'_n), x_1/t_1,\ldots,x_n/t_n, x'_1/t'_1\}\rangle \overset{\text{AS}}{\rightsquigarrow}$$
$$\ldots \overset{\text{AS}}{\rightsquigarrow} \ldots \overset{\text{AS}}{\rightsquigarrow}$$
$$\langle s_0 \;\&\; s_1 \;\&\ldots\&\; s_n); \{x/c'(t'_1,\ldots,t'_n), x_1/t_1,\ldots,x_n/t_n, x'_1/t'_1,\ldots,x'_n/t'_n\}\rangle$$

Figure 5. Proof of Theorem 3.3

discrimination), we need to write an underscore after the slash symbol (i.e., '`brother/_`').

Example 4.1: Consider now the following specification of a similarity relation:

```
mary ~ maria = 0.8.
sibling/1 ~ brother/1 = 0.9.
```

It is not necessary to add all similarity equations (for instance, the reflexive equation relating `mary` with `mary`), since the tool is able to "complete" the relation by performing the reflexive, symmetric and transitive closure of the given set of equations, as we will see in sub-section IV-B.

Note again that since our tool can work with different multi-adjoint lattices, similarity equations can be also described beyond the real interval $[0,1]$: the only required condition is that the similarity degrees of equations have to be members of the multi-adjoint lattice associated to the program or, in other words, with the lattice currently loaded into the system (see Figure IV-A).

B. Closure and Translation to PROLOG

Each similarity equation from the ".sim" file is translated into a PROLOG clause holding all its information. So, a similarity equation $\texttt{A/}n_A \sim \texttt{B/}n_B = \texttt{V}$ is coded as fact `r((A,`n_A`),(B,`n_B`),V)`, thus including the arity of each literal. The previous Example 4.1 (based on real numbers in the unit interval) should then be translated into:

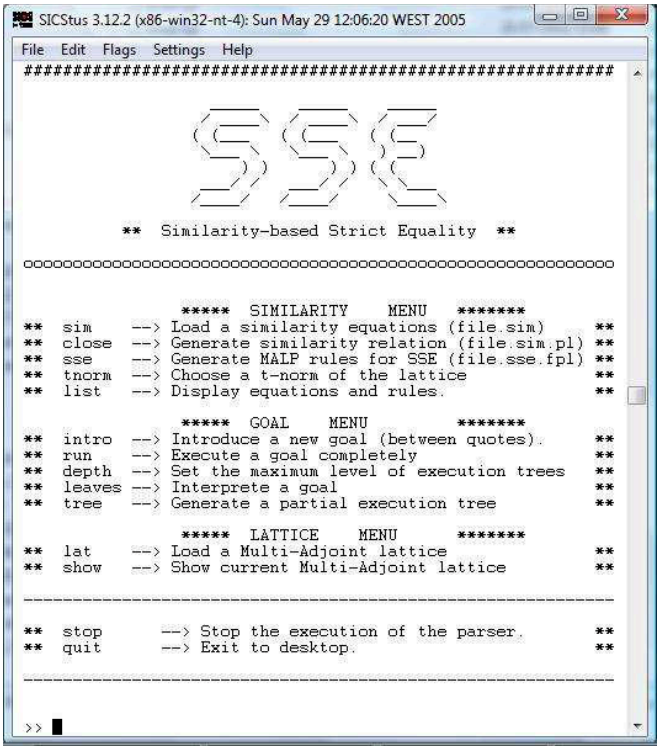


Figure 6. Main menu of the SSE application implemented with PROLOG

Algorithm 1

Require: An adjacency matrix $M = [m_{ij}]$, representing a fuzzy binary relation R on a set A, whose elements preserve transitivity and with all the elements of the superior triangular matrix set to $\bot$.

Ensure: The adjacency matrix $M^{\equiv}$ corresponding to the reflexive, symmetric, transitive closure of R.

for all $\langle i, i\rangle$ in M **do** {Build the reflexive closure}

2: $m_{ii} := \top$;

end for

4: **for all** $\langle i, j\rangle$ in M, such that $m_{ij} \neq \bot$ **do** {Build the symmetric closure}

$m_{ji} := m_{ij}$;

6: **end for**

for all column k and entry $\langle i, j\rangle$ in M **do** {Build the transitive closure}

8: $m_{ij} := m_{ij} \vee (m_{ik} \wedge m_{kj})$; where "$\vee$" and "$\wedge$" are, respectively, the supremum and infimum operators;

end for

10: $M^{\equiv} := M$

```
r((mary, 0), (maria, 0), 0.8).
r((sibling, 1), (brother, 1), 0.9).
```

All these facts are saved in their own PROLOG module "sim", in order to avoid collision of names. Also, the system saves the translated code into a file with the same name but extension "sim.pl".

Once this process has finished, the tool completes the intended similarity relation by performing the reflexive-symmetric-transitive closure according to Algorithm 1 (that we have just implemented in PROLOG) which is inspired by the one described in [29], [12], but generalizing it in order to deal with (multi-adjoint) lattices beyond the $[0, 1]$ case[2]:

As a result of performing this algorithm, the intended similarity relation is completed, and then, all similarity equations are successfully stored into module "sim" as PROLOG facts. The next step is to write these similarity equations into a file with the same name of that of the specification but extension "sim.pl", thus pointing out that the new file includes the same information, but using PROLOG syntax.

For instance, the closure of the similarity specification from Example 4.1 should return the relation of the following table, where a cell $\langle i, j\rangle$ gives the corresponding similarity degree between two symbols.

	maria	mary	brother	sibling
maria	1	0.8	0	0
mary	0.8	1	0	0
brother	0	0	1	0.9
sibling	0	0	0.9	1

This table is modeled by means of the following set of PROLOG facts resulting from the translation process previously described:

```
sim((maria,0),(maria,0),1).
sim((maria,0),(mary,0),0.8).
sim((mary,0),(maria,0),0.8).
sim((mary,0),(mary,0),1).
sim((brother,1),(brother,1),1).
sim((brother,1),(sibling,1),0.9).
sim((sibling,1),(brother,1),0.9).
sim((sibling,1),(sibling,1),1).
```

Example 4.2: For the following two similarity equations using degrees of the partially ordered lattice in Figure 1 (see again sub-section III-B), we show its corresponding table and associated PROLOG facts below (note that the 'top' element is the truth degree for all reflexive equations):

$$\mathrm{c} \sim \mathrm{d} = \mathrm{alpha}.$$
$$\mathrm{f}/2 \sim \mathrm{g}/2 = \mathrm{beta}.$$

	c	d	f	g
c	top	alpha	bot	bot
d	alpha	top	bot	bot
f	bot	bot	top	beta
g	bot	bot	beta	top

```
sim((c,0),(c,0),top).
sim((c,0),(d,0),alpha).
sim((d,0),(c,0),alpha).
sim((d,0),(d,0),top).
```

[2]Note that the algorithm can work with any particular multi-adjoint lattice: since any complete lattice (with supremum, infimum and a concrete ordering relation) is valid, then any multi-adjoint lattice is valid too.

Algorithm 2

Require: A set of similarity equations $S = \{S_i, i \in \{0,\ldots,N\}\}$ of the form $S_i = \{A/n_A \sim B/n_B = V\}$, where A and B are function symbols (possibly constants), n_A and n_B are their respective arities and V is the corresponding similarity degree.

Ensure: A set of MALP rules $R = \{R_i, i \in \{0,\ldots,N\}\}$.

for all $S_i = \{A/n_A \sim B/n_B = V\}$ in S **do**
  $body := \text{“}with\text{''} + V$;
  for all $j \in \{n_A,\ldots,1\}$ **do**
    $body := \text{“}, sse(X_j, Y_j)\text{''} + body$;
  end for
  if $n_A > 0$ **then**
    $body := \text{“} <- sse(X_1, Y_1)\text{''} + body$;
  end if
  $R_i := \text{“}sse(A(X_1,\ldots,X_{n_A}), B(Y_1,\ldots,Y_{n_B}))\text{''}$;
  $R_i := R_i + body$;
end for

```
sim((f,2),(f,2),top).
sim((f,2),(g,2),beta).
sim((g,2),(f,2),beta).
sim((g,2),(g,2),top).
```

C. *From similarities to* MALP *rules modeling* SSE

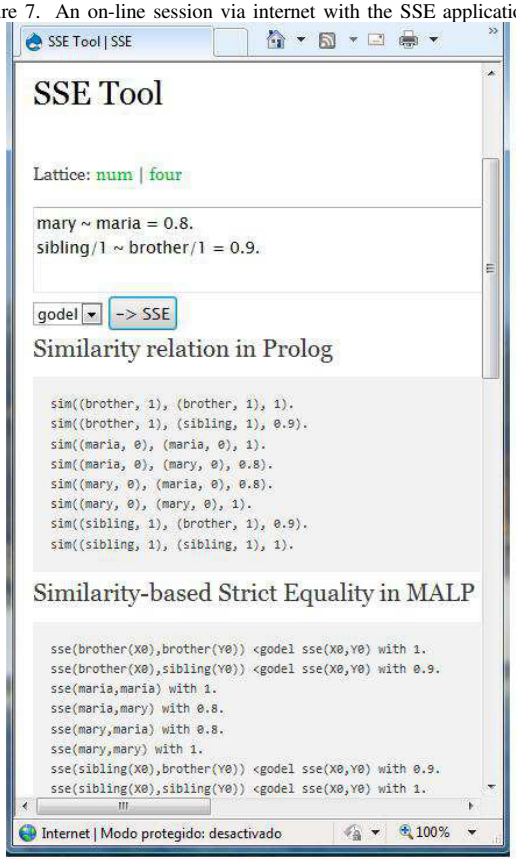
The last step consists on translating the similarity relation from its PROLOG syntax to the MALP syntax. Algorithm 2 performs such process, where the input is the set of PROLOG facts obtained after performing the closure, and the output is the intended MALP program:

Since $R = \{R_i, i \in \{0,\ldots,N\}\}$ is a set of MALP rules, it is also a valid fuzzy program, so it is located in a file with the same name of the original specification, and extension “sse.fpl”, thus implementing the notion of “Similarity-based Strict Equality SSE” as a MALP program. The resulting file can be naturally loaded into the $\mathcal{FLOPER}$ tool in order to run and debug goals, depicting evaluation trees, etc.

In order to illustrate this proccess, consider again the specification given in Example 4.1. Once we have the closure of the specification given in Section IV-A (expressed by means of PROLOG facts), the final MALP program has the following form:

```
sse(maria,maria)                    with 1.
sse(maria,mary)                     with 0.8.
sse(mary,maria)                     with 0.8.
sse(mary,mary)                      with 1.
sse(brother(X0),brother(Y0)) <- sse(X0,Y0)
                                    with 1.
sse(brother(X0),sibling(Y0)) <- sse(X0,Y0)
                                    with 0.9.
sse(sibling(X0),brother(Y0)) <- sse(X0,Y0)
                                    with 0.9.
sse(sibling(X0),sibling(Y0)) <- sse(X0,Y0)
                                    with 1.
```

Figure 7. An on-line session via internet with the SSE application



Moreover, regarding the similarity relation recasted from Example 4.2, we obtain the following set of MALP rules:

```
sse(c,c)                          with top.
sse(c,d)                          with alpha.
sse(d,c)                          with alpha.
sse(d,d)                          with top.
sse(f(X0,X1),f(Y0,Y1)) <- sse(X0,Y0) &
                          sse(X1,Y1) with top.
sse(f(X0,X1),g(Y0,Y1)) <- sse(X0,Y0) &
                          sse(X1,Y1) with beta.
sse(g(X0,X1),f(Y0,Y1)) <- sse(X0,Y0) &
                          sse(X1,Y1) with beta.
sse(g(X0,X1),g(Y0,Y1)) <- sse(X0,Y0) &
                          sse(X1,Y1) with top.
```

In addition to our desktop tool, we have developed too a comfortable on-line version of the application (so it is not necessary to download any file, but only work through the internet) which is located at the web page `dectau.uclm.es/sse`. We provide a link to download the PROLOG-based implementation of the tool but also, and more importantly, this URL enables the possibility of performing on-line work sessions, as illustrated in the screen-shot displayed in Figure 7.

V. CONCLUSIONS AND FUTURE WORK

In this paper we have recasted from [20] a static preprocess for improving the expressive power of a fuzzy declarative language in order to easily cope with similarity relations.

More exactly, we have adapted to the MALP framework our preliminary notion of SSE presented in [17], thus dealing with similarity relations by means of a simple but powerful method (somehow inspired by the -non fuzzy- functional paradigm) which surpass in some cases the effects obtained in other fuzzy languages which are not based on the simpler syntactic unification method of PROLOG. The main goal of this paper focused on proving some important formal properties of our technique for which we have shown some experimental results obtained by using our $\mathcal{FLOPER}$ platform as well as a preliminary PROLOG-based implementation of the technique (please, visit `http://dectau.uclm.es/sse/` for testing it on-line), which is nowadays being introduced inside the core of our system.

REFERENCES

[1] J. Lloyd, *Foundations of Logic Programming*. Springer-Verlag, Berlin, 1987, second edition.

[2] H. Nguyen and E. Walker, *A First Course in Fuzzy Logic*. Chapman & Hall/CRC, Boca Ratón, Florida, 2000.

[3] I. Bratko, *Prolog Programming for Artificial Intelligence*. Addison Wesley, 2000.

[4] R. Lee, "Fuzzy Logic and the Resolution Principle," *Journal of the ACM*, vol. 19, no. 1, pp. 119–129, 1972. [Online]. Available: http://doi.acm.org/10.1145/321679.321688

[5] M. Ishizuka and N. Kanai, "Prolog-ELF Incorporating Fuzzy Logic," in *Proceedings of the 9th Int. Joint Conference on Artificial Intelligence, IJCAI'85*, A. K. Joshi, Ed. Morgan Kaufmann, 1985, pp. 701–703. [Online]. Available: http://dl.acm.org/citation.cfm?id=1623611.1623612

[6] J. F. Baldwin, T. P. Martin, and B. W. Pilsworth, *Fril- Fuzzy and Evidential Reasoning in Artificial Intelligence*. John Wiley & Sons, Inc., 1995.

[7] D. Li and D. Liu, *A fuzzy Prolog database system*. John Wiley & Sons, Inc., 1990.

[8] J. Medina, M. Ojeda-Aciego, and P. Vojtáš, "Similarity-based Unification: a multi-adjoint approach," *Fuzzy Sets and Systems*, vol. 146, pp. 43–62, 2004. [Online]. Available: http://dblp.uni-trier.de/db/journals/fss/fss146.html#MedinaOV04

[9] M. Sessa, "Approximate reasoning by similarity-based SLD resolution," *Fuzzy Sets and Systems*, vol. 275, pp. 389–426, 2002. [Online]. Available: http://dx.doi.org/10.1016/S0304-3975(01)00188-8

[10] F. Arcelli and F. Formato, "Likelog: A logic programming language for flexible data retrieval," in *Proc. of the 1999 ACM Symposium on Applied Computing (SAC'99), February 28 - March 2, 1999, San Antonio, USA*. ACM, Artificial Intelligence and Computational Logic, 1999, pp. 260–267. [Online]. Available: http://doi.acm.org/10.1145/298151.298348

[11] R. Caballero, M. Rodríguez-Artalejo, and C. A. Romero-Díaz, "Similarity-based reasoning in qualified logic programming," in *Proceedings of the 10th Int. ACM SIGPLAN conference on Principles and practice of declarative programming*, ser. PPDP'08. New York, USA: ACM, 2008, pp. 185–194. [Online]. Available: http://doi.acm.org/10.1145/1389449.1389472

[12] P. Julián, C. Rubio, and J. Gallardo, "Bousi∼prolog: a prolog extension language for flexible query answering," *Electronic Notes in Theoretical Computer Science*, vol. 248, pp. 131–147, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.entcs.2009.07.064

[13] C. Rubio-Manzano and P. Julián-Iranzo, "A fuzzy linguistic prolog and its applications," *Journal of Intelligent and Fuzzy Systems*, vol. 26, no. 3, pp. 1503–1516, 2014. [Online]. Available: http://dx.doi.org/10.3233/IFS-130834

[14] M. Bröcheler, L. Mihalkova, and L. Getoor, "Probabilistic similarity logic," *Computing Research Repository*, vol. abs/1203.3469, 2012. [Online]. Available: http://arxiv.org/abs/1203.3469

[15] A. Kimmig, B. Demoen, L. D. Raedt, V. S. Costa, and R. Rocha, "On the implementation of the probabilistic logic programming language problog," *TPLP*, vol. 11, no. 2-3, pp. 235–262, 2011. [Online]. Available: http://dx.doi.org/10.1017/S1471068410000566

[16] G. Moreno and V. Pascual, "A hybrid programming scheme combining fuzzy-logic and functional-logic resources," *Fuzzy Sets and Systems*, vol. 160, pp. 1402–1419, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.fss.2008.11.028

[17] G. Moreno, "Similarity-based equality with lazy evaluation," in *Proc. of the 13th Int. Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems, IPMU'10, June 28-July 2, Dortmund, Germany*, E. Hullermeier, R. Kruse, and F. Hoffmann, Eds. Springer CCIS 80 (Part I), 2010, pp. 108–117.

[18] C. V. Hall, K. Hammond, W. Partain, S. L. P. Jones, and P. Wadler, "The glasgow haskell compiler: A retrospective," in *Functional Programming*, ser. Workshops in Computing, J. Launchbury and P. M. Sansom, Eds. Springer, 1992, pp. 62–71. [Online]. Available: http://dl.acm.org/citation.cfm?id=647557.729914

[19] M. Hanus (ed.), "Curry: An Integrated Functional Logic Language," Available at `http://www.informatik.uni-kiel.de/~mh/curry/`, 2003.

[20] G. Moreno, J. Penabad, and C. Vázquez, "SSE: Similarity-based strict equality for multi-adjoint logic programs," in *Proceedings 12th Int. Conference on Mathematical Methods in Science and Engineering, CMMSE'12. La Manga (Murcia), Spain, July 2-5*, J. Vigo-Aguiar, Ed., vol. III. ISBN: 978-84-615-5392-1, 2012, pp. 876–887.

[21] L. A. Zadeh, "Similarity relations and fuzzy orderings," *Information Sciences*, vol. 3, pp. 177–200, 1971. [Online]. Available: http://dx.doi.org/10.1016/S0020-0255(71)80005-1

[22] P. Morcillo, G. Moreno, J. Penabad, and C. Vázquez, "Dedekind-MacNeille completion and cartesian product of multi-adjoint lattices," *Int. Journal of Computer Mathematics*, vol. 89, no. 13-14, pp. 1742–1752, 2012. [Online]. Available: http://dx.doi.org/10.1080/00207160.2012.689826

[23] P. Morcillo, G. Moreno, J. Penabad, and C. Vázquez, "Declarative Traces into Fuzzy Computed Answers," in *Proc. of 5th Int. Symposium on Rules: Research Based, Industry Focused, RuleML'11. Barcelona, Spain, July 19–21*, N. Bassiliades, G. Governatori, and A. Paschke, Eds. Springer Verlag, LNCS 6826, 2011, pp. 170–185. [Online]. Available: http://dl.acm.org/citation.cfm?id=2032787.2032806

[24] J. Almendros-Jiménez, A. Luna, and G. Moreno, "A Flexible XPath-based Query Language Implemented with Fuzzy Logic Programming," in *Proc. of 5th Int. Symposium on Rules: Research Based, Industry Focused, RuleML'11. Barcelona, Spain, July 19–21*, N. Bassiliades, G. Governatori, and A. Pasckhe, Eds. Springer Verlag, LNCS 6826, 2011, pp. 186–193. [Online]. Available: http://dl.acm.org/citation.cfm?id=2032787.2032807

[25] G. Moreno and C. Vůzquez, "Fuzzy logic programming in action with floper," *Journal of Software Engineering and Applications*, vol. 7, pp. 237–298, 2014. [Online]. Available: http://dx.doi.org/10.4236/jsea.2014.74028

[26] P. Morcillo, G. Moreno, J. Penabad, and C. Vázquez, "A Practical Management of Fuzzy Truth Degrees using FLOPER," in *Proc. of 4nd Int. Symposium on Rule Interchange and Applications, RuleML'10*, M. D. et al., Ed. Springer Verlag, LNCS 6403, 2010, pp. 20–34. [Online]. Available: http://dl.acm.org/citation.cfm?id=1929574.1929580

[27] ——, "Fuzzy Computed Answers Collecting Proof Information," in *Advances in Computational Intelligence - Proc of the 11th Int. Work-Conference on Artificial Neural Networks, IWANN'11*, J. C. et al., Ed. Springer Verlag, LNCS 6692, 2011, pp. 445–452.

[28] P. Morcillo and G. Moreno, "Programming with Fuzzy Logic Rules by using the FLOPER Tool," in *Proc of the 2nd. Rule Representation, Interchange and Reasoning on the Web, Int. Symposium, RuleML'08*, N. B. et al., Ed. Springer Verlag, LNCS 3521, 2008, pp. 119–126. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-88808-6_1 4

[29] P. Julián, "A procedure for the construction of a similarity relation," in *Proc. of 12th Information Processing and Management of Uncertainty, IPMU'08,June 22-27, Málaga, Spain*, M. Ojeda, Ed. Springer CCIS 80 (Part I), 2008, pp. 489–ń496.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 129–134
DOI: 10.15439/2014F488
ACSIS, Vol. 2

Dynamic Weighting

New method of weighting panels with large numbers of weighting parameters

Marcin Pery
Military University of Technology
Faculty of Cybernetics
Warsaw, Poland
marcin@pery.pl

Abstract - The algorithm for dynamic weighing presented in this paper is a method used in research studies based on samples when due to the large number of weighting parameters it is not possible to establish a fixed set of sample weights without non-acceptable dispersion of weights.

Keywords: weighting methods, internet audience research, dynamic weighting, machine learning, multiple classifier systems

I. Introduction

Research studies for traditional media have a solid theoretical basis and well-established research methods [1], [2] based on the construction of samples and estimating the number of subsets of the population with specific characteristics. The first step is creating samples - panels [3], [4]. The second step is using known methods of solving systems of equations with many unknowns, determining weights of every element of sample (panelist) in order to make sample "representative" for the entire population [5], [6]. The panel is considered to have better quality when the dispersion of weights is small [7], [8], [9]. The estimation of the number of individuals in the population with the desired attributes (e.g. reading a press title or watching a TV channel) is the sum of the weights of the panelists having these attributes. Internet research studies have their unique characteristics resulting from the number and the type of data, which do not have researchers of other media (e.g. through site-centric type research studies) [10]. Each user's contact with the internet service leaves a trace in servers sending content to the user. Therefore, in the case of internet research studies, we have a vast number of hard data from the measurement of a large number of websites [11]. In practice, the estimation results for the samples (constructed the same way as for the other media) turned out to be unacceptable due to significant discrepancies between the results estimated by using a panel and hard data obtained from the measurement. If the number of weighting parameters is large it is very difficult to correctly determine the weights for a sample. Systems of equations are then unsolvable, or results are unacceptable because the dispersion of weights is too large.

The dynamic weighting algorithm was created in order to estimate coherent and useful results for the whole population (e.g. for audience research studies) even when it is not possible to determine one fixed set of weights of a sample. The algorithm can be used in machine learning and multiple classifier systems where there is a need to draw conclusions on the basis of samples [12], [13], [14], [15].

The very simple example of standard task for computing weights is as follow:

Let us assume that we have a population P of 100 objects which is a set of individuals with two attributes: $\{"is\ men", "is\ woman"\}$. Each of the individuals has exactly one of these two attributes. Let us assume that there are 40 individuals who have attribute "$is\ men$" and 60 with attribute "$is\ woman$". In a sample we have 4 objects, two of which have the attribute "is men" and the other two have the attribute "is woman". To compute weights of objects in the data sample we have to compute systems of equations with many unknowns: $\begin{cases} w_1 + w_2 \approx 40 \\ w_3 + w_4 \approx 60 \end{cases}$

There are infinitely many solutions to this system of equations. We need to find such a solution where weights of objects in a sample are as similar as it is possible. The simplest solution is: $w_1 = 20, w_2 = 20, w_3 = 30, w_4 = 30$. It means that the first object in the sample represents 20 objects in the population, the third object in the sample represents 30 objects in the population, etc.

Let us assume now that there are more attributes (not only sex but also a place of living, age, etc.) and the number of equations is much bigger. There could be no possibility to compute one set of weights which are similar. In a great number of cases some weights have to be set on zero (or close to zero). It means that the sample has very poor quality and as such is of no use. In such cases there is a need to take another approach to determine weights of objects in the sample - this new approach is presented in this very paper.

II. Model

A. Population

Let us define a population of objects P as follows:

$$P = \{p_1, p_2, \ldots, p_N\} \quad (1)$$

Let us define a set of objects' attributes A as follows:

$$A = \{a_0, a_1, a_2, \ldots, a_M\} \quad (2)$$

978-83-60810-58-3/$25.00

Let us define a function which assigns subsets of attributes from set A to objects from set P :

$$a: P \to 2^A \quad (3)$$

Values of function a are known only for some objects from set P. For most of the objects from set P values of function a are unknown.

Attributes are questions about some features of objects. There are only two possible answers: "yes" or "no". If $a_j \in a(o_i)$ it means that the answer is "yes" and the object o_i has the attribute a_j.

The attribute a_0 is special and it is the question: "Does the object belong to the population?". The attribute a_0 meets the following condition:

$$\forall_{o \in P}\left(a_0 \in a(o)\right) \quad (4)$$

Attributes are organized in the *Attributes tree*.

Let us define a function $parent$ which assigns to attribute a_i direct parent node in the *Attributes tree*:

$$parent: A \to A \quad (5)$$

Let us define a function $children$ which assigns a set of direct children nodes in the *Attributes tree* to attribute a_i :

$$children: A \to 2^A \quad (6)$$

Attributes tree meets the following conditions:

- The attribute a_0 is a root of the *Attributes tree*.
- For every pair of p_i and a_j object p_i has no more than one attribute which is the direct child of a_j:

$$\forall_{\mathrm{p_i \in P}} \forall_{\mathrm{a_j \in A}} \overline{\left(\mathrm{a(p_i) \cap children(a_j)}\right)} \leq 1 \quad (7)$$

If object p_i has attribute a_j then object p_i has attribute which is a direct parent of a_j (it does not apply to $\mathrm{a_0}$ as a root of the *Attributes tree*):

$$\forall_{\mathrm{p_i \in P}} \forall_{\mathrm{a_j \in A \setminus \{a_0\}}} \left(\mathrm{a_j \in a(p_i) \Rightarrow parent(a_j) \in a(p_i)}\right) \quad (8)$$

There is non-empty subset of attributes $A^{universe} \subset A$, $\mathrm{A} \neq \emptyset$ that contains only such attributes that we know how many objects from set P have them. Let us define a function $universe$ which assigns number of objects which have the following attribute to the attribute a_i :

$$univers: A^{universe} \to \langle 0, N \rangle \quad (9)$$

$A^{universe}$ meets the following conditions:

$$a_0 \in A^{universe} \quad (10)$$

$$\mathrm{universe(a_0) = N} \quad (11)$$

Let us define a function p which assigns subsets of objects from set P to attributes from set A:

$$p: A \to 2^P \quad (12)$$

$$\forall_{a_j \in A} \forall_{p_i \in p(a_j)} \left(a_j \in a(p_i)\right) \wedge \forall_{p_i \in P} \forall_{a_j \in a(p_i)} \left(p_i \in p(a_j)\right) \quad (13)$$

B. Sample

There is a non-empty subset $S \subset P$ (called *Sample*) that contains only such objects that we know values of function a for these objects. Let us define such a subset as follows:

$$S = \{s_1, s_2, \ldots, s_n\}, S \neq \emptyset \quad (14)$$

Let us define a function w (called *Weighing function*) which assigns the size of part of population which these objects "represent" (called *Weight*) to objects from set S:

$$\mathrm{w: S \to \langle 1, N \rangle} \quad (15)$$

Weighting function meets the following conditions:

- The sum of *Weights* of every element from set S is N:

$$\sum_{i=1}^{n} w(s_i) = N \quad (16)$$

- The sum of Weights of objects which have some attributes must be equal to values of function $universe$ for this attributes:

$$\forall_{a_j \in A^{universe}} \left(\sum_{s_i \in s(a_j)} w(s_i) = universe\left(a_j\right)\right) \quad (17)$$

Let us define *dispersion* e as a quality measure of *Weighting function* for *Sample*:

$$e = \frac{\sum_{i=1}^{n} \left(w(s_i) - \frac{N}{n}\right)^2}{n} \quad (18)$$

If e is big, the quality of sample data is bad and *Sample* cannot be used as a reliable source of information for the whole population P.

C. Task

Let us define *Question* as a subset $\mathrm{Q \subseteq A}$:

$$Q = \{q_1, q_2, \ldots, q_m\} \quad (19)$$

Having given:

- *Question* Q,
- Attributtes tree,
- function a,

- function p,
- *Sample* S,
- function $universe$,
- *Weighting function* w

find number R which is a size of a subset of objects from set P which have at least one attribute from set Q.

D. The trivial solution

If the model satisfies all the above assumptions, including the assumption (17), the number R is computing as follows:

$$R = \sum_{s_i \in \left(\cup_{j=1}^{m} p(q_j) \cap S\right)} w(s_i) \tag{20}$$

III. PROBLEM

If the number of attributes in set $A^{universe}$ is as big that it is almost impossible to satisfy all the above assumptions, including the assumption (17) with acceptable level of *dispersion* e. The set of equations needed to solve this case has no solution or final *Weights* are so different that it is not possible to draw conclusions on the basis of *Sample*.

To be able to compute any reliable results, the assumption (17) is satisfied only for some subset $\mathrm{A^{universe'}} \subset \mathrm{A^{universe}}$. In the internet research studies number of attributes in set $A^{universe'}$ is even tens of times greater than size of $A^{universe}$. Because of that, there is a need to use a different method of constructing *Weighting function* and computing number R than the trivial solution presented above.

IV. DYNAMIC WEIGHTING

A. Assumptions

The basic property of the algorithm is that the *Weights* are calculated based on the questions Q. Depending on what attributes belong to Q, the *Weighting function* will assign different values to objects from *Sample*. Dynamic *Weighting function* meets the following conditions:

1. Monotonicity:

$$\forall_{a_j \in A} R(Q,\ldots) \leq R\left(Q \cup \{a_j\},\ldots\right) \tag{21}$$

2. Additivity:

$$\forall_{a_j \in A} R(Q,\ldots) = R\left(Q \setminus \{a_j\} \cup children(a_j),\ldots\right) \tag{22}$$

3. Completeness:

$$\forall_{a_j \in A^{universe'}} \left(\sum_{s_i \in s(a_j)} w(s_i) = universe\left(a_j\right)\right) \tag{23}$$

B. Algorithm

Input: *Question* Q, *Attributes tree*, function a, function p, *Sample* S, function $universe$.

Task: determine the *Weighting function* w'.

The algorithm is an iterative algorithm providing the ultimate form of the function w' in the steps going from the bottom of the *Attributes tree* to its root.

Step 1. Calculate initial value of the *Weights* w' using classical methods based on the completeness condition (23).

Step 2. If set Q is empty it ends the algorithm.

Step 3. Determine the attribute a^{parent} from set A which is the parent of all the attributes from set Q with the longest path from the root. If Q contains only one element q_j, then $a^{parent} = q_j$.

Step 4. Create the set Q' including:

- all elements from Q ,
- attribute a^{parent},
- all children of attribute a^{parent} which are parents of any attribute from set Q .

Step 5. Each attribute q_j from a set Q' assigns a temporary *Weighting function* w', as follows:

- if $q_j \in A^{universe'}$:

$$\forall_{p_i \in p(q_j)} w'_j(p_i) = w'(p_i) * \frac{universe(q_j)}{\sum_{p_k \in p(q_j)} w'(p_k)} \tag{24}$$

- otherwise:

$$\forall_{p_i \in p(q_j)} w'_j(p_i) = w'(p_i) \tag{25}$$

Step 6. If the set Q' contains only one element q_j, $it\ w'$ assigns to w'_j and it stops the algorithm.

Step 7. Determine from a set Q' such attribute q' for which the distance from the root is the largest (in the case where there is more than one attribute of the longest path, select any of them by any means).

Step 8. Determine the attribute $\mathrm{a'}$ which is the direct parent of q':

$$\mathrm{a'} = \mathrm{parent(q')} \tag{26}$$

Step 9. Modify the temporary *Weighting function* w' for the attribute $\mathrm{a'}$ as follows:

a. create subsets $s'_{positive}$ i $s'_{negative}$ of *Sample* S as follows:

$$s'_{positive} = \cup_{a_j \in children(a') \cap Q'} p(a_j) \tag{27}$$

$$s'_{negative} = \left(\cup_{a_j \in children(a')} p(a_j)\right) \setminus s'_{positive} \tag{28}$$

b. create temporary *Weighting function* w'' as follows:

$$\forall_{p_i \in s'_{positive}} w''(p_i) = max_{\,a_j \in children(a')\cap Q'}\left(w'_j(p_i)\right) \quad (29)$$

$$\forall_{p_i \in s'_{negative}} w''(p_i) = min_{\,a_j \in children(a')\setminus Q'}\left(w'_j(p_i)\right) \quad (30)$$

c. Normalize the *Weighting function* w'' in the way that the sum of *Weights* w'' is $universe(a')$:

$$\forall_{p_i \in S}\, w''(p_i)^{new} = w''(p_i)^{old} * \frac{universe(a')}{\sum_{p_i \in S} w''(p_i)^{old}} \quad (31)$$

d. modify the temporary *Weighting function* w' for the attribute a' as follows:

$$corr = \frac{\sum_{p_i \in s'_{positive}} w''(p_i)}{\sum_{p_i \in s'_{positive}} w''(p_i) + \sum_{p_i \in s'_{negative}} w''(p_i)} \quad (32)$$

$$\forall_{p_i \in S}\, w'(p_i)^{new} = w''(p_i) + corr * (w'(p_i)^{old} - w''(p_i)) \quad (33)$$

Step 10. Delete from set Q' all children of node a' :

$$Q'^{new} = Q'^{old} \setminus children(a') \quad (34)$$

and back to step 6.

C. Solution

Having determined Weighting function w′for a given *Question* Q it is possible to answer the *Question* Q using analogous calculation as in the case of the classical selection of weights:

$$R = \sum_{s_i \in \left(\bigcup_{j=1}^{m} p(q_j) \cap s\right)} w'(s_i) \quad (35)$$

D. Known features of the algorithm

Where: *Sample* S is selected from the population P using the random function with known distribution, we know the size of the entire population N and the sample size M and the weight of the sample are selected by the classical method satisfying the assumption (17), we are able to calculate statistical errors of estimates of the function R. In the case of Dynamic weighting calculating statistical errors it is difficult due to the complexity and nonlinearity of the algorithm for determining *Weighting function* w'.

Due to the condition of monotonicity (21) the algorithm tends to overestimate the weights of objects that have more than one attribute and whose weights differ in the temporary weights significantly. The better the quality of the *Sample*, the less noticeable the phenomenon is.

E. Example

Input:

Size of population is:

$$\mathrm{N} = 1000 \quad (36)$$

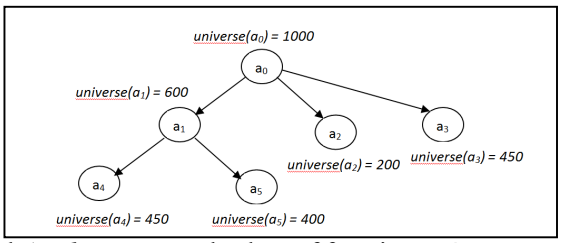
Set of attributes is:

$$\mathrm{A} = \{\mathrm{a}_0, \mathrm{a}_1, \mathrm{a}_2, \mathrm{a}_3, \mathrm{a}_4, \mathrm{a}_5\} \quad (37)$$

where:

$$\mathrm{A}^{\mathrm{universe}} = \{\mathrm{a}_0, \mathrm{a}_1, \mathrm{a}_2, \mathrm{a}_3, \mathrm{a}_4, \mathrm{a}_5\} \quad (38)$$

$$\mathrm{A}^{\mathrm{universe}'} = \{\mathrm{a}_0\} \quad (39)$$



and *Attributes tree* and values of function $universe$ are:

Fig. 1. Example: Atrributes tree and values of function universe

Sample is:

$$\mathrm{S} = \{\mathrm{s}_1, \mathrm{s}_2, \mathrm{s}_3, \mathrm{s}_4, \mathrm{s}_5, \mathrm{s}_6, \mathrm{s}_7, \mathrm{s}_8\} \quad (40)$$

Values of function a are described by the following table:

TABLE I. EXAMPLE: VALUES OF FUNCTION A FOR SAMPLE OBJECTS

Sample	***Attributes***					
	a_0	**a_1**	**a_2**	**a_3**	**a_4**	**a_5**
s_1	▪	▪			▪	
s_2	▪	▪			▪	
s_3	▪	▪	▪		▪	▪
s_4	▪	▪	▪			▪
s_5	▪		▪			
s_6	▪		▪	▪		
s_7	▪			▪		
s_8	▪			▪		

Question:

How many objects having the attribute a_3 or a_4 are in the population *P*?

$$\mathrm{Q} = \{\mathrm{a}_3, \mathrm{a}_4\} \quad (41)$$

Solution:

Step 1: Based on (32) initial values of the *Weights* w' were calculated as follows:

TABLE II. EXAMPLE: INITIAL VALUES OF WEIGHTS

	w'
s_1	125
s_2	125
s_3	125
s_4	125
s_5	125
s_6	125
s_7	125
s_8	125

Step 3: a^{parent} is determined as follows:

$$\mathrm{a}^{\mathrm{parent}} = \mathrm{a}_0 \tag{42}$$

Step 4: Set Q' was determined as follows:

$$\mathrm{Q}' = \{\mathrm{a}_0, \mathrm{a}_1, \mathrm{a}_3, \mathrm{a}_4, \mathrm{a}_5\} \tag{43}$$

Step 5: The temporary *Weighting function* w' was determined as follows:

TABLE III. EXAMPLE: INITIAL TEMPORARY WEIGHTS

Sample	Weights					
	w'_0	w'_1	w'_2	w'_3	w'_4	w'_5
s_1	125	150			150	
s_2	125	150			150	
s_3	125	150	50		150	200
s_4	125	150	50			200
s_5	125		50			
s_6	125		50	150		
s_7	125			150		
s_8	125			150		

Step 7: Attribute a_4 was determined as the node which has the longest path from the root in the *Attributes tree*.

Step 8: Attribute a_1 was determined as the direct parent of a_4:

$$\mathrm{parent}(\mathrm{a}_4) = \mathrm{a}_1 \tag{44}$$

Step 9: The temporary *Weights* w'_1 was modified as follows:

b.

TABLE IV. EXAMPLE: CREATING W''_1

	w''_1	w'_4	w'_5
s_1	150	150	
s_2	150	150	
s_3	150	150	100
s_4	100		100

c.

TABLE V. EXAMPLE: NORMALIZING W''_1

	w''_1
s_1	138
s_2	138
s_3	138
s_4	186

d.

$$\mathrm{corr} = \frac{414}{414 + 185} \cong 0{,}69$$

TABLE VI. EXAMPLE: MODIFYING W'_1

	w'^{old}_1	w'^{new}_1	w''_1
s_1	150	146	138
s_2	150	146	138
s_3	150	146	138
s_4	150	162	186

Step 10: Set Q' was modified as follows:

$$\mathrm{Q}' = \{\mathrm{a}_0, \mathrm{a}_1, \mathrm{a}_3\} \tag{45}$$

and went back to step 6.

Step 7^2: Attribute a_4 was determined as the node which has the longest path from the root in the *Attributes tree*.

Step 8^2: Attribute a_0 was determined as the direct parent of a_3:

$$parent(a_3) = a_0 \tag{46}$$

Step 9^2: The *Temporary weights* w'_0 was modified as follows:

b.

TABLE VII. EXAMPLE: CREATING W''_0

	w''_0	w'_1	w'_2	w'_3
s_1	146	146		
s_2	146	146		
s_3	146	146	50	
s_4	50	186	50	
s_5	50		50	
s_6	150		50	150
s_7	150			150
s_8	150			150

c.

TABLE VIII. EXAMPLE: NORMALIZING W''_0

	w''_0
s_1	148
s_2	148

	w''_0
s_3	148
s_4	51
s_5	51
s_6	152
s_7	152
s_8	152

d.

$$\text{corr} = \frac{899}{899+101} \cong 0{,}90$$

TABLE IX. EXAMPLE: MODIFYING w'_0

	w'^{old}_0	w'^{new}_0	w''_0
s_1	125	127	148
s_2	125	127	148
s_3	125	127	148
s_4	125	117	51
s_5	125	117	51
s_6	125	128	152
s_7	125	128	152
s_8	125	128	152

Step 10^2: Set Q' was modified as follows:

$$Q' = \{a_0\} \tag{47}$$

and went back to step 6.

Step 6: Since there is only one element in set Q' T*emporary weights* w'_0 are the final weights for *Sample*:

TABLE X. EXAMPLE: FINAL WEIGHTS FOR SAMPLE

	w'
s_1	127
s_2	127
s_3	127
s_4	117
s_5	117
s_6	128
s_7	128
s_8	128

Answer:

There are 765 objects in population P which have attributes a_3 or a_4.

$$R \cong 765 \tag{48}$$

V. CONCLUSIONS

There are cases where it is not possible to determine one constant *Weights* for *Sample* with acceptable *dispersion* e (and not only in internet research studies). In these cases Dynamic weighting algorithm may determine an individual set of *Weights* for each *Question*. The algorithm satisfies conditions (21) and (22) and (23), which makes the results reliable and useful for applied studies.

The presented algorithm is not protected against possible specific cases and possible incorrect input. Its practical implementation must take into account such cases like inconsistency of information ($e.g. universe(a_j) < \sum_{a_k \in children(a_j)} universe(a_k)$) or even contradictory information ($universe(a_j) > universe\left(parent(a_j)\right)$ at different levels of the *Attributes tree*.

The algorithm is presented in the simplest possible form. In the practice of its implementation in many places it can be optimized in terms of speed as well as memory resource consumption (e.g., through the use of temporary variables that store the temporary results) .

In order to simplify the algorithm, the questions Q are created as the sum of sets of attributes (corresponding to the logical operators OR). Practical implementations can also use the intersections (corresponding to the logical operators AND).

The assumptions of the presented algorithm are used by Gemius SA (research company) in the commercial online research studies in Poland, where the size of the population of internet users is several million people, the sample (panel of internet users) counts several thousand panelists and there are thousands of websites, internet services and web applications presented in final results of the audience research study.

In further work on the algorithm there seems to be a promising direction for estimating the statistical error of the results.

REFERENCES

[1] A. Stuart, Basic Ideas of Scientific Sampling, Hafner Publishing Company, New York, 1962.

[2] L. Kish, Survey Sampling, New York, 1965.

[3] K.W. Brown, P.C . Cozby, D.W. Kee, P.E. Worden, Research Methods in Human Development, Mountain View, CA , 1999.

[4] H. Lohr, Sampling: Design and Analysis, Duxbury, 1999

[5] W.G. Cochran, Sampling Techniques, New York, 1977.

[6] G. Kalton, Introduction to Survey Sampling. Sage Publications Series, No. 35, 1983.

[7] R. Lehtonen, E. J. Pahkinen, Practical Methods for Design and Analysis of Complex Surveys. New York, 1995.

[8] S. Levy, S. Lemeshow, Sampling of Populations: Methods and Applications, New York, 1999.

[9] S. Lohr, Sampling: Design and Analysis. Pacific Grove, 1999.

[10] S. Coffey, Internet audience measurement: a practitioner's view, Journal of Interactive Advertising, 2013

[11] P. Ejdys, T. Cisek, C. Modzelewski, Real Profilee, a new aproach to online media planning, Worldwide Audience Measurement 2003 - Online and Out-of-Home / Ambient Media, 2003

[12] Michal Wozniak, Manuel Graña, Emilio Corchado: A survey of multiple classifier systems as hybrid systems. Information Fusion 16: 3-17 (2014)

[13] Michal Wozniak, Bartosz Krawczyk: Combined classifier based on feature space partitioning. Applied Mathematics and Computer Science 22(4): 855-866 (2012)

[14] Bartosz Krawczyk, Gerald Schaefer: A hybrid classifier committee for analysing asymmetry features in breast thermograms. Appl. Soft Comput. 20: 112-118 (2014)

[15] Konrad Jackowski: Multiple Classifier System with Radial Basis Weight Function. HAIS (1) 2010: 540-547

[16] Dymitr Ruta, Bogdan Gabrys: Classifier selection for majority voting. Information Fusion 6(1): 63-81 (2005)

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 135–144
DOI: 10.15439/2014F494
ACSIS, Vol. 2

Election Algorithms Applied to the Global Aggregation in Networks of Comparators

Łukasz Sosnowski
Dituel Sp. z o.o.
ul. Ostrobramska 101 lok. 206,
04-041 Warsaw, Poland
Systems Research Institute,
Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
e-mail: l.sosnowski@dituel.pl

Andrzej Pietruszka
Institute of Mathematics,
University of Warsaw
ul. Banacha 2,
02-097 Warsaw, Poland
Dituel Sp. z o.o.
ul. Ostrobramska 101 lok. 206,
04-041 Warsaw, Poland
e-mail: a.pietruszka@dituel.pl

Stanisław Łazowy
Section of Computer Science,
The Main School of Fire Service
ul. Słowackiego 52/54,
01-629 Warsaw, Poland
e-mail: lazowy@inf.sgsp.edu.pl

***Abstract*—The paper shows the application of election algorithms in networks of comparators. We have described and adopted six election methods which have been used as an aggregator of partial results. We have performed experiments on the data gathered at the fire ground. All of them have been well described and results have been compared. The paper includes a discussion and interpretation of results obtained. It indicates the algorithm with the greatest potential to adapt and to obtain the best results.**

***Index Terms*—Networks of comparators, election algorithms, aggregation of partial results, similarity based reasoning, compound objects, fire rescue actions**

I. Introduction

SIMILARITY [1] is one of the fundamental aspects of reasoning methods used in AI. There are many techniques used by researchers to implement resemblance in practice. We can find many kinds of neural networks [2] which resolve pattern recognition problems, fuzzy sets [3] which are able to model complicated processes, rough sets [4] to perform knowledge discovery in data and many others. All of them are well-known and explored. There are many extensions of mentioned methods developed, e.g. neural networks with compound signals [5]. All these methods are specialized in resolving one of the defined problems.

In previous researches a common approach to similarity-based reasoning was developed. The same workflow is used to resolve various problems. The basic element of the authors' concept is a dedicated logical component called comparator [6]. It is responsible for examining the resemblance of a given feature between an input object and reference objects [7]. The comparator can be formally described as a function $C_B : A \rightarrow 2^{B\times[0,1]}$, where A is a set of input objects and B is a set of reference objects. Comparator outcome takes a form of weighted subsets of reference objects $C_B(a) = F(\{(b, g(\mu(a,b)) : b \in B\})$, where F is a function responsible for filtering partial results of a single comparator, e.g. min, max, top. Furthermore, $\mu(a,b)$ is a membership function of the fuzzy relation [3], which returns a similarity degree between $a \in A$ and $b \in B$, and $g(x)$ is an activation function which filters out results that are too weak. We put

$$g(x) = \begin{cases} 0 : x < p, \\ x : x \geq p \end{cases} \tag{1}$$

where p denotes the lowest acceptable similarity. One may also introduce some constraints which make $\mu(a,b) = 0$ based on the so-called exception rules [6].

The approach is based on a network of such comparators. This concept makes it possible to design a structure-driven solution as well as a flat one. The network consists of layers. They include comparators, aggregators and translators [8].

There are two types of aggregators: local and global. The functionality of a local aggregator comes down to selecting the best results for a given layer based on partial results. The functionality of the second one is focused on the synthesis of results of individual layers in order to calculate the final result.

The translator is an unit expressing the results of the one layer by objects existing in another layer. The general scheme of the type of network in question is shown in Figure 1. Complete information of the construction and operation of a network of comparators is not the subject of this article. It has been described well in the previous publication [8].

This article attempts to explore several different global aggregation methods and compare the results obtained by means of such methods. The authors also attempt to verify the importance of selection of the best results at the final stage of processing in this kind of networks. Quality level of this selection is expressed by the value of efficiency measures.

This research was motivated by authors' previous experiments. It was noticed that the final aggregation method may have a pronounced impact on results achieved. It can improve the efficiency of the model regardless of the way of evaluating

The research was supported by the Polish National Centre for Research and Development (NCBiR) - Grant No. O ROB/0010/03/001 in the frame of Defence and Security Programmes and Projects: "Modern engineering tools for decision support for commanders of the State Fire Service of Poland during Fire&Rescue operations in the buildings

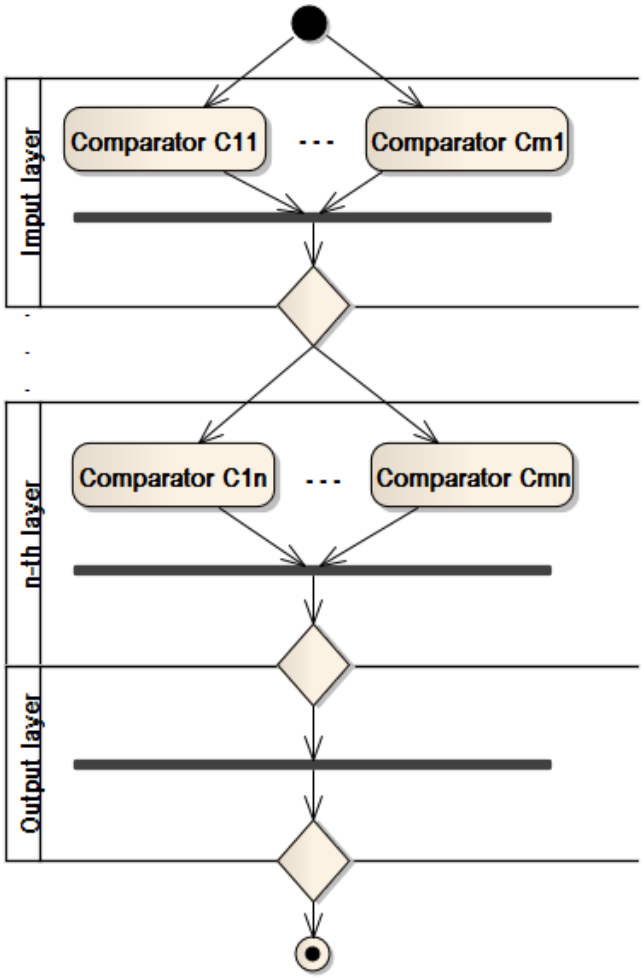


Fig. 1. General scheme of network of comparators. This is not an UML activity diagram. It only uses similar notation: oval boxes represent comparators, black horizontal lines are aggregators, the one in the output layer is a global aggregator. Diamonds between layers are translators.

similarity. Such comparison of voting algorithms has already been performed several times in the field of machine learning [9], [10]. The authors' work focuses strictly on the area of network comparators. Election algorithms were chosen as a group of methods which is well-known and described in literature. They are not the only methods which could be used, but they are easily accessible and implementable. Experiments use results obtained earlier [11] and take into account only the aspect of optimizing the process of global aggregation. This fact means that already calculated similarity results (from a previous research) are used and now attention is paid only to selecting a final set of objects. This is the main subject of this article called *global aggregation*.

The paper is organized as follows: the first section contains introductory information about context and a background of the problem presented. The second section presents the prime example used to perform the experiments, certain context required to be known to understand the results. The third section contains the description of methods used to conduct an experiment. It also contains a detailed description of elections algorithms used and their main properties. The fourth section presents data used in experiments performed and results obtained. The subsequent section discusses results and describes the best and the worst methods. It also introduces criticism of the methods used. The final part contains a brief summary.

II. Context and motivation example

The main example concerns the Fire&Rescue (F&R) actions and emerging risks during the rescue activity. Threats defined in the risk matrix [12] and objects that have vulnerabilities to these threats are taken into account. This matrix is used by fire brigades in certain countries to determine the prevalence of potential risks at the fire ground based on information collected (observation, interview, etc.).

TABLE I
Risk matrix used for evaluating the quality of fit between resembled actions. LEGEND: A1 - fear, A2 - toxic smoke, A3 - radiation, A4 - fire spreading, C - chemical substances, E1 - collapse, E2 - electricity, E3 - disease or injury, E4 - explosion

Risk/object	A1	A2	A3	A4	C	E1	E2	E3	E4
People (ME)									
Animals (T)									
Environment (U)	-								
Property (S)	-	-				-	-	-	
Rescuers (MA)								-	
Equipment (G)	-	-						-	

The previous research concerned the automation of acquiring potential risks using text descriptions and repository of historical F&R actions. The solution is based on the similarity of actions and assumption that the most resembled ones have a similar list of risks occurring. The resemblance takes into account domain knowledge acquired from experts and injected in a form of measures. In order to perform such reasoning a model of F&R action represented by a set of attributes is necessary. Data from the EWID[1] system are used to create it as well as domain knowledge from experts (e.g. division of F&R action into phases).

On this basis, the following division into stages can be distinguished: notification, disposal, recognition and activities. It is quite a rare division due to the limited number of available attributes.

Notification refers to the act of transfer of information about threat. This phase collects attributes related to time, place and approximate description of event. It contains basic information necessary to make a decision about what forces and resources should be disposed.

Disposals contain quantitative data of already disposed forces and resources. There is information on the number of rescuers, cars and equipment used for the event in question. In addition, there is a number of units of other services (medical, police, etc.) that took part in the F&R action.

[1]Polish Incident Data Reporting System used by Polish State Fire Service

Recognition is the stage at which rescuers carry out inspection, search and identification of the situation at the scene after arrival at the place of event. There are attributes describing dimensions of the place of event, building, size of event and also information about the existence of internal hydrants, smoke detectors, etc.

Actions form the actual start of firefighting operations. After collecting the required information at previous stages, firefighters start activities related to the neutralization of threats. They prepare an action strategy, assign tasks for rescuers and decide whether to use the specialized equipment.

These four stages are not independent. They are ordered in a sequence. If the notification materialized the disposal cannot be realized.

The solution mentioned earlier designates subsets of the most similar actions from the perspective of individual parts. The similarity ranking of each part may comprise a different list of preferred objects in order of significance. At this point s method has to be implemented, consisting in the combination of partial results into one coherent answer of the system which is a final result in form of subset of F&R objects with assigned risk labels.

This method is a decision problem of selecting the object that best meets the preferences of individual parts of F&R (based on similarity). In other words, it is analogous to the election case, where support for a candidate is expressed by voting. This analogy was the motivation for carrying out research on the application of elections algorithms to final selection of the results set (global aggregation). It was also an impulse to examining the importance of the choice of method. The authors were interested in the type of impact of the choice of election methods on final results achieved at the fire ground.

III. Methods

The global aggregation method is a part of a bigger solution mentioned in Section II. Therefore, the full path which must be passed in order to obtain the results described furtherwill be presented here.

The first step of the proposed solution is to design a network of comparators. The network is based on expert knowledge which should be part of the F&R ontology. It consists of concepts and their relations. It describes different aspects of action using evidenced concepts [13].

At the beginning division of an F&R action provided by expert is taken into account and described in Section II. The authors seek a possibility of comparing single parts and having partial results from each of them. This division consists of four parts: notification, disposal, recognition, actions. Each of them is represented by a composite comparator [14]. It means that each of them is independent sub-net. *Notification* consists of nine comparators responsible for examining different features associated with notification of fire. The complete list of these comparators is shown in Table II. The next phase is disposal. The sub-net of *disposal* consists of thirteen comparators. All these features are connected with a group of activities assigned to a disposal part. Other parts are constructed in an analogous manner. *Recognition* includes eleven comparators and *actions* consists of ten of them. Tables III, IV and V show respectively the detailed list of comparators for particular part of action. These tables also contain similarity measures applied in a particular comparator. This information is required for complete implementation.

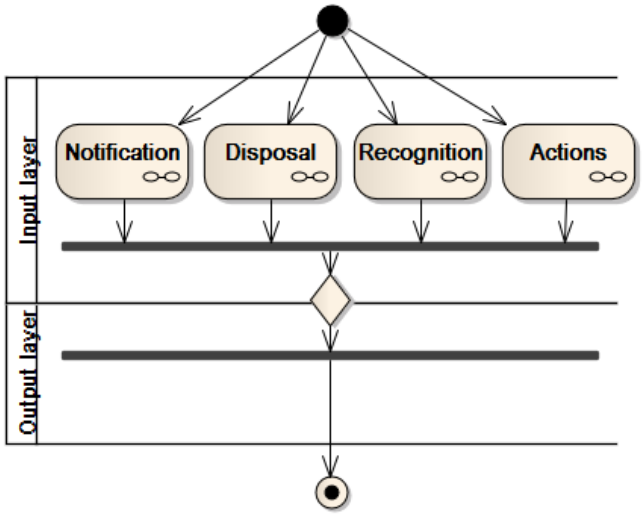


Fig. 2. The scheme of the network of comparators used for the designation of the most similar set of F&R actions. The oval boxes represent composite comparators. The translator in the output layer was skipped because there translation objects are not necessary in this case.

A particular comparator has a function to evaluate similarity between a given pair of objects. The pairs are created in combination of the input object and the reference object (taken from the F&R repository one by one). Each pair has assigned value of similarity determined by a given comparator. The calculation for a single input object implies a number of local results sets in the form of $\{(b_i, \mu(a, b_i) : b_i \in B\}$, where B is a set of reference objects.

Our solution consists of two parts. The first one is responsible for building a similarity model and computing partial results of resemblance. The second one is dedicated to optimization and aggregation of results. These two methods are classified as lazy [15] and eager types [16] respectively. The similarity model evaluates a given pair after calculating similarity for each comparator and it is a lazy one. On the other hand, aggregation of partial results requires learning certain parameters (in certain cases) or making certain decisions earlier. This part is classified as eager methods, e.g. learning weights for weighted average aggregator using genetic algorithm.

A *leave one out* [17] method is used to obtain four matrices with partial results. Each object is treated once as input objects against the remaining objects of the reference set. Such processing is performed for every single object one by one.

Our method can evaluate partial similarities and aggregate them to a higher level of resemblance and finally to a global one. The aggregator is a dedicated unit to handle such processes.

TABLE II
COMPARATORS OF FEATURES FOR NOTIFICATION PHASE

Comparator	Description	Comparison function
Time	F&R time. t_{a1}, t_{a2} - action time in compared objects, t_{e1}, t_{e2} - extinguish time in compared objects.	$\frac{t_1+t_2}{2}$, where $t_1 = \begin{cases} 1.0: & t_{d1}=0 \\ 1-\frac{t_{d1}}{\max(t_{a1},t_{a2})}: & t_{d1}\neq 0 \end{cases}$ $t_2 = \begin{cases} 1.0: & t_{d2}=0 \\ 1-\frac{t_{d2}}{\max(t_{e1},t_{e2})}: & t_{d2}\neq 0 \end{cases}$ where $t_{d1} = \lvert t_{a1} - t_{a2}\rvert$, $t_{d2} = \lvert t_{e1} - t_{e2}\rvert$.
Event place	F&R place. Places divided into a three level hierarchy. The first level is the most general.	$\mu(a,b) = \begin{cases} 1.0: & a_3 = b_3, \\ 0.8: & a_2 = b_2, a_3 \neq b_3, \\ 0.55: & a_1 = b_1, a_2 \neq b_2 \end{cases}$ where a_n, b_n, $n \in 1..3$ - level hierarchy.
Access	Access to F&R place. The Access is divided into a two level hierarchy. The first level is the most general.	$\mu(a,b) = \begin{cases} 1.0: & a_2 = b_2 \\ 0.55: & a_1 = b_1, a_2 \neq b_2 \end{cases}$ where a_n, b_n, $n \in 1..2$ - level hierarchy.
Building height	Building height for F&R place. Building heights are divided into clusters.	$\mu(a,b) = ms[C_a][C_b]$ where ms - matrix of similarity between clusters, C_a, C_b - cluster values for building height in compared objects.
Building type	Building type for F&R like a building low, a building medium, a building hight.	$\mu(a,b) = \begin{cases} 1.0: & a = b \\ 0.0: & a \neq b \end{cases}$ where a, b - building types in compared objects.
Notification channel	Notification channel about fire, e.g. by phone, by radio.	The same as above.
Entity notifier	Entity notifier about fire, e.g. an employees.	The same as above.
ZL category	Building category risk to humans.	$\mu(a,b) = ms[a][b]$ where ms - matrix of similarity, a, b - building category risk to humans.
Distance	Distance between the first units and the F&R.	$\mu(a,b) = \begin{cases} 1.0: & a = b \\ 1-\frac{\lvert a-b\rvert}{6}: & a \neq b \end{cases}$ where a, b - distance in compared objects.

TABLE III
COMPARATORS OF FEATURES FOR DISPOSAL PHASE

Comparator	Description	Comparison function
Firefighters	How many firefighters were sent to F&R.	$\mu(a,b) = \begin{cases} 1.0: & a = b \\ 1-\frac{\lvert a-b\rvert}{\max(a,b)}: & a \neq b \end{cases}$ where a, b - number of firefighters in compared objects.
Vehicles	How many vehicles were sent to F&R.	The same as above.
Firefighting cars	How many firefighting cars were sent to F&R.	The same as above.
Special cars	How many special cars were sent to F&R.	The same as above.
Additional vehicles	How many additional vehicles were sent to F&R.	The same as above.
Ambulances	How many ambulances were sent to F&R.	The same as above.
Power emergencies	How many power emergencies were sent to F&R.	The same as above.
Gas emergencies	How many gas emergencies were sent to F&R action.	The same as above.
Forest services	How many forest services were sent to F&R.	The same as above.
Police	How many police services were sent to F&R.	The same as above.
City guards	How many city guards were sent to F&R.	The same as above.
Other services	How many other services were sent to F&R.	The same as above.
Other region vehicles	How many vehicles from other regions were sent to F&R.	The same as above.

It has been noted in the research that the elections algorithms theory [18] is one of the approaches available, dealing with methods of selection of the best candidate for a group of voters. In this case, similarity results can be treated as voting results, comparators as voters and reference objects as candidates. The optimization task is to select the best candidate who will became a winner acceptable for the majority of voters.

There are a lot of known and available election algorithms. Six of them have been shosen, those which are the most popular, well-known or fulfil important criteria (e.g. majority, condorcet winner, condorcet looser, etc.) [19].

A. Plurality voting system

This is a single winner voting system. Each voter votes for one candidate only [20]. The winner is the one who receives the highest number of votes. This voting system is very easy for voters and it is easy to implement. However, it fails to provide information about preferences and support for individual candidates only for the most supported one. This is quite a popular method in real elections.

Implementation performs voting for each input object iteratively. For a given input object the candidate with the highest score is selected. The score is calculated from results of comparators. The winner for given comparator is the first candidate with the maximum similarity value. After this selection there are four votes. The global winner is the one with the highest global score. In the case of a tie, the first candidate is obtained.

TABLE IV
COMPARATORS OF FEATURES FOR RECOGNITION PHASE

Comparator	Description	Comparison function
Dust explosion	During F&R there was the dust explosion.	$\mu(a,b) = \begin{cases} 1.0: & a = b \\ 0.0: & a \neq b \end{cases}$ where a, b - the dust explosion in compared objects.
Gas explosion	During F&R there was the gas explosion.	The same as above.
Object dimensions	Object dimensions for F&R place. The object dimension was divided into clusters.	$\mu(a,b) = \begin{cases} 0,8 + (0,2 * f): & diff = 0 \\ 0,5 + (0,3 * f): & diff = 1 \\ 0: & diff > 1 \end{cases}$ where $f = (1 - \frac{\lvert a-b \rvert}{\max(a,b)})$, $diff$ - difference between clusters, a, b - dimensions in compared objects.
Event size	Event size for F&R.	$\mu(a,b) = \begin{cases} 1.0: & a = b \\ 1 - \frac{\lvert a-b \rvert}{\max(a,b)}: & a \neq b \end{cases}$ where a, b - event size of compared objects.
Without PSP	The fire was extinguished without PSP.	$\mu(a,b) = \begin{cases} 1.0: & a = b \\ 0.0: & a \neq b \end{cases}$ where a, b - fire was extinguished without PSP in compared objects.
Fire cause	Fire cause. Fire causes are divided into clusters.	$\mu(a,b) = ms[C_a][C_b]$ where ms - matrix of similarity between clusters, C_a, C_b - cluster values for fire cause in compared objects.
Internal hydrants	Status internal hydrants in a building during F&R.	$\mu(a,b) = ms[a][b]$ where ms - matrix of similarity, a, b - statuses of internal hydrants.
Smoke devices	Status smoke devices in a building during F&R.	The same as above.
Fire extinguishing	Status fire extinguishing in a building during F&R.	The same as above.
Auto transmission	Status auto transmission system in a building during F&R.	The same as above.
Fire detection	Status fire detection system in building during F&R.	The same as above.

TABLE V
COMPARATORS OF FEATURES FOR ACTIONS PHASE

Comparator	Description	Comparison function
Extinguishing on offensive	Flag indicates than during F&R extinguishing was used in offensive activities.	$\mu(a,b) = \begin{cases} 1.0: & a = b \\ 0.0: & a \neq b \end{cases}$ where a, b - extinguishing in offensive activities in compared objects.
Extinguishing in defense	Flag indicates than during F&R extinguishing was used in defensive activities.	The same as above.
Fire extinguisher	Fire extinguishers used during F&R with values.	$\mu(a,b) = \begin{cases} 1.0: & a = b \\ 1 - \frac{\lvert a-b \rvert}{\max(a,b)}: & a \neq b \end{cases}$ where a, b - number of fire extinguishers in compared objects.
Extinguishing media	Used extinguishing media during F&R.	The same as above.
Medical	Medical assistance provided during F&R.	The same as above.
Actions taken	Set of activities taken during F&R. Matrix of similarity defined by an expert.	$\mu(A,B) = \forall r \in R, r \in A, r \in B, \frac{\sum ms[a][b]}{n}$ where ms - matrix of similarity, A, B - sets of activities in compared objects, R - a reference set, n - size the reference set.
Medical actions	Set of medical activities taken during F&R.	The same as above.
Activities place	Place of activities taken during F&R.	The same as above.
Equipment used	Equipment used during F&R.	The same as above.
Water supply	Water supply methods during F&R.	The same as above.

This method has a computational complexity estimated at $O(N)$. This provides for good computational properties in terms of time and computational power consumption.

B. Borda count

This is also a single winner voting method where voters produce a ranking of candidates in order of preferences [21]. The candidate receives a number of points connected with a position in ranking. The higher position in the ranking the more points candidate gets. There are various scoring methods: promoting higher place more than the lower, linear or specific one (e.g. only first three places are scored). The winner is the candidate with the highest point result.

Implementation treats comparators as four voters. Each voter gives a ranking for all candidates. Rankings are created on the basis of similarity value for a given input object and particular candidates. Candidates are reference objects. Each candidate receives points. Linear scale is used. The candidate of the first place takes maximum number of points (in this case 406). The score function is given by the following formula:

$$score(a) = n(C) + 1 - RankPos(a), \tag{2}$$

where $n(C)$-quantity of candidates, $RankPos(a)$ - position in ranking of candidate a. After that all points from particular rankings are summed up for each candidate. The last part is selecting a candidate with a maximum number of points. In the case of a tie, the first candidate with the maximum score is obtained.

This method is characterized by a very good computational complexity estimated at $O(N)$. One of the variable parameter

is a method of allocating points. It is an important procedure because it favors higher ranking position.

C. Copeland's method

This is a condorset [19] method which returns results in form of a ranking. The score for setting a ranking position is calculated from the number of wins in pairs minus the number of defeats [22]. This is a round-robin tournament methods which is easy understandable and easy implementable in software solutions. The winner is a candidate with the highest score. Implementation performs this method for each input object independently. For a given input object it starts with creating the cartesian product of candidates (F&R objects). The winner is calculated for each pair. Every winner is computed with data from four comparators (voters). Each of them specifies which similarity value (between input object and given candidate) is bigger. The local winner gets one point, the local looser gets minus one point. In the case of a tie, both take a zero point. After resolving this problem for all comparators, points for both candidates are summed up. The winner of a single duel is the one who scores more points. This procedure is repeated for each pair of candidates to create full tournament table with score for each candidate. The global winner for the current input object is the candidate with the maximum global score calculated in a following way:

$$score(a) = wins(a) + defeats(a) \tag{3}$$

where $wins(a)$ is a number of points for wins (positive value) for candidate a, $defeats(a)$ is a number of points for defeats (negative value).

This method has a computational complexity estimated at $O(N^2)$. It does not provide any parameters to set.

D. Approval voting

This is a single winner voting method where each voter may approve or disapprove of any candidate from a ballot. It means that the voter has to specify approval or disapproval as regards each candidate. Consequently, the ballot designates the accepted set of candidates for a particular voter. The winner is the candidate who has the highest number of votes of approval [23].

This algorithm has been implemented by means of four comparators as voters and 405 reference objects as candidates. Approval voting is performed for each input object (406 times). This method has been adapted to the requirements in the following way: the approval factor is found, starting from the biggest one (1.0); then it is reduced by 0.01 in the case of failure to obtain majority in voting. Then, the approval factor is the threshold for similarity value of pair of objects. If the resemblance value is greater than or equal to the threshold, the vote is interpreted as approved or as disapproved for a candidate who is the reference object in given pair. In the case of a tie, first candidate with the highest number of *approved* votes is obtained. For each input object the winner is calculated in the same way.

This method is characterized by a very good computational complexity estimated at $O(N)$. This allows for very efficient calculation of the final results. Implementation presented could have various values of the reduction factor. It has impact on the quality of results and the speed of calculation.

E. Range voting

This is a single winner voting system. The voting is realized by rating ballots. A rate scale is specified, e.g. $[0, 1]$ or $[0, 100]$. Voters rate each candidate with own score matched with a fixed scale [24]. All candidates scores are summarized. The winner is a candidate with the highest sum of points. If certain candidates are not scored, the zero value is assigned. In general, all candidates should be rated.

Implementation takes on the form of calculating the mean value for each pair (input object and reference object). There are four comparators. Each of them provides similarity value for a given pair. The mean value is calculated from these four similarities for each pair.

This method has a computational complexity estimated at $O(N)$. It does not provide any parameters.

F. Weighted voting

This is a voting system which makes it possible to favour certain voters. In real life, one can find this kind of voting on boards of companies, where there are different shares or stocks [25]. The weight connected with a vote makes it stronger or weaker depending on the value of the weight. In this particular case, if all weights are equal, the system is identical to *range voting*.

In this case, voting is implemented in a similar way to *range voting*. This method works only if weights are given. Weights are indicated as a w_i where i is a number of comparator.

The solution has been expanded by adding the sub-optimal learning procedure determining weights for voting. A genetic algorithm [26] is used in order to find weights which give the highest evaluation score for this kind of voting. There are four weights described by the following formulas:

$$w_1 = \frac{n}{n+d+r+a} : (n+d+r+a) \neq 0 \tag{4}$$

$$w_2 = \frac{d}{n+d+r+a} : (n+d+r+a) \neq 0 \tag{5}$$

$$w_3 = \frac{r}{n+d+r+a} : (n+d+r+a) \neq 0 \tag{6}$$

$$w_4 = \frac{a}{n+d+r+a} : (n+d+r+a) \neq 0 \tag{7}$$

where n - factor responsible for *notification*, d - factor responsible for *disposal*, r - factor responsible for *recognition*, a-factor responsible for *actions*. Each of them is in the range of $[0, 63]$, but the sum cannot be zero. The chromosome in this representation contains four parts dedicated to each factor. Each of them is coded in six bits, i.e. there is a twenty four bits chromosome. As genetic operations tournament crossover with probability 0.5 and mutation with probability 0.065 is

TABLE VI
THE PERFORMANCE COMPARISON FOR THE NETWORK OF COMPARATORS (NOC) WITH DIFFERENT METHODS OF AGGREGATION

Algorithm	Precision	Recall	F1-score
NoC with Approval voting	0.73	0.68	0.65
NoC with Borda count	0.77	0.73	0.69
NoC with Copeland's method	0.77	0.73	**0.70**
NoC with Plurality voting	0.68	0.63	0.61
NoC with Range voting	0.76	0.73	0.69
NoC with Weighted voting	**0.78**	**0.74**	**0.70**

used. As fitting function the measure in the following form is implemented:

$$f(ch) = \frac{3 * RQ + sum(F1score)}{4 * ARQ} : ARQ \neq 0. \quad (8)$$

where ch is a chromosome, *RQ* is a number of identified risks, *ARQ* is a total quantity of all considered risks and *sum(F1score)* is a sum of all F1score values.

Parameters for genetic algorithm were chosen in an experimental way. Population stands at one hundred individuals. Weights have been learnt by means of a training set consisting of 136 out of 406 F&R actions (33%). The procedure was repeated ten times. Every time the population was initialized with random values. Termination condition was reached for a one hundred generation. The final weights are the ones with maximum evaluations. The evaluation function (8) took into account the number of recognized risks, as well as the overall prediction quality.

This method allows to use many different types of crossover operations, mutations and successions of a population. A number of combinations of particular parts of procedures in question may be considered.

IV. RESULTS

This research is based on data available in the EWID system. The data describe F&R after it had already finished. This is one of the disadvantages of this set of data. There is no information about the point in time when something has happened for a given F&R. The only information is whether it had happened over the duration of the whole action. The full set of data consists of 291 683 F&Rs. A subset of 406 F&Rs is used in these experiments. These actions have been labeled with risks from Table I by experts. The data consist of 506 binary, numeric, multi-value attributes and descriptive attributes. This research does not take into account the descriptive ones. This set has been divided into four subsets according to four stages of action (mentioned in the previous section). Partial results data set is obtained and it contains 164430 pairs of objects after calculating the similarity by composite comparators. The pair consists of two objects representing F&R actions. This is not a complete Cartesian product. Some pairs have been eliminated by activation functions.

Partial results were achieved by means of default parameters of the network. Activation functions parameters have the *0.5* value which limited the results to the ones with similarity value greater than or equal to that value. The aggregations method inside all composite comparators was implemented as a mean function.

TABLE VIII
STATISTICAL MEASURES OF THE RESULTS GROUPED BY RISKS ABBREVIATIONS: MIN - MINIMUM, MAX - MAXIMUM, AVG - MEAN, MED - MEDIAN, STD - STANDARD DEVIATION, RAN - RANGE. VALUES PRESENTED IN *F1-score*

Risk	MIN	MAX	AVG	MED	STD	RAN
A1_MA	0.26	0.38	0.32	0.33	0.05	0.12
A1_ME	0.90	0.92	0.91	0.90	0.01	0.02
A1_T	0.07	0.19	0.11	0.10	0.05	0.12
A2_MA	0.85	0.91	0.87	0.86	0.02	0.06
A2_ME	0.45	0.89	0.77	0.88	0.18	0.44
A2_S	0.07	0.22	0.14	0.13	0.06	0.15
A2_T	0.02	0.20	0.12	0.13	0.06	0.18
A2_U	0.22	0.45	0.33	0.33	0.08	0.23
A4_G	0.11	0.34	0.19	0.16	0.09	0.23
A4_MA	0.16	0.46	0.28	0.27	0.10	0.30
A4_ME	0.16	0.34	0.23	0.22	0.07	0.18
A4_S	0.18	0.38	0.28	0.28	0.07	0.20
A4_T	0.17	0.67	0.41	0.40	0.25	0.50
E1_MA	0.20	0.44	0.31	0.31	0.09	0.24
E1_ME	0.07	0.40	0.25	0.27	0.17	0.33
E2_MA	0.02	0.14	0.07	0.07	0.04	0.12
E2_ME	0.05	0.13	0.09	0.09	0.04	0.08
E2_S	-	-	-	-	-	-
E3_G	-	-	-	-	-	-
E3_MA	0.08	0.17	0.13	0.14	0.03	0.09
E3_ME	0.13	0.36	0.24	0.24	0.10	0.23
E4_MA	0.02	0.22	0.11	0.14	0.09	0.20
E4_ME	-	-	-	-	-	-
E4_S	-	-	-	-	-	-

The procedure of global aggregation using the Weighted Voting algorithm was performed for the following weights: $\frac{2}{70}, \frac{48}{70}, \frac{1}{70}, \frac{19}{70}$ for notification, disposals, recognition and actions comparators respectively. These values have been achieved in the learning process described above. The described partial results set is a data source for experiment in this particular paper.

The experiment consists of selecting the best reference object for each input object from leave-one-out method in such a way as to ensure the greatest similarity within each part of the F&R action (according to the adopted division). Here, six algorithms in question were applied (one by one) and efficiency of the overall solution was measured by the quality of final results.

In the experiment, measures dedicated for classifiers such as: *precision*, *recall* and *F1-score* are used. Two types of measurements were performed for each of the algorithms. The first was on the assessment of individual pairs as classification results while the other evaluated the efficacy from the perspective of individual risks.

In the first case, the relevant set is a collection of risks labels associated with the input object, and the retrieved set is the one assigned to the reference object identified as the most similar. *Precision*, *Recall* and *F1-score* are calculated for each best pair. Lastly, all these three factors are averaged. Final results of this measurement are presented in Table VI.

In the second case, classification effectiveness measures are

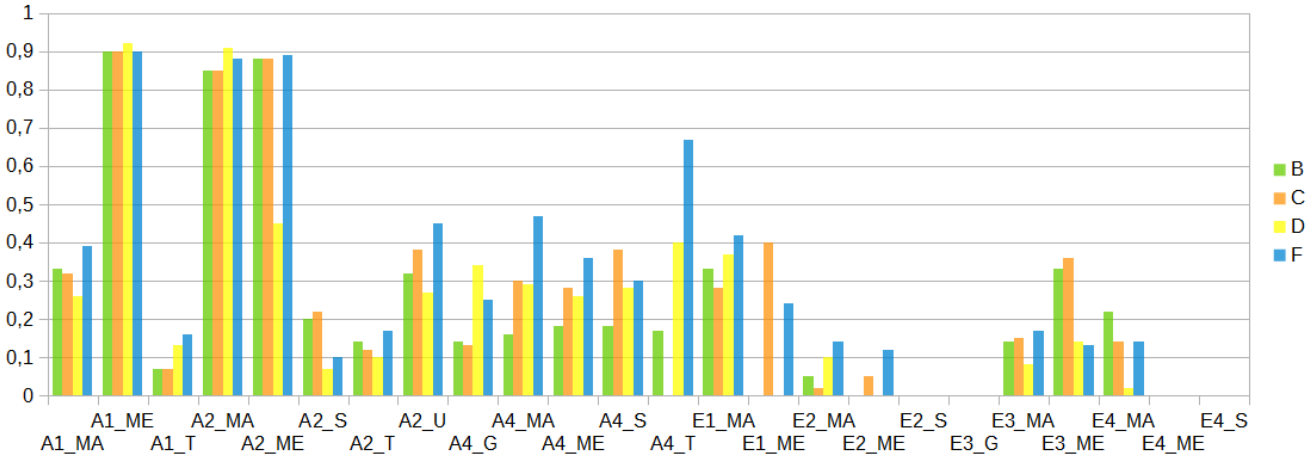


Fig. 3. The best algorithms grouped by risks evaluated by the *score* function. Abbreviations: B- Borda count, C-Copeland's method, D-Approval, F-Weighted voting. The left axis shows the *F1-score* value, the bottom one selected risks from the risk matrix.

TABLE VII
EXPERIMENTAL RESULTS FOR ALL AGGREGATION ALGORITHMS IN QUESTION EXPRESSED IN F1-SCORE, PRECISION AND RECALL VALUES. ABBREVIATIONS: A - PLURALITY, B - BORDA COUNT, C - COPELAND'S METHOD, D - APPROVAL, E - RANGE VOTING, F - WEIGHTED VOTING

	F1-score						Precision						Recall					
Risk	*A*	*B*	*C*	*D*	*E*	*F*	*A*	*B*	*C*	*D*	*E*	*F*	*A*	*B*	*C*	*D*	*E*	*F*
A1_MA	0.27	0.33	0.32	0.26	0.36	**0.38**	0.37	0.37	0.36	**0.46**	0.40	0.41	0.21	0.30	0.28	0.18	0.33	**0.36**
A1_ME	0.90	0.90	0.90	**0.92**	0.91	0.90	0.87	**0.91**	**0.91**	0.87	**0.91**	**0.91**	**0.93**	0.89	0.89	0.97	0.91	0.89
A1_T	0.07	0.07	0.07	0.13	0.12	**0.19**	0.08	0.08	0.10	**0.21**	0.11	0.20	0.06	0.06	0.06	0.09	0.13	**0.19**
A2_MA	0.87	0.85	0.85	**0.91**	0.85	0.86	0.86	**0.87**	**0.87**	0.86	0.86	**0.87**	0.89	0.83	0.84	**0.96**	0.84	0.85
A2_ME	0.66	0.88	0.88	0.45	**0.89**	0.88	0.88	0.89	0.89	**0.91**	0.89	0.89	0.53	0.87	0.87	0.30	**0.89**	0.87
A2_S	0.07	0.20	**0.22**	0.07	0.17	0.09	0.11	0.33	**0.50**	0.07	0.22	0.14	0.07	**0.14**	**0.14**	0.07	**0.14**	0.07
A2_T	0.02	0.14	0.12	0.10	0.13	**0.20**	0.03	0.14	0.13	0.20	0.12	**0.22**	0.02	0.14	0.11	0.07	0.14	**0.18**
A2_U	0.22	0.32	0.38	0.27	0.34	**0.45**	0.35	0.34	0.38	**0.52**	0.37	0.49	0.16	0.31	0.39	0.18	0.32	**0.42**
A4_G	0.18	0.14	0.13	**0.34**	0.11	0.25	0.25	0.14	0.13	**0.40**	0.09	0.22	0.14	0.14	0.14	**0.29**	0.14	**0.29**
A4_MA	0.22	0.16	0.30	0.29	0.24	**0.46**	0.29	0.20	0.35	0.32	0.25	**0.42**	0.18	0.14	0.27	0.27	0.23	**0.50**
A4_ME	0.18	0.18	0.28	0.26	0.16	**0.34**	0.21	0.21	0.30	**0.33**	0.18	0.32	0.15	0.15	0.26	0.22	0.15	**0.37**
A4_S	0.28	0.18	**0.38**	0.28	0.23	0.31	0.35	0.24	**0.46**	0.41	0.25	0.32	0.24	0.15	**0.33**	0.21	0.21	0.30
A4_T	-	0.17	-	0.40	-	**0.67**	-	0.14	-	0.33	-	**0.50**	-	0.22	-	0.50	-	**1.00**
E1_MA	0.20	0.33	0.28	0.37	0.22	**0.44**	0.18	**0.50**	0.25	0.43	0.22	0.44	0.22	0.25	0.33	0.33	0.22	**0.44**
E1_ME	0.07	-	**0.40**	-		0.27	0.11	-	**1.00**	-	-	0.28	0.05	-	0.25	-	-	**0.26**
E2_MA	0.06	0.05	0.02	0.10	0.07	**0.14**	0.09	0.05	0.03	**0.21**	0.07	0.13	0.05	0.05	0.02	0.07	0.07	**0.15**
E2_ME	-	-	0.05	-	0.09	**0.13**	-	-	0.06	-	0.09	**0.17**	-	-	0.05	-	**0.10**	**0.10**
E2_S	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
E3_G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
E3_MA	**0.17**	0.14	0.15	0.08	0.13	0.12	0.21	0.29	**0.40**	0.17	0.25	0.18	**0.14**	0.09	0.09	0.05	0.09	0.09
E3_ME	0.19	0.33	**0.36**	0.14	0.28	0.13	0.17	**0.67**	1.00	0.20	0.40	0.17	**0.22**	**0.22**	**0.22**	0.11	**0.22**	0.11
E4_MA	0.14	**0.22**	0.02	0.02	-	0.15	0.01	**0.50**	0.14	0.01	-	0.17	**0.29**	0.14	0.14	**0.29**	-	0.14
E4_ME	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
E4_S	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Score	0.62	0.62	0.66	0.61	0.59	**0.70**	0.62	0.63	0.68	0.63	0.59	**0.70**	0.61	0.62	0.66	0.62	0.58	**0.70**
Mean	0.19	0.23	0.26	0.22	0.22	**0.31**	0.23	0.29	**0.34**	0.29	0.24	0.31	0.19	0.21	0.24	0.22	0.21	**0.32**

calculated for each individual risk from the threat matrix. The calculation uses a set of actions containing a given risk as a

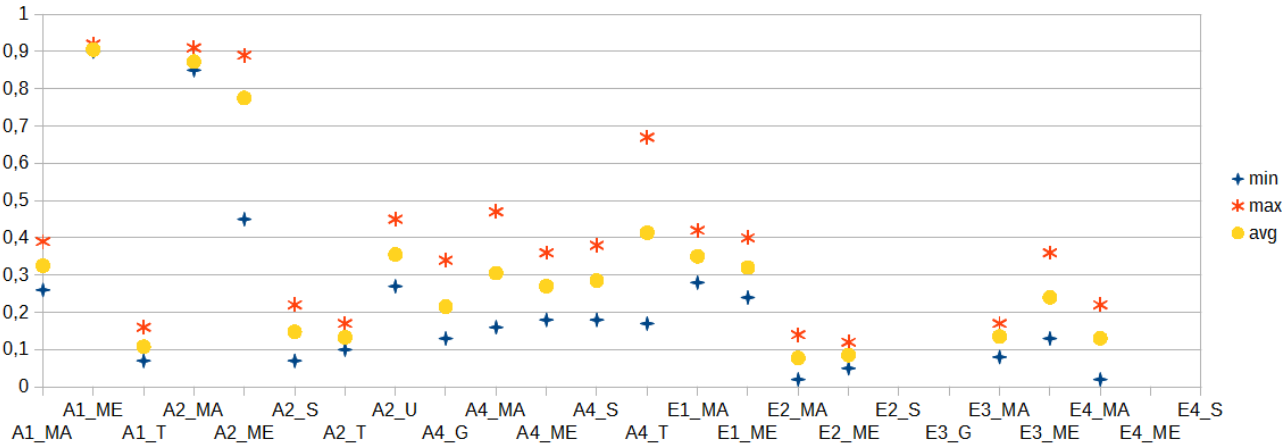


Fig. 4. Chart with the minimum, maximum and average results value for particular risks. It shows the range of values between the minimum and maximum followed by the mean value for all algorithms.

relevant set (on the basis of assignment of labels by an expert). The retrieved set consists of actions (reference objects) found by means of our solution. Table VII contains the performance comparison of all classification methods expressed by *F1-score*, *Precision* and *Recall*. Additionally, there are the *score* and *mean* rows at the bottom. They show the values of global scoring of algorithms. The first one is a normalized value of the number of recognized risks boosted tree times and summed with a total sum of *F1-score*. The second one is a mean value but calculated for all risks from the domain. It means that the zero value is taken for those risks which have not been identified. The *score* function is used to evaluate global efficiency of a particular algorithm in application in this experiment.

Table VIII contains statistical factors of the achieved results. They have been calculated on the basis of rows of Table VII limited to the *F1-score* measure. These are the most frequently used factors such as: min, max, mean, etc. The important one is the range factor which shows the difference between the best and the worst result for a given risk.

Versatility in identifying risks is a desired feature that should characterize the best solutions. Therefore, the function of evaluation algorithms (score) highly rewards those algorithms that identify the broadest spectrum of risks. Evaluation of a particular algorithm using only a number of identified risks is presented in Figure 5.

Results presented are extensions of results presented in the previous publication of the authors [11]. Three other methods are described in the publication: Naive Bayes, ESA, kNN Canberra. Authors' experiments are based on the same set of data. The results in relation to the previous ones show progress in terms of growth performance measures.

V. Discussion

It can be noted in Tables VI and VII that the best results for the whole described solution have been achieved by Weighted

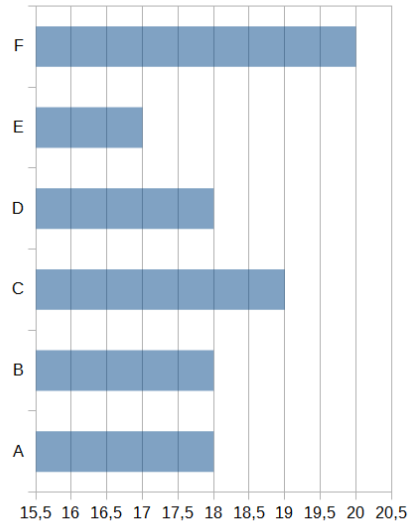


Fig. 5. The comparison of the number of identified risks by particular method. X-axis represents the number of different risks which were at least once properly assigned to an F&R action. The entire space consists of 24 risks. Abbreviations: A - Approval voting, B - Borda count, C - Copeland's method, D - Plurality voting system, E - Range voting, F - Weighted voting.

Voting algorithm. They have been obtained in terms of the largest number of identified risks as well as the highest values of effectiveness measures. Network of Comparators (NoC WA) with global aggregation implemented by means of this algorithm has identified 20 risks among all 24 existing in the threat matrix. All other algorithms obtained worse results. The second in this ranking is the result of the Copeland's method with a value of one less.

In terms of classification effectiveness from the risk point

of view it has achieved the highest value of score function calculated for all the three considered measures (0.7). Also, the mean value is the highest for both *F1-score* and *Recall* case and a little bit lower in the case of *Precision*. Similarly to the averaged measure for each best pair. The NoC WA has achieved 0.78, 0.74 and 0.70 for *Precision*, *Recall* and *F1-score* respectively.

Results presented prove that the problem of selecting the global aggregation method in a network of comparators is important and may have big impact on achieved results. They also show that election algorithms are a very good solution for optimization problems described. From the point of view of the arithmetic mean (range voting), all the methods used have obtained a higher value of the *score* function. Table VIII shows statistical factors for results generated from the risk perspective. Figure 4 demonstrates we can see exactly what the extreme values achieved by algorithms for individual risks are. This shows that in this range results can vary depending on the selection of an algorithm. It is clear that in certain cases it is the value of 0.5 which is a 50% of the scale. Additionally, it demonstrates how the average value fits the range.

Weak points of experiments are the tie solving methods. In experiments, the authors have used the method of getting the first candidate as a winner. This is quite a random method not showing real preferences.

VI. Conclusions

This paper shows similarity-based solution for recognition and identification purposes. This article in particular raises the problem of selecting the optimal results set. It presents theory and working solution on the example of recognition risks at the fire ground.

In order to investigate the problem of selecting the aggregation method for the network of comparators, six methods have been tested. The theory part for each of them is presented in detail as well as adaptation for our concrete problem. One of the strongest conclusions is that the election algorithms are a very powerful methods for implementation in such cases.

The results achieved are very promising. They are very good in comparison with different attempts to resolve risk recognition problem using the same set of data [12]. Based on these results we recommend a Weighted Average election algorithm to use as a global aggregator in the network of comparators. This method makes it possible to adapt to a specific problem by learning weights. For the purposes of the NoC solution, we have developed a learning method using a genetic algorithm.

The NoC is a very useful solution in the field of AI applications. It allows to build complex networks of comparators that can be used as ensemble classifiers [27]. In this way, very complex decision support problems can be solved as well as classification, recognition and identification problems.

The future work should concentrate on propagating the optimization method on the level of a single comparator and particular layer of the network (not only for output layer). Our research has showed that this might be a large field of improvement of performance for the whole solution.

References

[1] A. Tversky and E. Shafir, *Preference, Belief, and Similarity: Selected Writings*, ser. Bradford books. MIT Press, 2004.

[2] C. M. Bishop, *Neural Networks for Pattern Recognition*. New York, NY, USA: Oxford University Press, Inc., 1995.

[3] J. Kacprzyk, *Multistage Fuzzy Control: A Model-based Approach to Fuzzy Control and Decision Making*. John Wiley & Sons, Limited, 2012.

[4] S. O. Kuznetsov and D. Slezak, "Data mining and soft computing," *Int. J. General Systems*, vol. 42, no. 6, pp. 543–545, 2013.

[5] M. S. Szczuka and D. Slezak, "Feedforward neural networks for compound signals." *Theor. Comput. Sci.*, vol. 412, no. 42, pp. 5960–5973, 2011.

[6] D. Ślęzak and Ł. Sosnowski, "SQL-based Compound Object Comparators: A Case Study of Images Stored in ICE," in *Proc. of FGIT-ASEA 2010*, ser. Communications in Computer and Information Science, vol. 117, 2010, pp. 303–316.

[7] L. Sosnowski and D. Slezak, "How to design a network of comparators," in *Brain and Health Informatics*, 2013, pp. 389–398.

[8] ——, "Networks of compound object comparators," in *FUZZ-IEEE*, 2013, pp. 1–8.

[9] L. I. Kuncheva and J. J. Rodríguez, "A weighted voting framework for classifiers ensembles," *Knowl. Inf. Syst.*, vol. 38, no. 2, pp. 259–275, 2014.

[10] M. Wozniak and K. Jackowski, "Some remarks on chosen methods of classifier fusion based on weighted voting." in *HAIS*, ser. Lecture Notes in Computer Science, E. Corchado, X. Wu, E. Oja, Ã. Herrero, and B. Baruque, Eds., vol. 5572. Springer, 2009, pp. 541–548.

[11] L. Sosnowski, A. Pietruszka, A. Krasuski, and A. Janusz, "A resemblance based approach for recognition of risks at a fire ground," in *Active Media Technology - 10th International Conference, AMT 2014, Warsaw, Poland, August 11-14, 2014. Proceedings*, 2014, pp. 559–570.

[12] A. Krasuski and A. Janusz, "Semantic tagging of heterogeneous data: Labeling fire & rescue incidents with threats," in *FedCSIS*, 2013, pp. 77–82.

[13] S. Staab and A. Maedche, "Knowledge Portals: Ontologies at Work," *AI Magazine*, vol. 22, no. 2, pp. 63–75, 2001.

[14] L. Sosnowski, "Applications of comparators in data processing systems," *Technical Transactions, Automatic Control*, 2014, to appear.

[15] D. W. Aha, Ed., *Lazy Learning*. Norwell, MA, USA: Kluwer Academic Publishers, 1997.

[16] T. M. Mitchell, *Machine Learning*, 1st ed. New York, NY, USA: McGraw-Hill, Inc., 1997.

[17] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection." Morgan Kaufmann, 1995, pp. 1137–1143.

[18] P. Faliszewski, E. Hemaspaandra, and L. A. Hemaspaandra, "Using complexity to protect elections." *Commun. ACM*, vol. 53, no. 11, pp. 74–82, 2010.

[19] E. Baharad and Z. Neeman, "The asymptotic strategy proofness of scoring and condorcet consistent rules," *Review of economic design : RED.*, 2002.

[20] X. Lin, S. Yacoub, J. Burns, and S. Simske, "Performance analysis of pattern classifier combination by plurality voting," *Pattern Recogn. Lett.*, vol. 24, no. 12, pp. 1959–1969, Aug. 2003.

[21] M. Barbie, C. Puppe, and A. Tasnadi, *Non-manipulable domains for the borda count*, ser. Bonn econ discussion papers, 2003, no. 13.

[22] D. G. Saari and V. R. Merlin, "The copeland method i; relationships and the dictionary," *Economic Theory*, vol. 8, 1996.

[23] S. J. Brams and P. C. Fishburn, "Going from theory to practice: the mixed success of approval voting." *Social Choice and Welfare*, vol. 25, no. 2-3, pp. 457–474, 2005.

[24] M. Balinski and R. Laraki, "A theory of measuring, electing, and ranking," *Proceedings of the National Academy of Sciences*, vol. 104, no. 21, pp. 8720–8725, May 2007.

[25] G. Levitin and A. Lisnianski, "Reliability optimization for weighted voting system." *Rel. Eng. & Sys. Safety*, vol. 71, no. 2, pp. 131–138, 2001.

[26] D. Whitley, "A genetic algorithm tutorial," *Statistics and Computing*, vol. 4, no. 2, pp. 65–85, Jun. 1994.

[27] M. Wozniak, M. Graña, and E. Corchado, "A survey of multiple classifier systems as hybrid systems," *Inf. Fusion*, vol. 16, pp. 3–17, Mar. 2014.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 145–153
DOI: 10.15439/2014F491
ACSIS, Vol. 2

MITC: An Intention-Based Model for Cooperative Resolution of Traffic Conflicts

Alejandro Triana Castañeda
Computer Science and System Engineering
Pontifical Javeriana University
Bogotá, Colombia
f.triana@javeriana.edu.co

Enrique González Guerrero
Computer Science and System Engineering
Pontifical Javeriana University
Bogotá, Colombia
egonzal@javeriana.edu.co

***Abstract*—Urban traffic problems have become a quotidian problem that affects many cities in the world. This problem, caused by the exponential increase of vehicles, leads to the appearance of different complications such as environmental pollution, accidents and slow mobility. This work formulates MITC, a model of cooperation focused to conflict resolution for the traffic agents, considering explicit communication of their intentions, allowing them to adjust their decisions intelligently, so as to reduce the conflicts and mitigate traffic congestion.**

***Keywords*—Intelligent Traffic Systems; Conflict Resolution; Game Theory; Multiagent Systems**

I. Introduction

URBAN traffic problems have become an everyday problem that affects many cities in the world. The total amount of vehicles in the world is calculated to be about 600 million, with an annual increase of 50 million [18]. Different factors such as the inefficiency in the infrastructure and its planning or a weak public awareness of traffic have increased the complexity of the problem [18]. Traffic problems can be divided into three kinds [10][6]: 1) Mobility issues, related to traveling time, 2) Safety issues, specially focused in preventing accidents, and 3) Environmental issues, generally caused by CO2 emissions.

Intelligent Transportation Systems (ITS) have emerged as an answer to traffic problems becoming one of the most interesting and promising alternatives within the scientific community [4][5][6]. The ITS aim to apply different artificial intelligence techniques such as Fuzzy Logic [3][7], Neuronal Network [15][16], Evolutionary Computation [14] and, in a more general way, the Agent and Multiagent System paradigm [9][12]. The works on ITS based on Multiagent Systems have covered a great quantity of fronts, among which these can be found: road traffic [7][13], urban traffic control (UTC) [1][2] [4][6][8], and decision support systems [7][17]. In all these solutions, the agents make decisions in an intelligent and cooperative way based in their knowledge of their surroundings.

This paper describes the Intentional Model for Cooperative Traffic (MITC for its name in Spanish). This solution is a traffic model based on Multiagent Systems in which agents cooperate explicitly communicating their intentions in order to solve traffic conflicts. The communication of intentions allows agents to adjust their decisions in an intelligent way to reduce the conflicts generated by the scarcity of resources (highway network) and non-compatible goals (antagonism between vehicles). The conflict resolution is inspired in the *benevolence* concept, namely the traffic agents with best traffic culture are prioritized. The second section introduces the agent's model and the proposed interaction mechanisms between them. The third section exposes the cooperative model, specifically the conflict resolution protocol. The four section describes the decision making system game theory based, which aims to reduce the traffic conflicts. The experiments that were carried out to evaluate this model are detailed in the last section. Finally, the conclusions are exposed from the perspective of reduction of conflicts between the traffic agents.

II. Traffic Multiagent System

This section describes the architecture of the Multiagent System. Initially, the design precepts are presented to describe the general characteristics of architecture. Follow, the characterization of the agents is defined in terms of their main goal. The final part of this section characterizes the agents' interactions and the existing means of communication.

A. Multiagent System's Basic Characteristics

Urban Traffic systems are highly complex, inherently distributed and have to deal with limited infrastructural resources. Due to these restrictions, the proposed Multiagent model exhibits the following characteristics:

1. Focused in Congestion Problems: its components and its relations aim to lower the conflicts among the traffic agents.
2. Highly Concurrent: it supports the great number of interactions among the agents, which are usually simultaneous.
3. Robust: it controls the handling of exceptional situations such as the damaging of sensors and traffic lights, among others.
4. Scalable: it allows the deployment in cities of different size and complexity.

B. Agents

This work proposes a model with five agents is describing in the Table 1. Each of the agents of the system are characterized in terms of their main goal, namely their principal function

978-83-60810-58-3/$25.00

inside the system. Likewise, each agent has an alias for quickly reference in the document.

TABLE I. SYSTEM AGENTS

Name	*Main Goal*
Traffic Intersection Agent – TIA	Controls vehicles in crossing intersections; for instance traffic lights.
Traffic Sensor Agent – TSA	Provides traffic information and generates of metrics of vehicle flow performance in a vehicle segment. The vehicle segment refers to the structure proposed in a Linear Based System (LBS).
Driver Control Agent – DA	Controls the motion of a vehicle going from a origin point to a destination point in the shortest possible time.
Traffic Area Monitor Agent – TAMA	Delivers information concerning a determinate on very large traffic area.
Traffic Supervisor Agent – TSUA	Supervision, support and control of the decisions of human controllers.

Accordingly, Fig. 1 illustrates the agent interactions in the proposed model. These interactions involve the existence of last generation technologies such as the detection of pedestrian flow, the presence of sensors (such as GPS), among others. However, some of the mentioned technologies are optional (for example the sensor for pedestrian flow), if available it allows a higher efficiency for the proposed model.

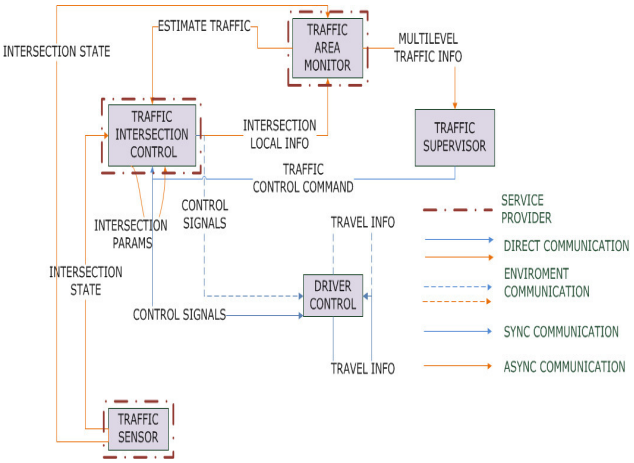


Fig. 1 Agent Interaction

Agent interactions can have different communication mechanisms, as follows:

1. Service Provider: It refers to an agent that provides an specific type of information to different agents. This information can be supplied in two ways:
   - Agents can subscribe to the Service Provider and receive the information asynchronously (Async Communication), whenever it is available.
   - Agents carry out demands to obtain information synchronously (Sync Communication).
2. Direct Communication: direct and explicit communication, usually through mechanisms like flashing headlights or other non-conventional ones, such as wireless networks.
3. Environmental Communication: Indirect communication across the environment. For instance, the horn or the turn signal lever.

III. CONFLICT RESOLUTION

In this section, the protocol of the MITC conflict resolution mechanism is described. Firstly, the concept and the categorization of the intentions that are used in this model are defined. Secondly, the intentions for each of the traffic agents are described. Finally, the characterizations of the traffic conflicts together with the proposed resolution protocol are introduced.

A. Definition of Intentions

This paper defines an intention as the goals that a traffic agent can have. These goals can follow a hierarchal and recursive classification, as follows:

1. Global Purpose Intentions: corresponds to the global aim of the agent.
2. Deliberative Intentions: refers to those intentions that are subject to the sequence of actions included in the plan of the agent. For example, the driver, according to his knowledge of traffic and exogenous information (news, weather forecast etc.), selects a path that includes several routes to go from his point of origin to his destination point.
3. Immediate Intentions: real-time actions carried out according to nearby traffic conditions. These intentions are motivated by the environment, the agent manage them in a reactive fashion; for instance, a traffic accident or a blockage due to adverse weather conditions.

Accordingly, consider a Multiagent System with N agents and $1 \le i \le N$.

- For every agent a_i one Global Intention G_i exists.
- G_i is achieved by a sequence of Deliberative Intentions included in a Plan $Pi = (P_{i1}, P_{i2}, \cdots P_{iP})$ of size p and $1 \le p \le P$.
- P_{ip} is carried out by a sequence of T Immediate Intentions $(I_{i1}, I_{i2,} \ldots It_{iT})$ and $1 \le t \le T$.

In this sense, the previous definition for the intentions can be to apply for the traffic agents´ model as shown in TABLE

II. These definitions let classify the traffic agents like: 1.) Expressive: agents can communicate the intentions for conflict resolution (TIA y DA are expressive agents) and 2.) Support Agents: agents provide information to use in conflict resolution.

TABLE I. TRAFFIC INTENTION AGENTS

Agent	*Global Purpose Intentions*	*Deliberative Intentions*	*Immediate Intentions*
TSA	Obtain traffic information and give rise to measures that can determine its performance.		
TIA	Mitigate vehicle time delay.	Control parameters adjustments according to historical acquisition.	High-beam switch
DA	Going from point A to point B in the least possible time.	Travel route selection.	Right or left turn.
			Move forward
			Brake
			Accelerate
			Change of lane
TAMA	Deliver multilevel information of a determined traffic area.		
TSUA	Support human controlers' decisions	Establish control rule per period.	

B.

C. Conflict Resolution Protocol

As previously mentioned, traffic conflicts are framed within road infrastructure shortages and agents' antagonist goals. Such conflicts happen in a defined geographic area (e.g. an intersection) and have a limited time duration. Accordingly, in order to solve conflicts the MITC model proposes the following:

1. A conflict has a scope CS denominated *conflict set*. The scope refers to the set of agents that intervene in the conflict, that is $CS=(a_1, a_2, a_k \cdots a_M)$ with $1 \leq k \leq M$.

2. Every agent a_k has a credit c_k. The credit represents the accumulated benefit that an agent has received when a conflict is solved to his favor.

3. There is an agent initiator of the a_i conflict protocol (an agent initiator is any agent traffic that can express its intentions), who communicates an immediate intention I_{it} of the set of available immediate intentions, that is $I_{it} \in (I_{i1}, I_{i2}, I_{i3} \cdots I_{iT})$ with $0 < t \leq T$ and $a_i \in CS$.

4. For every agent a_i, a possibility function $f_{pos}(I_{it})$ exists, which, given the I_{it} intention, evaluates the possibility of causing a conflict. The f_{pos} function complies with the following characteristics:

   - It is defined within the range [0,1]. Values close to 1 present a higher possibility of the intention causing a conflict.
   - If the value of the function f_{pos} exceeds a predefined threshold, the dialogue to prevent the conflict is initiated.

5. A CS *conflict set* has an a_m mediator agent associated to it, where $a_m \in CS$. The mediator agent is a virtual agent that emerges for to arbitrate the conflict resolution.

6. Each one of the a_k agents included in the conflict set generates an bid value b_k, calculated by a function f_{bid} such that $f_{bid}(I_{kt}) = b_k$.

7. For each a_k agent there is an associated unit value u_k obtained as the result in the conflict resolution process.

8. Every conflict has a unique identifier t_m. Every message that belongs to the conflict resolution dialogue has to include the identifier associated to it.

Taking into account these definitions and conditions, the proposed conflict resolution protocol is presented in the Fig. 2 and its formulation include the following steps or phases:

1. If the possibility function $f_{pos}(I_{it})$ of an agent a_i exceeds the *U* threshold, it creates a mediating agent a_m. The initiating agent sends a *conflict init* message to the mediating agent attaching the intention I_{it}, the offer b_i, the accumulated credit c_i, and the identifiers of the agents within the *conflict set* CS.

2. The mediating agent a_m forwards the message *conflict init* to all of the concerned agents a_k (with $k \neq i \ and \ k \neq m$) within the *conflict set* CS.

3. The agents a_k receive the request of a dialogue initiation for conflict resolution and answer with a message c*onflict response,* including their accumulated credit c_k, and their bid b_k. Notice that some agents may not respond to the petition of Conflict Resolution because of errors inherent to the communication channel.

4. The mediating agent a_m calculates the utility u_k for the agent a_i and for each agent a_k. Likewise, a_m sends the message *conflict result* announcing the utility u_i to agent a_i and the utility u_k for each agent a_k.

5. Finally, the initiating agent sends the message *conflict ACK* confirming the implementation of the intention.

In this sense, due to the fact that an agent can be involved in different conflicts simultaneously, the following considerations concerning the concurrency issues must be taken into account:

- When an agent a_i initiates a conflict dialogue, or an agent a_k receives a resolution request, he blocks his *availability* to participate in any other conflict resolution dialogue. This guarantees that an agent can only participate in one dialogue of conflict resolution at the same time.
- The participation of an agent a_i in a conflict is temporary and delimited in time. When time expires, the agent activates his *availability* in order to participate in any other resolution dialogue.
- For every agent a_i, $a_k \in CS$ two queues of handling messages exist:
  - One queue of incoming messages Qin, which stores the initial resolution messages. Each message received by the agent is stored in the queue using t_m as an identifier.
  - One queue to handle the events of a respective conflict $Qman$. This queue handles the messages for only one conflict simultaneously.
- When the conflict dialogue ends, the agents activate their *availability* to participate in any other resolution dialogue.

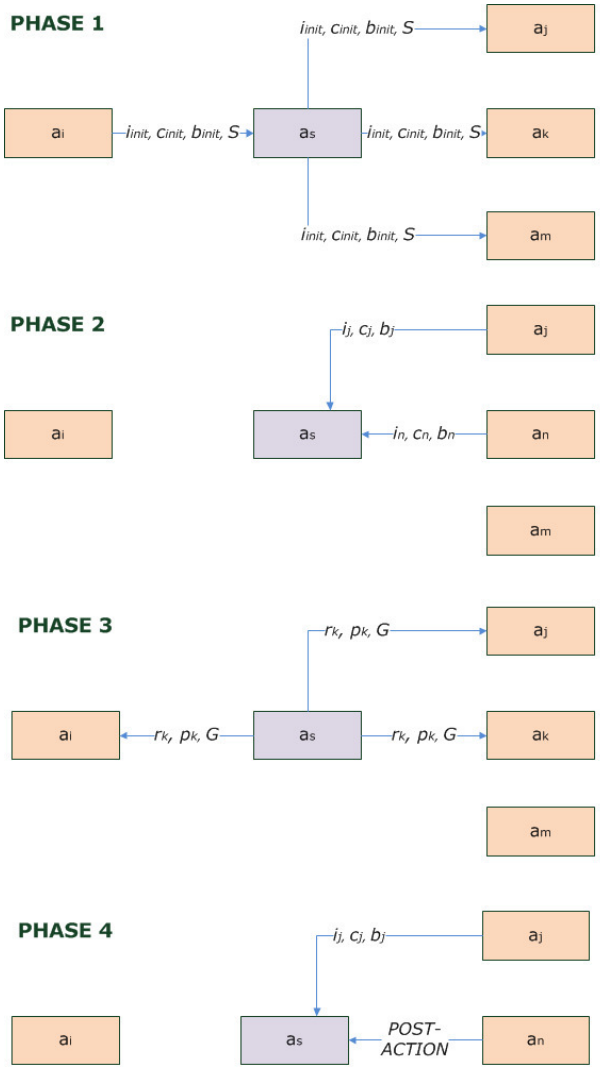


Fig. 2 Resolution Protocol Steps

IV. Decision Model by Game Theory

This section describes the model of decision making for traffic agents. First, the traffic conflicts are characterized as a game model. Then, the formalization of the game for a traffic conflict is carried out. Finally, the steps for the solution of the conflict are described in terms of the approach of a game in a normal-form.

A. General Assumptions for Traffic Conflict as a Non-Cooperative Model of Game Theory

MITC proposes a model based on Game Theory to find utility values u_k for agents a_k that are involved in a *conflict set* CS such that $CS = (a_1, a_2, a_k \cdots a_M)$ and $1 \leq k \leq M$.

Traffic conflicts can be described as a model of Game Theory according to the following considerations:

1. These are games of both complete information (the players know *completely* the strategy of the others, since they communicate their intentions) and perfect information (there is no uncertainty regarding the decisions of the agents).

2. These are games of simultaneous interaction. In other words, each conflict is independent of previous events that happen among the agents.

3. Players: every traffic agent $a_k \in CS$.

4. Actions: the vector $IC_k = (I_{kt}, -I_{kt})$, corresponds to one agent a_k where I_{kt} corresponds to the agent's immediate intention and $-I_{kt}$ corresponds to the non-carrying out of such intention. Be noted that this chapter refers to the terms of action and immediate intention indistinctively.

5. Utility Value: corresponds to the utility value u_k obtained by agent a_k.

B. Traffic Conflict as a Normal-form Game

One traffic conflict can be characterized as a Normal-form Game, as a tuple (CS, IC, u) where:

1. CS (*conflict set*) is the finite set of agents a_k that take part in the game.

2. $IC = (IC_1, IC_2, IC_3 \cdots IC_M)$ is a vector such that $IC_{kt} = (I_{kt}, -I_{kt}), \quad with \quad 0 < k \leq M$. The intentions IC_{kt} correspond to the set of available actions for agent a_k and is denominated *action profile* for agent a_k.

3. $u = u_1, u_2 \ldots\ldots\ldots\ldots u_M$, where the utility u_k is defined in terms of accumulated credit c_k and bid value b_k.

A more intuitive way to represent a game in a normal-form is the bimatrix mechanism. In the bimatrix, the cells contain the utility of each agent for the possible combinations of strategies. Each cell contains two numbers (and therefore the origin of its name), which represent the utility of the agents in such strategy. The figure 3 illustrates the bimatrix for a conflict of change of lanes between two vehicle agents (DA – Driver Agents) $a_1 and\ a_2$.

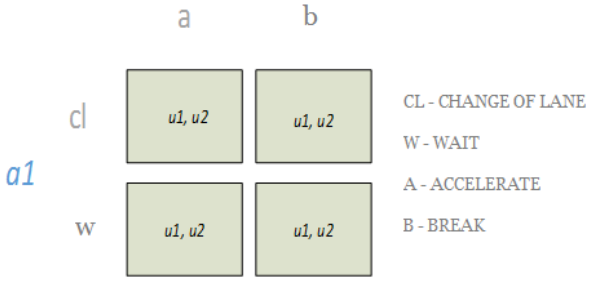


Fig. 3 Bimatrix of a Traffic Game

In general terms, for a traffic conflict with *M* agents, the actions $I^*_{1t}, I^*_{2t} \cdots I^*_{Mt}$ form a Nash Equilibrium if for each agent a_k, the immediate action I^*_{kt} is the best action that can be taken by agent a_k for the actions of the other $k-1$ players $I^*_{1t}, I^*_{2t}, I^*_{|k-1|t}, I^*_{|k+1|t} \cdots I^*_{Mt}$. That is:

$$u(I^*_{1t}, I^*_{2t}, I^*_{|k-1|t}, I^*_{kt}, I^*_{|k+1|t} \cdots I^*_{Mt}) > u(I^*_{1t}, I^*_{2t}, I^*_{|k-1|t}, I_{kt} I^*_{|k+1|t} \cdots I^*_{Mt})$$

For each possible action I_{kt} in IC_{kt}, I_{kt} is a solution for:

$$\max(I^*_{1t}, I^*_{2t}, I^*_{|k-1|t}, I^*_{kt}, I^*_{|k+1|t} \cdots I^*_{Mt})$$

C. Strategy of Nash Equilibrium Calculation

During steps 3 and 4 of the Conflict Resolution Protocol the mediating agent receives the offers of the agents in the *conflict set* and must solve the conflict calculating the corresponding Nash equilibrium. The strategy for this calculation given an immediate intention, consists of the following steps:

1. The bids $b_1, b_2, b_3, b_4 \cdots b_M$ are obtained. The bid function $f_{bid}(I_{kt}) = b_k$ for an agent a_k follows these criteria:

   a. Each intention I_{kt} has one base bid value associated w_{kt} with $w_{kt} \in Z$. The base bid value can be seen as the importance of the intention within the system's context (e.g. an ambulance can have a greater importance in its intentions than private cars).

   b. Each agent a_k has a benevolence coefficient v_k with $v_k \in Z$. For more

information about the benevolence calculation see section 5.3.

c. Each bid is attenuated by the accumulated credit c_k. The higher the accumulated credit value, the lesser the bid value b_k is. This approach allows controlling those agents that intend to abuse of their benevolence to accumulate excessive credit.

Accordingly, the function for the bid is calculated as:

$$b_k = f_{bid}(I_{kt}) = (w_{kt} * v_k)\left(\frac{1}{c_k}\right) \quad para \quad c_k > 0$$

2. The mediating agent a_m calculates the game bimatrix T_{ik} for every couple of agents a_i (initiating agent) and a_k with $i \neq k$. The calculation of the utilities for the bimax is based on the following conditions:

- If the intention I_{il} is chosen and not I_{kl}, then $u_i = b_i - b_k, \; u_k = b_k + b_i$
- If the intention I_{kl} is chosen and not I_{il}, then $u_k = b_k - b_i, \quad u_i = b_i + b_k$
- If I_{il} and I_{kl} are chosen, then $u_i = -b_i, \quad u_k = -b_k$
- If I_{il} and I_{kl} are not chosen, then $u_i = 0, \; u_k = 0$

1. For each matrix T_{ik} the corresponding Nash equilibrium is calculated as Eq_{ik}.

2. All equilibriums Eq_{ik} are obtained, where the intention I_{it} of the initiating agent a_i is selected. Afterwards, the equilibrium Eq_{ik} is selected as the one that produces the maximum utility u_i.

3. The agents a_i and a_k modify their accumulated credit as follows:

$$c_i = c_i + u_i, \quad c_k = c_k + u_k$$

The calculation of Nash equilibrium includes certain characteristics to be taken into account:

- There can be situations in which there is no equilibrium, in which case there is no conflict resolution.
- The state for $u_i = -b_i \; and \; u_k = -b_k$ can never correspond to an equilibrium; namely, the state in which both agents comply with their intention simultaneously is omitted.

V. RESULTS

This chapter describes the experiments carried out to validate the proposed MITC model. First the scenario design for the simulation is presented in order to, later on, present the results of the conducted experiments.

A. Simulation Scenario

The validation experiments use the scenario of Crossing on a Slow Lane, which is described in terms of figure 4 as follows:

- From point **1** to point **2** there is a distance of 1400 meters.
- From point **1** to point **3** there is a distance of 1408.93 meters.
- The scenario is comprised of six roads R1, R2, R3, R4, R5, y R6.
- R1 and R5 have a length of 800 meters with two lanes in each one of them.

- R2 has a length of 200 meters and 5 lanes.
- R4 has a length of 408.1 meters and 1 lane. It has a connection at the final point of R2.
- R3 and R6 have a length of 400 meters.

B. Results

During the first part of the protocol, 5 of the experiments were carried out using the native behavior included in the simulator and 5 experiments were carried out using the cooperative model with the following characteristics:

- 10 repetitions were carried out for each experiment.
- The flow of vehicles was varied (HIGH, MEDIUM-HIGH, MEDIUM-LOW, LOW) with a duration of the experiment of 3600 seconds and a 0.2 time step for the simulation[1].
- The variables Resolution Activation and Benevolence were established in HIGH.

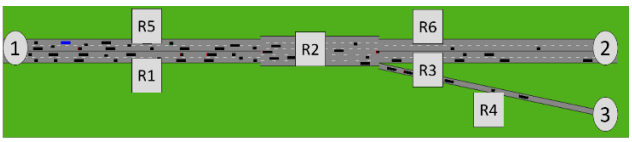


Fig. 4 Simulation Scenario

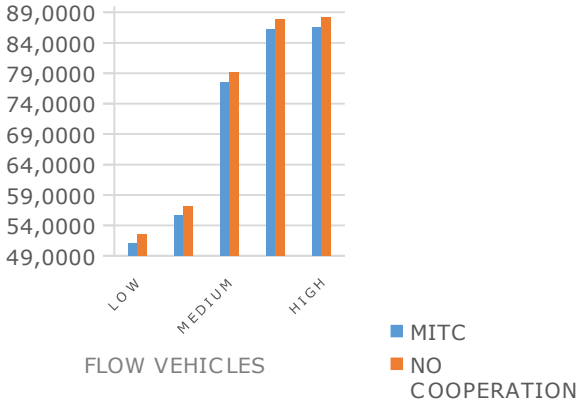


Fig. 5 Time Comparison

[1] The duration constitutes the time that the simulation lasts until it reaches 3600 (it represents one hour in MovSim) with an increasing value by cycle of 0.2. That is to say, in this case 18000 cycles of simulation would be carried out.

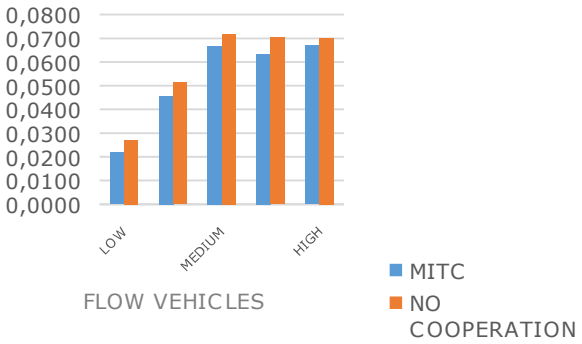


Fig. 6 Consumption Comparison

According to figures 5 and 6, the MITC model provides an improvement of travel time of 2.17% per vehicle and of 8.8% for energy consumption. At first glance these indicators pose an improvement in a local scenario of 1400 meters and, taking it to a scenario with higher dimensions (e.g. a metropolitan area), it may represent great benefits.

The second part of the experiments was focused in the analysis of the Activation Resolution and Benevolence variables. These variables determine the behavior of the model towards exceptional situations, such as infrastructure communication errors and traffic agents with lack of collaborative culture.

As can be observed in Fig. 7 and Fig. 8, the MITC model has a similar behavior in comparison with the results observed in the basic behavior (the simulation without the cooperative model included), when the Activation of the Resolution has values LOW and MEDIUM. This means that the model can cohabit, without negatively affecting the performance of the system, in mixed scenarios that not only include cooperation but also indifference.

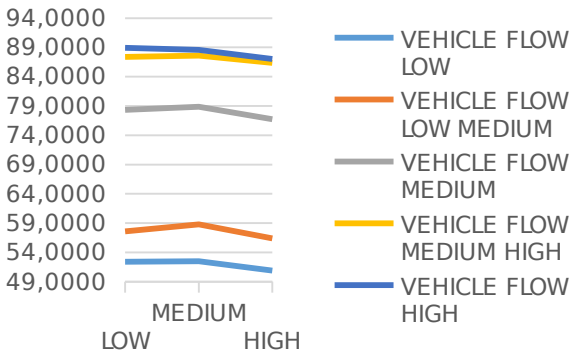


Fig. 7 Results Comparison Activation Resolution – Travel Time

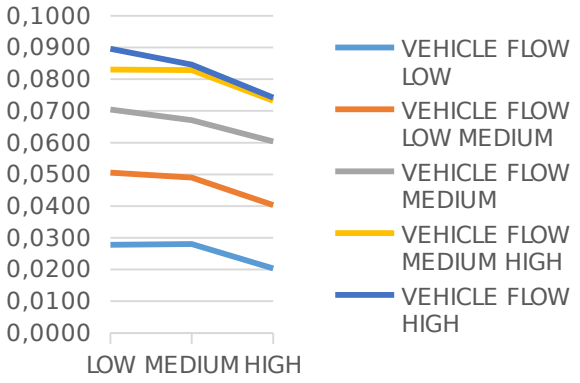


Fig. 8 Results Comparison Activation Resolution – Consumption

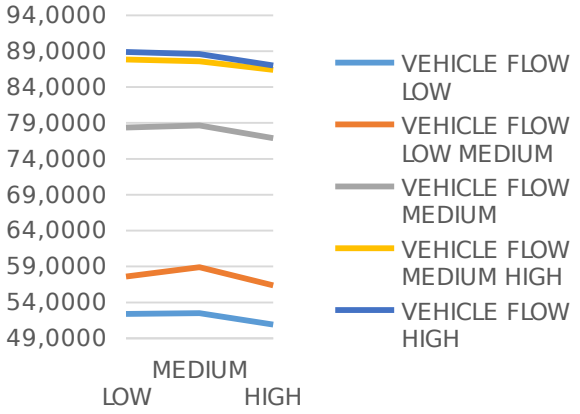


Fig. 9 Comparison Results Benevolence – Travel Time

Likewise, the results for the Benevolence variable can be observed in Fig. 9 and 10. Only when there is higher benevolence (specifically, for HIGH values), traffic conditions improve significantly. This highlights the importance of traffic culture in a city.

VI. Conclusions

The design and construction of an Intention-Based Model for Cooperative Resolution of Traffic Conflicts offers a different approach in order to solve congestion problems. MITC allows explicitly expressing the cooperative mechanisms between traffic agents in scenarios where conflicts are presented, according to shared resources and according to conflicting goals. In this sense, MITC provides a set of advantages:

1. The Multiagent Model is designed to tolerate different requirements given by an Urban Traffic System, such as concurrency, scalability and its complexity. In computational terms, MITC allows approaching traffic problems in a distributed way, which brings important advantages in terms of availability and fault tolerance.
2. The Multiagent Model poses a Conflict Resolution Protocol that supports the essential characteristics such as concurrency and its temporality.
3. The Multiagent Model covers transversally every aspect of a Traffic System, including control areas, administration and supervision.
4. The definition of the concept of Intention and its three level hierarchies allows modeling the characteristics of traffic agents in a manner that is natural and closer to reality.
5. The Decision Making Model based in Game Theory guarantees the solution of conflict in a rational and balanced way. Additionally, its inspiration in the concept of benevolence allows analyzing essential aspects such as public traffic conscience.

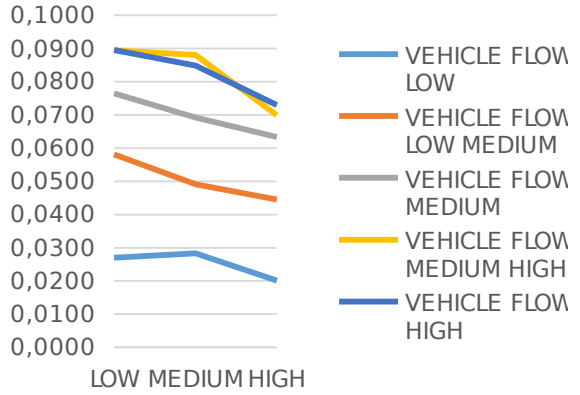


Fig. 10 Comparison Results Benevolence – Consumption

This work opens the door to a great number of applications. For instance, in the development of campaigns for intelligent traffic and traffic culture, it provides formal and measurable elements in terms of travelling time. Likewise, it would be very helpful to include experiments that determine the decrease of accidental rates through MITC on behalf of incorporated elements of road safety that give as a result a pedagogic frame for urban traffic. In this sense and although in Latin American cities traffic networks are far from implementing technologies such as inter-vehicular nets or smart vehicles, MITC is an adaptable model that can be implemented partially. Therefore, designing a System of Urban Traffic Control becomes interesting, posing options for intentional traffic lights and its respective strategies for conflict solving. In a similar way, it is possible to generate a system of traffic recommendations in real time that can

assist users while they drive their vehicles. Likewise, the model can be extended with a realistic Deliberative Intentions implementation, where the model express a coherent traceability across of intention hierarchy.

MITC is a model with a significant impact when implemented in real traffic scenarios, since it can mitigate some of the factors that have negative effects in the quality of life of people. For example, it can decrease atmospheric pollution, given that vehicles would spend less time on the streets, and it can decrease environmental noise, as it prevents the emission of sound signals of vehicles, as conflicts can be solved automatically.

REFERENCES

[1] J. L. Adler and V. J. Blue. A cooperative multi-agent transportation management and route guidance system. Transportation Research Part C-emerging Technologies, 10(5-6):433–454, October 2002.

[2] J. L. Adler, G. Satapathy, V. Manikonda, B. Bowles, and V. J. Blue. A multi-agent approach to cooperative traffic management and route guidance. Transportation Research Part B-methodological, 39(4):297–318, May 2005.

[3] T. Akiyama and M. Okushima. Advanced fuzzy traffic controller for urban expressways. International Journal of Innovative Computing Information and Control, 2(2):339–355, April 2006.

[4] P. G. Balaji, X. German, and D. Srinivasan. Urban traffic signal control using reinforcement learning agents. Iet Intelligent Transport Systems, 4(3):177–188, September 2010.

[5] M. Bielli, G. Ambrosino, and M. Boero. Artificial intelligence applications to traffic engineering. Vsp, 1994.

[6] J. C. Burguillo-Rial, P. S. Rodriguez-Hernandez, E. CostaMontenegro, and F. Gil-Castineira. History-based selforganizing traffic lights. Computing and Informatics, 28(2):157–168, 2009.

[7] B. Chen and H. H. Cheng. A review of the applications of agent technology in traffic and transportation systems. Ieee Transactions On Intelligent Transportation Systems, 11(2):485–497, June 2010.

[8] R. S. Chen, D. K. Chen, and S. Y. Lin. Actam: Cooperative multi-agent system architecture for urban traffic signal control. Ieice Transactions On Information and Systems, E88D(1):119–126, January 2005.

[9] M. C. Choy, D. Srinivasan, and R. L. Cheu. Cooperative, hybrid agent architecture for real-time traffic signal control. Ieee Transactions On Systems Man and Cybernetics Part A-systems and Humans, 33(5):597–607, September
2003.

[10] R. Fernandes. A BDI-based approach for assessment of drivers decision-making in commuter. PhD thesis, Universidade Federal Do Rio Grande Do Sul, nov 2002.

[11] Enrique Gonzales and Cesar Bustacara. Desarrollo de Aplicaciones Basadas en Sistemas Multiagentes. 2007.

[12] J. Z. Hernandez, S. Ossowski, and A. Garcia-Serrano. Multiagent architectures for intelligent traffic management systems. Transportation Research Part C-emerging Technologies, 10(5-6):473–506, October 2002.

[13] J. Hillenbrand, A. M. Spieker, and K. Kroschel. A multilevel collision mitigation approach - its situation as sessment, decision making, and performance tradeoffs. Ieee Transactions On Intelligent Transportation Systems, 7(4):528–540, December 2006.

[14] T. Ma and B. Abdulhai. Genetic algorithm-based optimization approach and generic tool for calibrating traffic microscopic simulation parameters. *Intelligent Transportation Systems and Vehicle-highway Automation 2002: Highway Operations, Capacity, and Traffic Control*, (1800):6–15, 2002.

[15] D. Srinivasan, M. C. Choy, and R. L. Cheu. Neural networks for real-time traffic signal control. *Ieee Transactions On Intelligent Transportation Systems*, 7(3):261–272, September 2006.

[16] H. B. Yin, S. C. Wong, J. M. Xu, and C. K. Wong. Urban traffic flow prediction using a fuzzy-neural approach rid a-7258-2008. *Transportation Research Part C-emerging Technologies*, 10(2):85–98, April 2002.

[17] N. Zhang, F. Y. Wang, F. H. Zhu, D. B. Zhao, and S. M. Tang. Dynacas: Computational experiments and decision support for its. *Ieee Intelligent Systems*, 23(6):19–23, November 2008.

[18] D. Zhao, Y. Dai, and Z. Zhang. Computational intelligence in urban traffic signal control: A survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, PP(99):1 –10, 2011

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 155–161
DOI: 10.15439/2014F377
ACSIS, Vol. 2

Fully Informed Swarm Optimization Algorithms: Basic Concepts, Variants and Experimental Evaluation

Szymon Łukasik[1,2]

[1]Department of Automatic Control and IT
Cracow University of Technology
ul. Warszawska 24, 31-155 Krakow, Poland
Email: {szymonl,pkowal}@pk.edu.pl

Piotr A. Kowalski[1,2]

[2]Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
Email: {slukasik,pakowal}@ibspan.waw.pl

***Abstract*—Particle swarm optimization constitutes currently one of the most important nature-inspired metaheuristics, used successfully for both combinatorial and continuous problems. Its popularity has stimulated the emergence of various variants of swarm-inspired techniques, based in part on the concept of pairwise communication of numerous swarm members solving optimization problem in hand. This paper overviews some examples of such techniques, namely Fully Informed Particle Swarm Optimization (FIPSO), Firefly Algorithm (FA) and Glowworm Swarm Optimization (GSO). It underlines similarities and differences among them and studies their practical features. Performance of those algorithms is also evaluated over a set of benchmark instances. Finally, some concluding remarks regarding the choice of suitable problem-oriented optimization technique along with areas of possible improvements are given as well.**

I. Introduction

PARTICLE Swarm Optimization introduced by Kennedy, Eberhart and Shi in 1995 [1] is at the moment one of the most noteworthy nature-inspired metaheuristics used for variety of tasks, both in science and engineering. It was induced by the observation of flocking and schooling patterns of birds and fish. An idea to represent each solution of the optimization problem at-hand as a member of the virtual swarm – communicating with others and modifying its position under the influence of best individuals – proved to be extremely successful. The degree of this success can be represented by the significant amount of contributions employing PSO in real-world problems e.g. in data analysis [2], resource allocation [3] etc. It can be also quantified through a number of related algorithms, based on the idea of intelligent swarms. One of recent examples of such include: Quantum-behaved Particle Swarm Optimization [4] and Multi-Swarm PSO [5].

The general goal of continuous optimization is to find x^* which satisfies:

$$f(x^*) = \min_{x \in S} f(x), \tag{1}$$

where $S \subset R^N$, and $f(x)$ constitutes solution's x cost function value. Therefore actual task of the optimizer is to find argument minimizing f.

Initial PSO algorithm's behavior was built on the assumption that each individual member of the swarm, i.e. solution of the optimization problem (1), changes its velocity vector in the consecutive algorithm's iteration as a result of the influence of two specific solutions: the best one found so far by the swarm and the top solution identified by this individual. Fully Informed Particle Swarm Optimization (FIPSO) presented first by Mendes, Kennedy and Neves [6] constitute a modification of this approach. In the most general variant of FIPSO velocity update is constructed using weighted average position of all swarm members. Creators of the algorithm considered however alternative communication topologies, e.g. ring or cluster, as well as different schemes of assigning weights to prioritize individuals' inputs. Firefly Algorithm (FA) created by Xin She Yang in 2008 is constructed on similar assumptions [7]. The position of swarm member x_m within feasible solution space S is determined by all other individuals' fitness – better solutions will attract those which are worse, in the sense of selected cost function f value [8]. Glowworm Swarm Optimization (GSO) developed by Krishnanand and Ghose [9] exhibits similar behavior however swarm member is attracted here by its better-performing neighbors found only within given radius. Considering those similarities FA and GSO like FIPSO can be perceived as most representative members of broader family of techniques named here Fully Informed Particle Swarm algorithms. It can be characterized by building solution space exploration process on exchanging information between swarm members regarding local fitness landscape and modifying their position accordingly. For other examples of such methods one can refer to [10], [11], [12].

The goal of this contribution is to provide synthetic comparative perspective on major Fully Informed Particle Swarm algorithms, introduced above, both on conceptual and performance-based grounds. It is organized as follows. First the description of all techniques studied here in their basic variants is provided, along with similarities between them. It also contains brief discussion of selected technical aspects, examples of applications and possible modifications. Then

978-83-60810-58-3/$25.00

the results of performed comparative experimental studies are given, both in the context of optimization performance, algorithms' convergence and computational demands. Finally general remarks concerning the choice of suitable problem-oriented optimization techniques and planned further studies are under consideration.

II. Fully-Informed Swarm Algorithms

First let us introduce the notation which will be used in the following subsections. To solve the optimization problem (1) the swarm, consisting of M members will be used. It will be represented by a set of N-dimensional vectors – equivalent to individuals' positions – within the iteration k denoted by:

$$x_1(k), x_2(k), ..., x_M(k). \tag{2}$$

Euclidean distance between two swarm members, indexed p and q is denoted here by $d(x_p, x_q)$. The best position found by given swarm member m prior to iteration k is given by $x_m(k)^*$ with cost function value $f(x_m(k)^*)$. At the same time:

$$x(k)^* = \arg \min_{m=1,...,M} f(x_m(k)), \tag{3}$$

corresponds to the best solution found by the algorithm in its k iterations, with $f(x(k)^*)$ representing its related cost function value. Swarm optimization algorithms, based in particular on PSO paradigm, employ frequently a concept of individual's m velocity, denoted here in iteration k as $v_m(k)$. It is used to update particles' positions and can be initialized randomly within given bounds.

The following part of the paper will provide comprehensive description of Fully Informed Particle Swarm algorithms, referring to the notation given above.

A. Fully-Informed Particle Swarm Optimization (FIPSO)

Fully Informed Particle Swarm Optimization constitutes one of heuristic algorithms derived from the basic PSO paradigm. The idea of using information from a group of particle's K_m neighbors, rather than just the best one – as in traditional canonical PSO – was first proposed by Suganthan in 1999 [13]. It was also included in complete Fully Informed PSO procedure suggested by Mendes, Kennedy and Neves. In each iteration of the algorithm particle's position $x_m(k)$ is updated by moving it iteratively along vector $v_m(k)$ with the coordinates $n = 1, ..., N$ adjusted as follows:

$$v_{mn}(k+1) = \chi \left[v_{mn}(k) + \frac{1}{K_m} \sum_{j=1}^{K_m} U(0,\varphi)(x_{N_m(j)_n}(k)^* - x_{mn}(k)) \right], \tag{4}$$

with χ known as a constriction factor, whereas $U(0,p)$ corresponds to the uniformly distributed random number in $(0,p)$, φ constitutes an acceleration coefficient, and finally, $N_m(j)$ is a function which returns the index of j-th nearest neighbor of particle m. Complete FIPSO procedure was provided below in the form of pseudocode (Algorithm 1).

Algorithm 1 Fully Informed Particle Swarm Optimization algorithm

```
k ← 1 {initialization}
for m = 1 to M do
    Generate_Solution(x_m(k))
    Initialize_Velocity(v_m(0))
    f(x_m(0)*) ← ∞
end for
{main loop}
repeat
    {evaluate and update best solutions}
    for m = 1 to M do
        f(x_m(k)) ← Evaluate_quality(x_m(k))
        if f(x_m(k)) < f(x_m(k−1)*) then
            x_m(k)* ← x_m(k)
        else
            x_m(k)* ← x_m(k−1)*
        end if
        if f(x_m(k)) < f(x(k)*) then
            x(k)* ← x_m(k)
        else
            x(k)* ← x(k−1)*
        end if
    end for
    for m = 1 to M do
        for n = 1 to N do
            c_mn(k) ← 0
            for all K_m nearest neighbors (index p) of m do
                c_mn(k) ← c_mn(k) + (U(0,φ)(x_pn(k)* − x_mn(k))
            end for
            v_mn(k) ← χ * [v_mn(k−1) + 1/K_m * c_mn(k)]
            x_mn(k+1) ← x_mn(k) + v_mn(k)
        end for
    end for
    stop_condition ← Check_stop_condition()
    k ← k + 1
until stop_condition = false
return f(x(k)*), x(k)*, k
```

One of the findings of initial study on FIPSO was that increasing the size of the "informing" neighborhood seems to deteriorate the performance of the swarm. FIPSO with a fully connected topology, i.e., when each particle has all the particles in the swarm as its neighbors, shows a particularly bad performance in comparison with the one attained with other topologies, e.g. ring or square. In addition to that FIPSO convergence was thoroughly studied in [14]. Authors observe there that for highly connected topologies, the particles explore a region close to the centroid of the swarm. It may bring positive results for some specific functions however the algorithm in that case is prone to becoming trapped in local minima. The algorithm in the form introduced above was successfully applied for engineering problems like power

systems optimization [15]. It was also used as a starting point for other similar approaches [16], [17] as well as a component of hybrid algorithms [18]. Here we consider most general FIPSO with fully-connected particles to study its performance when referencing it to two other more recent approaches.

B. Firefly Algorithm (FA)

Firefly Algorithm developed by Xin-She Yang [7] is inspired by mechanisms of firefly communication via luminescent flashes. This swarm intelligence optimization technique is based on the assumption that solution of an optimization problem can be perceived as agent (firefly) which "glows" proportionally to its quality in a considered problem setting. Consequently each brighter firefly attracts its partners (regardless of their sex), which makes the search space being explored more efficiently [8].

Each firefly has its distinctive attractiveness β which implies how strong it attracts other members of the swarm. For attractiveness in FA an exponential function of the distance $r_j = d(x_m, x_j)$ to the chosen firefly j is used:

$$\beta = \beta_0 e^{-\gamma r_j} \tag{5}$$

where β_0 and γ are predetermined algorithm parameters: maximum attractiveness value and absorption coefficient, respectively. Every member of the swarm is also characterized by its light intensity I_m which can be directly expressed as a inverse of a cost function $f(x_m)$.

To effectively explore considered search space S it is assumed that each firefly m is changing its position iteratively taking into account two factors: attractiveness of other swarm members with higher light intensity i.e. $I_j > I_m, \forall j = 1, ...M, j \neq m$ – which is varying across distance – and a fixed random step vector $U(min, max)$. It should be noted as well that if no brighter firefly can be found only such randomized step is being used [8].

Algorithm 2 presents generic Firefly Algorithm which includes all aforementioned elements. For a recent overview of FA modifications, variants and applications one can refer to [19]. We employ here standard FA algorithm with uniform random number generator and scaling factor related to search space size S [8].

Algorithm 2 Firefly Algorithm

$k \leftarrow 1$ {initialization}
for $m = 1$ to M **do**
  Generate_Solution($x_m(k)$)
end for
$f(x(0)^*) \leftarrow \infty$
{main loop}
repeat
  {evaluate and update best solution}
  for $m = 1$ to M **do**
    $f(x_m(k)) \leftarrow$ Evaluate_quality($x_m(k)$)
    if $f(x_m(k)) < f(x(k-1)^*)$ **then**
      $x(k)^* \leftarrow x_m(k)$
    else
      $x(k)^* \leftarrow x(k-1)^*$
    end if
  end for
  for $m = 1$ to M **do**
    for $p = 1$ to M **do**
      if $f(x_m(k)) < f(x_p(k))$ **then**
        $r_p \leftarrow Calculate\_Distance(x_m(k), x_p(k))$
        $\beta \leftarrow \beta_0 e^{-\gamma r_p}$
        for $n = 1$ to N **do**
          $x_{mn}(k) \leftarrow (1-\beta)x_{mn}(k) + \beta x_{pn}(k) + U_n(min, max)$
        end for
      end if
    end for
  end for
  {best moves randomly}
  for $m = 1$ to M **do**
    if $Was\_Moved(x_m(k))$ = **false then**
      for $n = 1$ to N **do**
        $x_{mn}(k) \leftarrow x_{mn}(k) + U_n(min, max)$
      end for
    end if
  end for
  $stop\_condition \leftarrow Check\_stop\_condition()$
  $k \leftarrow k + 1$
until $stop\_condition$ = **false**
return $f(x(k)^*)$, $x(k)^*$, k

C. Glowworm Swarm Optimization (GSO)

Glowworm Swarm Optimization is another optimization strategy which was stimulated by the observation of fireflies' social behavior. In contrast to Firefly Algorithm agents in GSO depend only on information available in their strict neighborhood to make decisions [9]. What is more GSO uses an adaptive neighborhood range in order to successfully deal with multimodal functions landscapes. Both luciferin quantity $\iota_m(k)$, which predetermines the probability of individual's movement, and neighborhood radius $r_m(k)$ are updated on per-iteration basis. It is realized using the following formulas:

$$\iota_m(k) = (1-\rho)\iota_m(k-1) + \gamma f(x_m(k))^{-1}, \tag{6}$$

$$r_m(k+1) = \min\{r_s, \max\{0, r_m(k) + \beta(N_{set} - |N_m(k)|)\}\}, \tag{7}$$

with ρ representing luciferin decay parameter, γ constituting luciferin enhancement constant, r_s - maximum sensor range, N_{set} - parameter controlling number of neighbors and finally, $N_m(k)$ denoting a set of neighbors of $x_m(k)$ located within radius $r_m(k)$:

$$N_m(k) = \{x_j(k) : d(x_m(k), x_j(k)) < r_m(k) : \iota_m(k) < \iota_j(k)\}. \tag{8}$$

Probability of glowworm movement towards one of other individuals in the neighborhood $N_m(k)$ is proportional to its

luciferin quantity, related to the sum of luciferin values for all neighbors found in $N_m(k)$. It is denoted for all neighbors by a vector $p_m(k)$. The Algorithm 3 presents plain description of GSO procedure including most important technical details.

Algorithm 3 Glowworm Swarm Optimization algorithm

$k \leftarrow 1$ {initialization}
$f(x(k)^*) \leftarrow \infty$
for $m = 1$ to M **do**
  Generate_Solution($x_m(k)$)
  $f(x_m(k)) \leftarrow$ Evaluate_quality($x_m(k)$)
  $\iota_m(0) \leftarrow \iota_0$
  $r_m(0) \leftarrow r_0$
end for
{main loop}
repeat
  {update luciferin quantity}
  for $m = 1$ to M **do**
    $\iota_m(k) \leftarrow (1-\rho)\iota_m(k-1) + \gamma f(x_m(k))^{-1}$
  end for
  {move glowworms}
  for $m = 1$ to M **do**
    $N_m(k) \leftarrow Find\_Neighborhood(x_m(k))$
    {sum selection probalities for all p neighbors in $N_m(k)$}
    $P_{sum} \leftarrow sum(\iota_p(k) - \iota_m(k))$
    for all $x_j(k)$ in $(N_m(k))$ **do**
      $p_{mj}(k) \leftarrow (\iota_j(k) - \iota_m(k))/P_{sum}$
    end for
    $q \leftarrow Select\_neighbor\_index(p_m(k))$
    {move towards selected}
    $x_m(k+1) \leftarrow x_m(k) + s(x_q(k) - x_m(k))$
      $/(\|x_q(k) - x_m(k)\|)$
    $r_m(k+1) \leftarrow \min\{$
    $r_s, \max\{0, r_m(k) + \beta(N_{set} - |N_m(k)|)\}\}$
  end for
  for $m = 1$ to M **do**
    if $Was\_Moved(x_m(k)$ = **false then**
      $x_m(k+1) \leftarrow x_m(k)$
    end if
    $f(x_m(k)) \leftarrow$ Evaluate_quality($x_m(k)$)
    if $f(x_m(k)) < f(x(k)^*)$ **then**
      $x(k)^* \leftarrow x_m(k)$
    else
      $x(k)^* \leftarrow x(k-1)^*$
    end if
  end for
  $stop\_condition \leftarrow Check\_stop\_condition()$
  $k \leftarrow k+1$
until $stop\_condition$ = **false**
return $f(x(k)^*)$, $x(k)^*$, k

GSO like FA attracted much attention resulting in several contributions improving the general scheme of the algorithm [20], studying theoretical properties [21] or employing GSO for real-life problems [22]. Here, for comparative studies we utilize standard GSO developed by Krishnanand and Ghose.

D. Summary

Techniques covered in this Section are employing mutual information exchange between all members of the swarm or its selected groups (depending on precise variant and parameter values). It is used for modifying swarm member position (GSO and FA) or its velocity vector (FIPSO).

In case of FIPSO individual's movement is influenced by a set of best solutions obtained by other swarm members. For GSO and FA latest position of other individuals can be used, however it must be better than the one of solution currently under consideration.

Algorithms studied here employ a randomization component, either in the form of implicit randomized movement (FA), by random selection of informing agent (GSO) or determining strength of each neighbor's influence (FIPSO). All mechanisms tend to improve algorithms' abilities to escape local minima. In this aspect additional dynamics contained within FIPSO technique could be extremely beneficial.

Every technique studied here possesses significant number of parameters, with GSO being most parameter-rich and FIPSO parameter-free one. All required parameters are listed in Table I. For most of them some guidelines have been already worked out in the related contributions - they are listed in the table as well.

TABLE I
ALGORITHMS' PARAMETERS AND THEIR SUGGESTED VALUES

Algorithm	Parameter	Suggested value/range	Source
FIPSO	M	[20,50]	[23]
	χ	0.72984	[24]
	φ	4.1	[24]
	K_m	[3,5]	[6]
FA	M	[20,50]	[25]
	α	[0.1,0.2]	[25]
	β_0	1	[8]
	γ	[1,30]	[8], [25]
GSO	M	[10,500]	[21]
	ρ	0.4	[21]
	γ	0.6	[21]
	β	0.08	[21]
	N_{set}	5	[21]
	s	0.03	[21]
	ι_0	5	[21]
	r_s	use pilot runs	[21]
	r_0	$r_0 = r_s$	[21]

As for computational complexity of algorithms studied here it is in all cases significant. With regards to swarm size M and iteration number K it can be expressed by notation $O(KM^2)$. The actual relative time needed for execution of all algorithms as other performance measures will be studied in the following Section.

III. EXPERIMENTAL STUDIES

One of the main goals of conducted experiments was to examine dynamics of swarm's performance for all considered techniques during the optimization process. Running times

TABLE II
BENCHMARK FUNCTIONS USED FOR EXPERIMENTAL STUDIES

f	Name	Expression	Feasible bounds	N	f^*
f_1	Sphere	$f_1(x) = \sum_{i=1}^{M} z_i^2 + f_1^*$ $z = x - o$	$[-100, 100]^N$	10	-1400
f_2	Different Powers	$f_2 = \sqrt{\sum_{i=1}^{N} \lvert z_i \rvert^{2+4\frac{i-1}{N-1}}} + f_2^*$ $z = x - o$	$[-100, 100]^N$	10	-1000
f_3	Rotated Rastrigin	$f_3(x) = \sum_{i=1}^{N} (z_i^2 - 10\cos(2\pi z_i) + 10) + f_3^*$ $z = M_1 \Lambda^{10} M_2 T_{asy}^{0.2}(T_{osz}(M_1 \frac{5.12(x-o)}{100}))$	$[-100, 100]^N$	10	-300
f_4	Schwefel	$f_4(x) = 418.9829N - \sum_{i=1}^{N} g(z_i) + f_4^*$ $z = \Lambda^{10}(\frac{1000(x-o)}{100}) + 4.209687462275036e + 002$ $g(z_i) = z_i \sin(\lvert z_i \rvert^{1/2})$	$[-100, 100]^N$	10	-100
f_5	Rotated Katsuura	$f_5(x) = \frac{10}{N^2} \prod_{i=1}^{N} (1 + i \sum_{j=1}^{32} \frac{\lvert 2^j z_i - round(2^j z_i) \rvert}{2^j})^{\frac{10}{N^{1.2}}} - \frac{10}{N^2} + f_5^*$ $z = M_2 \Lambda^{100}(M_1 \frac{5(x-o)}{100})$	$[-100, 100]^N$	10	200

Symbols:
$o = [o_1, o_2, ..., o_N]$ – shifted global optimum, randomly distributed in $[-80, 80]^N$,
M_1, M_2 – orthogonal (rotation) matrix generated from standard normally distributed entries by Gram-Schmidt orthonormalization.
Λ^α – diagonal matrix in N dimensions with the i^{th} diagonal element $\lambda_{ii} = \alpha^{\frac{i-1}{2(N-1)}}$ for $i = 1, 2, ..., N$.
T_{asy}^β – if $x_i > 0, x_i = x_i^{1+\beta\frac{i-1}{N-1}\sqrt{x_i}}$ for $i = 1, 2, ..., N$.
T_{osz} – for $x_i = sign(x_i)\exp(\hat{x}_i + 0.049(sin(c_1\hat{x}_i) + sin(c_2\hat{x}_i)))$ for $i = 1, 2, ..., N$.
where:
$\hat{x}_i = log(\lvert x_i \rvert)$ for $x_i \neq 0$, otherwise $\hat{x}_i = 0$,
$c_1 = 10$ if $x_i > 0$, otherwise $c_1 = 5.5$,
$c_2 = 7.9$ if $x_i > 0$, otherwise $c_2 = 3.1$.

were also carefully studied as well as final cost function values, which were additionally compared by means of statistical tests. The following subsections are covering the details of algorithms' numerical evaluation.

A. Problems and Experimental Setting

For computational experiments set of benchmark problems considered in CEC'13 competition was used [26]. Table II lists those functions along with their mathematical expressions, dimensionality and optimum values.

The experiments were conducted for fixed number of iterations $K = 10000$ $(1000 * N)$, and 30 trials for each function. Population size $M = 40$ was used for all algorithms. FIPSO was configured with fully-connected topology and parameter values as suggested in Table I. For FA we used $\alpha = 0.15$ and $\gamma = 10$ following the suggestions found in related literature. Random step size was also scaled to the size of search space S. GSO was configured with parameters given in Table I, with modified step size $s = 0.8$. For r_s we used half of maximum distance in S (that is $r_s = 315$) and for r_0 the value of 90 was selected. Both settings were established during a set of pilot runs to adopt neighborhood size properly to the domain of given optimization task.

As a performance measure mean optimization error $\overline{E}(k)$ was used (with $E(k) = |f(x(k)^*) - f^*|$) along with its standard deviation $\sigma_{E(k)}$. We have also studied mean execution time $\overline{t}$ in seconds needed for one algorithm's run.

B. Algorithms' Search Process Dynamics

First set of experiments was aimed at establishing dynamics of swarm performance in the function of execution time (iterations). For all algorithms mean optimization errors in 30 trials during 10000 iterations were reported. The results of this study for all investigated techniques are shown on Figures 1-5.

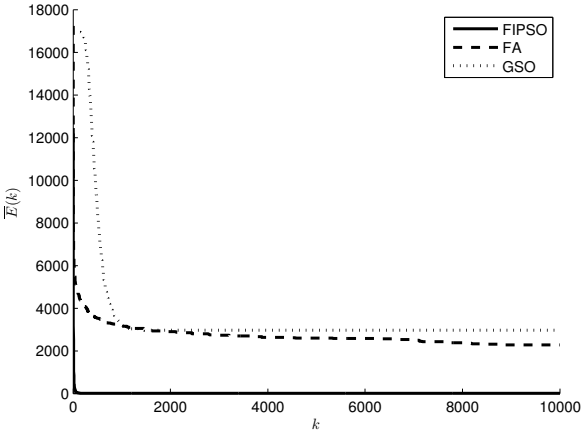


Fig. 1. Mean error values obtained within 10000 iteration of optimization process for f_1 (Sphere) function

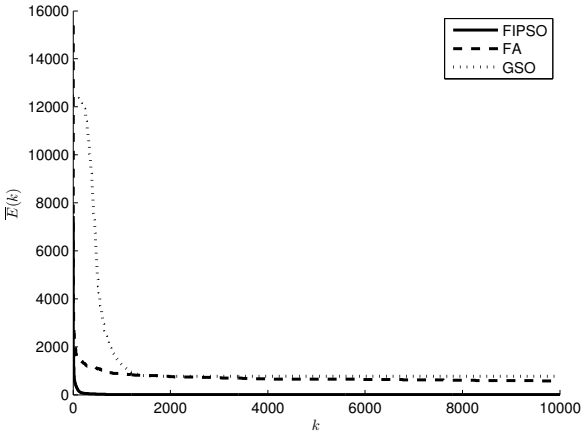


Fig. 2. Mean error values obtained within 10000 iteration of optimization process for f_2 (Different Powers) function

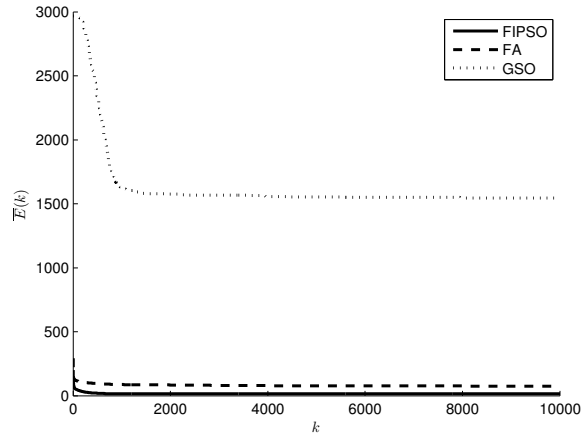


Fig. 3. Mean error values obtained within 10000 iteration of optimization process for f_3 (Rotated Rastrigin) function

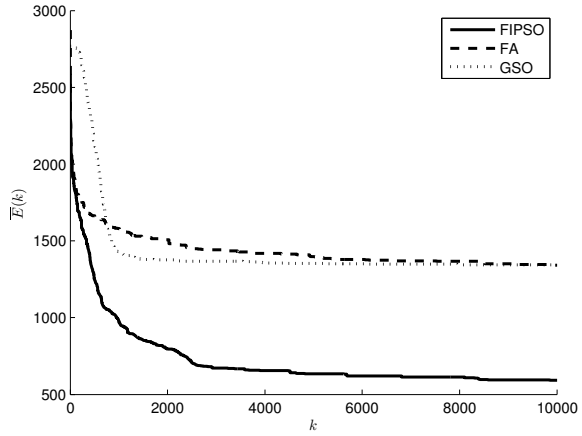


Fig. 4. Mean error values obtained within 10000 iteration of optimization process for f_4 (Schwefel) function

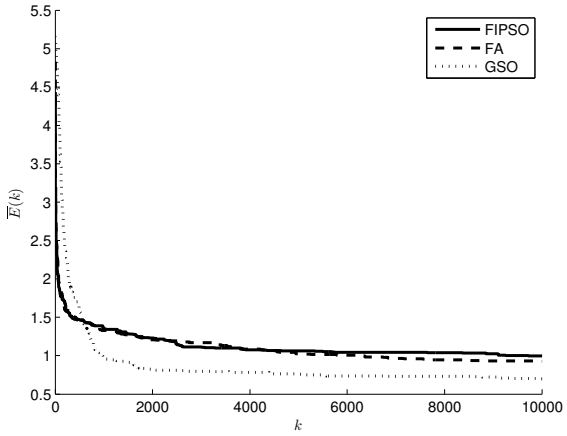


Fig. 5. Mean error values obtained within 10000 iteration of optimization process for f_5 (Rotated Katsuura) function

It can be seen that in the framework of GSO swarm members tend to find local minima and to stop exploring problem's search space afterwards. This tendency will demonstrate itself again in the next analysis. As for dynamics of FA it is also prone for being trapped in local minima. Most intense progress of error values' minimization during first optimization phase was observed for FIPSO, for all but one (f_5) function.

C. Algorithms' Performance

To compare performance of all considered algorithms we examined final cost function values obtained through a set of runs described above. In this quantitative comparison, both means, standard deviations and best obtained cost values were reported. Relation between algorithm's performance indicators was studied by means of pairwise T-tests. In most cases differences proved to be statistically meaningful at 0.95 significance level. We ranked all algorithms according to conclusive results of those tests.

Among studied algorithms FIPSO was found to be best-performing one. What is more its execution for fully-connected topology requires the least computational effort. Among two other algorithms FA was found to perform better than GSO, however here the difference – both in terms of final cost function values and required execution time – proved to be not very substantial. The most difficult problem for all the algorithms to tackle was minimization of Schwefel function. At the same time surprisingly Sphere function proved to be problematic for FA and GSO algorithms.

IV. Conclusion

The paper examined practical features of a set of important nature-inspired metaheuristics: Fully Informed Particle Swarm Optimization, Firefly Algorithm and Glowwworm Swarm Optimization. They were for the first time considered within the same methodological context, with their performance being examined on a set of benchmark functions.

We found that when using the most basic variants of aforementioned techniques Fully Informed Particle Swarm Optimization proved to be most effective and least-computationally expensive. At the same time Firefly Algorithm and Glowworm Swarm Optimization were found to be similar, both in terms of performance and computational resources' requirements. From our observations GSO is a very powerful technique for discovering local minima especially when it employs large swarm population. To use it as a tool of global optimization however one should complement it by randomized local search algorithm (e.g. Simulated Annealing [27]) or other procedure which would allow swarm members to escape from cost function valleys. In case of FA we see a suitable choice of random step generation method (e.g. Lévy Flight [28]) as a crucial element for successful applications in the field of continuous optimization.

Finally it should be noted that performed experiments do not include recent modifications of all analyzed optimization strategies, described briefly in Section 2, and more specific benchmarks. Their results however can be used as a baseline for enriching this study by including those aspects. It is planned within forthcoming follow-up contribution [29].

Acknowledgment

First author is thankful to Marco A. Montes de Oca for helpful suggestions and inspiration provided during his stay at Université Libre de Bruxelles.

References

[1] J. Kennedy and R. Eberhart, "Particle Swarm Optimization," in *Proceedings of IEEE International Conference on Neural Networks, vol. IV*, 1995. doi: 10.1109/ICNN.1995.488968 pp. 1942–1948. [Online]. Available: http://dx.doi.org/10.1109/ICNN.1995.488968

[2] T. Cura, "A particle swarm optimization approach to clustering," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1582–1588, 2012. doi: 10.1016/j.eswa.2011.07.123. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2011.07.123

[3] Y. Morsly, N. Aouf, M. Djouadi, and M. Richardson, "Particle Swarm Optimization Inspired Probability Algorithm for Optimal Camera Network Placement," *IEEE Sensors Journal*, vol. 12, no. 5, pp. 1402–1412, 2012. doi: 10.1109/JSEN.2011.2170833. [Online]. Available: http://dx.doi.org/10.1109/JSEN.2011.2170833

TABLE III
ALGORITHMS' PERFORMANCE COMPARISON

Function	FIPSO				FA				GSO				Pairwise T-test results (order)
	$\overline{E}(K)$	$\sigma_{E(K)}$	$\overline{t}$	$\min E(K)$	$\overline{E}(K)$	$\sigma_{E(K)}$	$\overline{t}$	$\min E(K)$	$\overline{E}(K)$	$\sigma_{E(K)}$	$\overline{t}$	$\min E(K)$	
f_1	0.00	0.00	26.33	0.00	2273.98	362.74	194.71	1516.96	2972.20	1715.90	233.95	946.33	1.FIPSO, 2.FA, 3.GSO
f_2	0.01	0.01	28.48	0.00	575.11	105.46	193.26	341.09	774.01	501.62	217.58	178.25	1.FIPSO, 2.FA, 3.GSO
f_3	14.80	5.20	30.84	4.00	73.02	8.20	196.82	51.98	80.04	26.55	238.45	21.24	1.FIPSO. 2.FA,GSO
f_4	592.85	220.73	31.87	261.83	1341.30	124.44	234.44	937.65	1343.50	339.56	194.29	641.79	1.FIPSO, 2.FA,GSO
f_5	1.00	0.14	32.55	0.68	0.93	0.14	249.83	0.67	0.70	0.22	212.90	0.31	1.GSO, 2.FIPSO, 3.FA

[4] W. Fang, J. Sun, Y. Ding, X. X. Wu, and W. Xu, "A Review of Quantum-behaved Particle Swarm Optimization," *IETE Technical Review*, vol. 27, pp. 336–348, 2010.

[5] A. Röhler and S. Chen, "Multi-swarm hybrid for multi-modal optimization," in *Proceedings of the IEEE Congress on Evolutionary Computation*, 2012. doi: 10.1109/CEC.2012.6256566 pp. 1759–1766. [Online]. Available: http://dx.doi.org/10.1109/CEC.2012.6256566

[6] R. Mendes, J. Kennedy, and J. Neves, "The Fully Informed Particle Swarm: Simpler, Maybe Better," *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 2, pp. 204–210, 2004. doi: 10.1109/TEVC.2004.826074. [Online]. Available: http://dx.doi.org/10.1109/TEVC.2004.826074

[7] X. Yang, *Nature-Inspired Metaheuristic Algorithms*. Frome: Luniver Press, 2008.

[8] S. Łukasik and S. Żak, "Firefly Algorithm for Continuous Constrained Optimization," *Lecture Notes in Artificial Intelligence*, vol. 5796, pp. 97–106, 2009. doi: 10.1007/978-3-642-04441-0_8. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-04441-0_8

[9] K. Krishnanand and D. Ghose, "Glowworm Swarm Optimization for Simultaneous Capture of Multiple Local Optima of Multimodal Functions," *Swarm Intelligence*, vol. 3, no. 2, pp. 87–124, 2008. doi: 10.1007/s11721-008-0021-5. [Online]. Available: http://dx.doi.org/10.1007/s11721-008-0021-5

[10] T. Havens, C. Spain, N. Salmon, and J. Keller, "Roach infestation optimization," in *Proceedings of the IEEE Swarm Intelligence Symposium 2008*, 2008. doi: 10.1109/SIS.2008.4668317 pp. 1–7. [Online]. Available: http://dx.doi.org/10.1109/SIS.2008.4668317

[11] D. Pham, A. Ghanbarzadeh, E. Koç, S. Otri, S. Rahim, and M. Zaidi, "The Bees Algorithm - A Novel Tool for Complex Optimisation Problems," in *Proceedings of IPROMS 2006 Conference*, 2006, pp. 454–461.

[12] X. Yang and A. Gandomi, "Bat algorithm: a novel approach for global engineering optimization," *Engineering Computations*, vol. 29, no. 5, pp. 464–483, 2012. doi: 10.1108/02644401211235834. [Online]. Available: http://dx.doi.org/10.1108/02644401211235834

[13] P. Suganthan, "Particle swarm optimiser with neighbourhood operator," in *Proceedings of the 1999 Congress on Evolutionary Computation, Vol. 3*, vol. 3, 1999. doi: 10.1109/CEC.1999.785514 pp. 1958–1962. [Online]. Available: http://dx.doi.org/10.1109/CEC.1999.785514

[14] M. A. M. de Oca and T. Stutzle, "Convergence Behavior of the Fully Informed Particle Swarm Optimization Algorithm," in *Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation*, 2008. doi: 10.1145/1389095.1389106 pp. 71–78. [Online]. Available: http://dx.doi.org/10.1145/1389095.1389106

[15] H. Tehzeeb-Ul-Hassan, R. Zafar, S. Mohsin, and O. Lateef, "Reduction in power transmission loss using fully informed particle swarm optimization," *International Journal of Electrical Power & Energy Systems*, vol. 43, no. 1, pp. 364 – 368, 2012. doi: 10.1016/j.ijepes.2012.05.028. [Online]. Available: http://dx.doi.org/10.1016/j.ijepes.2012.05.028

[16] L. Hongliang, E. Howely, and J. Duggan, "Particle Swarm Optimisation with Gradually Increasing Directed Neighbourhoods," in *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation*, 2011. doi: 10.1145/2001576.2001582 pp. 29–36. [Online]. Available: http://dx.doi.org/10.1145/2001576.2001582

[17] M. Setayesh, M. Zhang, and M. Johnston, "Effects of Static and Dynamic Topologies in Particle Swarm Optimisation for Edge Detection in Noisy Images," in *Proceedings of the 2012 IEEE Congress on Evolutionary Computation*, 2012. doi: 10.1109/CEC.2012.6256104 pp. 1–8. [Online]. Available: http://dx.doi.org/10.1109/CEC.2012.6256104

[18] M. A. M. de Oca, T. Stutzle, M. Birattari, and M. Dorigo, "Frankenstein's PSO: A composite particle swarm optimization algorithm," *IEEE Transactions on Evolutionary Computation*, vol. 13, no. 5, pp. 1120–1132, 2009. doi: 10.1109/TEVC.2009.2021465. [Online]. Available: http://dx.doi.org/10.1109/TEVC.2009.2021465

[19] X. Yang and X. He, "Firefly algorithm: recent advances and applications," *International Journal of Swarm Intelligence*, vol. 1, no. 1, pp. 36 – 50, 2013. doi: 10.1504/IJSI.2013.055801. [Online]. Available: http://dx.doi.org/10.1504/IJSI.2013.055801

[20] P. Oramus, "Improvements to Glowworm Swarm Optimization algorithm," *Computer Science*, vol. 11, pp. 7–20, 2010. doi: 10.7494/csci.2010.11.0.7. [Online]. Available: http://dx.doi.org/10.7494/csci.2010.11.0.7

[21] K. Krishnanand and D. Ghose, "Theoretical Foundations for Rendezvous of Glowworm-inspired Agent Swarms at Multiple Locations," *Robotics and Autonomous Systems*, vol. 56, no. 7, pp. 549–569, 2008. doi: 10.1016/j.robot.2007.11.003. [Online]. Available: http://dx.doi.org/10.1016/j.robot.2007.11.003

[22] A. Karegowda and M. Prasad, "A Survey of Applications of Glowworm Swarm Optimization Algorithm," in *IJCA Proceedings of the International Conference on Computing and Information Technology 2013*, 2013, pp. 39–42.

[23] L. Zhang, H. Yu, and S. Hu, "Optimal choice of parameters for particle swarm optimization," *Journal of Zhejiang University Science A*, vol. 6, no. 6, pp. 528–534, 2005. doi: 10.1007/BF02841760. [Online]. Available: http://dx.doi.org/10.1007/BF02841760

[24] M. Clerc and J. Kennedy, "The particle swarm - explosion, stability, and convergence in a multidimensional complex space," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 1, pp. 58–73, 2002. doi: 10.1109/4235.985692. [Online]. Available: http://dx.doi.org/10.1109/4235.985692

[25] M. Yuan-bin, M. Yan-zhui, and Z. Qiao-yan, "Optimal Choice of Parameters for Firefly Algorithm," in *Proceedings of the Fourth International Conference on Digital Manufacturing and Automation (ICDMA)*, 2013. doi: 10.1109/ICDMA.2013.210 pp. 887–892. [Online]. Available: http://dx.doi.org/10.1109/ICDMA.2013.210

[26] J. Liang, B. Qu, P. Suganthan, and A. Hernandez-Diaz, "Problem Definitions and Evaluation Criteria for the CEC 2013 Special Session and Competition on Real-Parameter Optimization," 2013, technical Report 201212, Computational Intelligence Laboratory, Zhengzhou University, Zhengzhou China and Technical Report, Nanyang Technological University, Singapore.

[27] Y. Zhou, G. Zhou, and J. Zhang, "A Hybrid Glowworm Swarm Optimization Algorithm for Constrained Engineering Design Problems," *Applied Mathematics and Information Sciences*, vol. 7, no. 1, pp. 379–388, 2013.

[28] X. Yang, "Firefly Algorithm, Lévy Flights and Global Optimization," in *Research and Development in Intelligent Systems XXVI*, M. Bramer, R. Ellis, and M. Petridis, Eds. Springer London, 2010, pp. 209–218. [Online]. Available: http://dx.doi.org/10.1007/978-1-84882-983-1_15

[29] S. Łukasik and P. Kowalski, "Reviewing Fully Informed Swarm Optimization Algorithms," 2014, in preparation.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 163–169
DOI: 10.15439/2014F107
ACSIS, Vol. 2

Ladder Tagger — Splitting Decision Space to Boost Tagging Quality

Mariusz Paradowski, Adam Radziszewski
Institute of Informatics, Wrocław University of Technology
Wybrzeże Stanisława Wyspiańskiego 27, 50-370 Wrocław
Poland

***Abstract*—This paper describes a part of speech tagger. The tagger is based on a set of probability mixture models. Each mixture model is responsible for tagging of a specific class of words, sharing similar context properties. Probability mixture models contain 25 various mixture components. The tagger is tested on Polish language and compared to other available taggers.**

I. Introduction

PART-OF-SPEECH (POS) tagging is a common and well-researched Natural Language Processing (NLP) task. It is the process of assigning POS tags to words and word-like units (*tokens*) in text. In languages with rich morphology the tags usually include significantly more information than just parts-of-speech, e.g. nouns may be specified for values of number, gender and case, adverbs may be specified for degree. In such a setting, the task is referred to as morphosyntactic tagging.

In this paper we present a novel approach to Part-of-Speech tagging (POS tagging), where the labelling is performed in a non-sequential manner using an array of simple probability mixture models. The models are derived directly from the training data. The approach has been developed for Polish, an inflective language that is typically described with a rich, positional tagset. However, very little language-dependent information is in fact employed and it is reasonable to expect that the approach will work equally well for other positional tagsets.

The presented version of the tagger deals only with disambiguation of the grammatical class (roughly corresponding to Part-of-Speech). In spite of that, it already makes use of the information available in the whole tags.

II. Tagging of inflective languages

Tagging of inflective languages is a not an easy task. One of the main difficulty is the relatively free word order, which makes sequences of n-grams much less frequent than in English [1]. The other major difficulty is related to the size and characteristics of tagsets. English tagsets usually define 40–200 different tags, while the 1-million-token manually annotated part of the National Corpus of Polish [2] contains about 1000 different tags. This is a frequently cited reason of low performance of taggers for inflective languages [3]–[5].

Tagging of inflective languages is usually performed in two stages: morphological analysis and morphological disambiguation [6]. The first stage is essentially dictionary look-up, resulting in attaching sets of tags attached to each token. The proper tagging happens during the disambiguation phase — tags that are recognised as contextually inappropriate are removed from the sets. The other technique that is commonly used is called *tiered tagging* (originating from a work by Tufiş [7]). This technique assumes splitting of positional tags according to some groups (tiers) of grammatical categories that the tags consist of. A sentence is disambiguated iteratively, one tier at a time.

Both techniques are used in three taggers made for Polish language during the last five years. PANTERA [8] is built upon an adaptation of the Brill's transformation-based learning algorithm. The tagger employs three tiers, but also makes uses of modified rule templates that operate on the level of particular tagset attributes (grammatical categories) instead of whole tags, which is an important enhancement when dealing with positional tagsets. Also, the sets of tags assigned in the course of morphological analysis are used to constrain possible transformations: if any transformation would result in generation of a tag not accounted for by the morphological dictionary, the transformation is cancelled.

WMBT [9] is a memory-based tagger that introduces morphological analysis and tiered tagging to the standard memory-based tagging framework, namely MBT [10]. The tagger uses as many tiers as there are attributes in the tagset (plus one for the grammatical class). The algorithm iterates over tiers; tagging of one tier involves classification of subsequent tokens with a k-Nearest Neighbour classifier. The classification process benefits from a rich feature set, including values of particular attribute (grammatical class, number, gender and case for tokens surrounding the token being tagged), but also tests for morphological agreement on number, gender and case. WMBT was later enhanced with a simple handling for unknown words [11]. The procedure assumes collecting separate case bases for known and unknown words.

WCRFT [11] is a modification of the WMBT tagger, where a linear-chain first-order Conditional Random Field (CRF) is used instead of the k-NN classifier. Also, instead of classifying independently each token, the CRF model is used to classify a whole sentence at a time. A separate model is trained for each tier and these models are run sequentially when performing disambiguation. The feature set is taken directly from WMBT. Also, similar unknown word handling procedure is employed.

Concraft [12] is based on a new mathematical model

978-83-60810-58-3/$25.00

devised for the purpose of morphosyntactic tagging, namely Constrained Conditional Random Fields (CCRF). CCRF is an extension of the CRF model where additional constraints are imposed on the set of labels that may be produced for each token in the output sequence. The constraints are used to enforce that each output tag belongs to the set of tags generated during morphological analysis. The model used for disambiguation is second-order. An interesting feature of the tagger is that disambiguation is performed on all layers simultaneously instead of using the standard tiered tagging scheme. Concraft also contains a separate model for handling of unknown words, which is based on first-order CCRF.

III. Ladder Tagger

In this paper we propose a new part-of-speech tagger called *Ladder Tagger*. Formally, part-of-speech tagging may be modelled as a sequence I of dependent multi-class decision problems $w \in I$. Decision problems w are statistically dependent because previously made decisions (tag assignments) influence further ones.

Natural language structure is highly complex, often very flexible, full of exceptions. The main idea of the proposed method is to divide the sequence of decision problems I into n subsets of problems with similar properties:

$$D_1 \subseteq I, D_2 \subseteq I, ..., D_n \subseteq I, \tag{1}$$

$$D_1 \cap D_2 \cap ... \cap D_n = \emptyset. \tag{2}$$

Each of these subsets $D_i : i \in \{1, ..., n\}$ should model a part of language complex structure. Statistical dependence of decision problems $w \in I$ makes the final tagging result dependent on the order in which they are solved. Our idea is to solve decision problems from the easiest to the most difficult ones (hence the name of the proposed tagger) in a non-sequential manner. The major advantage of this approach is that both **preceding** and **succeeding** tagging results may influence the current result.

The second contribution of the proposed tagger is highly intense usage of tagging context information. In probabilistic tagging smoothing plays a key role. Smoothing models may be more or less complex, but they should provide statistically significant information regarding token context. Of course smoothing parameters are highly dependent on the decision problem to solve. In the proposed tagger we use as much as 25 probability smoothing components, both taking into account preceding and succeeding tokens. Smoothing parameters are independently estimated for each set of decision problems.

The general outline of the tagging process is presented in Fig. 1. First, we assign each token (decision problem) to a set of decision problems according to its predefined properties. The first set D_1 contains the easiest, unambiguous tokens (they are trivial single-class cases). Remaining sets $D_2, ..., D_n$ contain difficult, multi-class decision problems. Sets of decision problems are now solved according the their difficulty.

Require: I – sequence of decision problems (tokens to be POS-tagged)
Ensure: t – part-of-speech tagging of I
1: $D_1 \leftarrow \emptyset$, $D_2 \leftarrow \emptyset$, ..., $D_n \leftarrow \emptyset$
2: **for all** $w \in I$ **do**
3: $m \leftarrow$get-problem-type(w)
4: $D_m \leftarrow D_m \cup \{w\}$
5: **end for**
6: **for all** $D \in \{D_1, D_2, ..., D_n\}$ **do**
7: **for all** $w \in D$ **do**
8: $t_w \leftarrow$solve-problem(w)
9: **end for**
10: **end for**
11: **return** t

Figure 1. Ladder Tagger algorithm

Our proposal follows the two-stage scheme described in the previous section: tagger's input must be first subjected to morphological analysis and the core algorithm is responsible for morphological disambiguation. In other words, it is assumed that each token is attached a set of *possible tags*[1] and each decision problem consists in selecting one of its elements. The information on *possible tags* is used during both training and normal tagger performance (disambiguation) to designate *ambiguity classes*.

A. Subsets of decision problems

The main goal of decision problems division is to build better, more specific classifiers. One of the key issues in case of POS tagging is proper usage of the available training information. Different type of information should be used in case of e.g. frequent known words, rare known words and unknown words. Thus, the following **classes of decision problems** are defined:

1) specific *known words* (SKN),
2) frequent *known words* (FKN),
3) possible *grammatical class (POS) ambiguity classes* of known words (GKN),
4) rare *known words* (RKN),
5) *word suffixes or prefixes* for unknown words (UNK).

Decision problem subsets are instances of the above classes. Specific known words (SKN) class contains very frequent words with lots of grammar exceptions. List of these words should be pre-specified by expert linguists. The list should not be too long, because classifiers for each specific known word are trained separately. Frequent known words (FKN) class contains frequently used words, generally following grammar rules. Due to their high frequency in the corpus they have a large training data available. *Ambiguity classes* class is much less specific, it often covers multiple known words. Number

[1]Possible according to the morphological analyser employed. Even if a word form is known to the analyser, the returned set of tags may be incomplete or erroneous. Given the evolving nature of language, it is impossible to create an exhaustive morphological dictionary that will be valid for any text.

of known instances of a given word is another important criterion. If a word has only one instance in the training set, any parameter estimation is not possible for such a word. Such words are considered rare and are tagged according to ambiguity classes.

The employed two-stage scheme is bound to face the problem of unknown words. Word forms of some tokens will be not recognised by the morphological analyser, hence no set of possible tags (ambiguity class) will be available. Such tokens are treated as a separate class (UNK). Tagging of such words is language-specific. For Polish (and also for many Slavic languages), word suffixes and, to some extent, prefixes, give useful inflectional information that helps make the correct decision.

During tagging process, a token is treated as an unknown word if the analyser failed to provide a set of possible tags. For our model to work, the UNK class must also be explicitly present in the training data. Fortunately, this information is easy to obtain. To make sure that the sets of possible tags present in the training corpus reflect the actual behaviour of the morphological analyser, we followed the procedure called *reanalysis of the training data* that was proposed in [11]. It assumes feeding the training data through the morphological analyser actually used to update the information on tags *possible* for each token. In case of word forms that were unknown to the analyser, the set of possible tags always consists of the proper tag (taken from the original training data) and the *unknown word tag* (`ign`). What results is consistent assignment of tokens to ambiguity classes, but also, explicit marking of unknown words.

B. Decision model

The proposed tagger applies Bayesian decision model. Maximum a posteriori decision rule is used:

$$t_w^* = \arg\max_{t \in g_w} p(t|w), \tag{3}$$

where: w – a word being currently tagged, t_w^* – chosen tag, g_w – set of possible POS tags for word w.

Class probability $p(t|w)$ is estimated from available training corpus data. Each decision problem subset has its own subset of corpus. Corpus subsets are built according to predefined rules. Rules of corpus subset generation for possible classes of decision problems are given in Tab. I.

Table I
RULES OF CORPUS SUBSET GENERATION FOR VARIOUS DECISION PROBLEM CLASSES.

problem class	word form	grammatical classes	prefix suffix
SKN	match	match	—
FKN	match	match	—
GKN	—	match	—
RKN	—	match	—
UNK	—	—	match

Exemplary, word *to* (eng. *it*) is belongs to the class of specific known word problems. The training set consists of all instances of word *to* with matching set of possible grammatical classes.

C. Probability smoothing

Probability smoothing is one of the key components in statistical approaches to tagging. There is no sufficient data to get well estimated probability distributions of word sequences. One of the most common approach to probability smoothing is Jelinek-Mercer smoothing [13]. The basic model is defined as follows:

$$p(w_0|w_1)_J = \lambda p(w_0|w_1)_D + (1-\lambda)p(w_0)_D, \tag{4}$$

where: $\lambda \in \langle 0;1 \rangle$ and $p(w_0|\cdot)_D$ is estimated directly from the data.

The Jelinek-Mercer model is further extended to Witten-Bell [13] smoothing and it takes a recurrent form:

$$p(w_0|s_i)_R = \lambda_i p(w_0|s_i)_D + (1-\lambda_i)p(w_0|s_{i-1})_R, \tag{5}$$

where: $\lambda_i \in \langle 0;1 \rangle$, s_i is the context of word w_0.

The above model $p(w_0|s_i)_R$ can be simply rewritten into non-recurrent equation and it takes the basic form of a well known probability mixture model. Thus, the probability smoothing used in the proposed approach is defined as follows:

$$p(w_0|s_i)_M = \sum_{j=0}^{i} \alpha_j p(w_0|s_j)_D, \tag{6}$$

where:

$$\alpha_i = \lambda_i \prod_{j=0}^{i-1}(1-\lambda_j), \quad \sum_{j=0}^{i} \alpha_j = 1. \tag{7}$$

Estimation of α_i parameters is not straightforward, and can be done in several different ways. In more complex language models (like the one proposed), the number of parameters may be high.

D. Mixture model probabilities

One of the key components to successful tagging is the definition of smoothing probabilities. As shown above, smoothing may be represented as a mixture probability. Component probabilities are estimated on four different types of data: word forms (w_i), grammatical class (POS) ambiguity classes (g_i), ambiguity classes of whole tags (c_i) and *tagging results* (grammatical classes assigned by the tagger) (t_i), where i stands for relative position of a word in the tagged corpus. In case the word in not yet tagged and a *tagging results* (t_i) is used, the probability is estimated as 0. Given the above events, various conditional probabilities can be defined. We use total 25 probability mixture model components, as shown in Tab. II. For instance, $p(t_0|x_0, g_{-1}, w_{+1})$ represents probability of tag t_0, given that:

1) decision class specific parameters x_0 on position 0 are matched,
2) grammatical class (POS) ambiguity classes g_{-1} on position -1 matches,
3) word form w_{+1} on position $+1$ matches.

Table II
PROBABILITY MIXTURE COMPONENTS USED FOR PROBABILITY SMOOTHING, x_0 REPRESENTS A DECISION CLASS SPECIFIC EVENT. FOR *SKN* AND *FKN* $x_0 = (w_0, g_0)$, FOR *GKN* AND *RKN* $x_0 = g_0$ AND FOR *UNK* x_0 IS DEFINED IN TERMS OF SUFFIXES AND PREFIXES.

$p(t_0\|\cdot)$	word forms (w)	g-class(g)	f-class(c)	tags(t)
$p(t_0\|x_0, w_{+1}, w_{+2})$	two succ.	–	–	–
$p(t_0\|x_0, w_{-1}, w_{-2})$	two prec.	–	–	–
$p(t_0\|x_0, w_{-1}, w_{+1})$	two neigh.	–	–	–
$p(t_0\|x_0, w_{-1}, g_{+1})$	first prec.	first succ.	–	–
$p(t_0\|x_0, g_{-1}, w_{+1})$	first succ.	first prec.	–	–
$p(t_0\|x_0, w_{-1}, t_{+1})$	first prec.	–	–	first succ.
$p(t_0\|x_0, t_{-1}, w_{+1})$	first succ.	–	–	first prec.
$p(t_0\|x_0, t_{-2}, t_{-1})$	–	–	–	two prec.
$p(t_0\|x_0, t_{-1}, t_{+1})$	–	–	–	two neigh.
$p(t_0\|x_0, c_{-2}, c_{-1}, c_{+1})$	–	–	three neigh.	–
$p(t_0\|x_0, w_{-1}, c_{+1})$	first prec.	–	first succ.	–
$p(t_0\|x_0, c_{-1}, w_{+1})$	first succ.	–	first prec.	–
$p(t_0\|x_0, c_{-2}, c_{-1})$	–	–	two prec.	–
$p(t_0\|x_0, c_{-1}, c_{+1})$	–	–	two neigh.	–
$p(t_0\|x_0, w_{+1})$	first succ.	–	–	–
$p(t_0\|x_0, w_{-1})$	first prec.	–	–	–
$p(t_0\|x_0, t_{-1}, g_{+1})$	–	first succ.	–	first prec.
$p(t_0\|x_0, g_{-1}, g_{+1})$	–	two neigh.	–	–
$p(t_0\|x_0, c_{+1})$	–	–	first succ.	–
$p(t_0\|x_0, c_{-1})$	–	–	first prec.	–
$p(t_0\|x_0, t_{+1})$	–	–	–	first succ.
$p(t_0\|x_0, t_{-1})$	–	–	–	first prec.
$p(t_0\|x_0, g_{+1})$	–	first succ.	–	–
$p(t_0\|x_0, g_{-1})$	–	first prec.	–	–
$p(t_0\|x_0)$	–	–	–	–

For the easiness of reading, mixture model components shown in Tab. II are sorted according to their assumed generality. More specific general estimates are shown at the top, more general ones are shown at the bottom of the table. The main idea of probability smoothing is preserved. Mixture components from the top of the table will have low recall, however high precision. Those from the bottom will have high recall and lower precision.

E. Tagger parameter estimation

The proposed tagger has a set of parameters, as described above, which need to be estimated before the tagging process. Parameter estimation is done at the training set. Given the smoothing model has n components and that there are m sets of decision problems, there are total nm parameters to be estimated. For each set of decision problems its smoothing model parameters are estimated independently.

Mixture parameters estimation is considered as a validation set tagging quality optimization problem. Validation sets are built on top of the training set using Leave-One-Out routine. Tagging quality is defined as a classic *recognition accuracy*. Training process is defined as maximization of recognition accuracy:

$$[\alpha]^* = \arg\max_{[\alpha]} \frac{1}{|I|} \sum_{w \in I} |\{t_w^*\} \cap \{r_w\}|, \tag{8}$$

where: r_w – reference tag for word w, t_w^* – tagging result of word w from the validation set.

Random-restart hill climbing method is used as the optimization tool. Exemplary estimates of smoothing parameters are given in Tab. III. Estimated mixture weights may provide an interesting insight into the language structure. They may show the importance of different contexts in various cases of language usage. Exemplary, word *to* (eng. *it*) has a very high importance of succeeding context, while in case of word *i* (eng. *and*) proceeding context is much more important. Thus, results of training may provide further information (and have done so in the past) how to extend available set of contexts.

F. Computational and memory complexity analysis

Proposed tagging approach is can be classified as a *lazy recognition method*, because it estimates probabilities directly from the data during the recognition phase. Thus, it requires memorization of the whole training data, i.e., the tagged corpus. Estimation of mixture probabilities requires iteration through the data, but can be limited only to its small subsets. The whole training set is organized as a *hash map*. Each element of the hash map represents a *decision class specific event* x_0. It contains a list of all indexes of event x_0 appearance in the training set. For example, to estimate $p(t_0|x_0, w_{+1})$ where: $x_0 = (w_0, g_0)$, w_0 is a specific word and g_0 a specific grammatical class, the tagger accesses the hash map with key (w_0, g_0) and extracts a list of relevant indexes. Given the list of indexes, it iterates through the training set and checks only for w_{+1} match condition. This results in a large speedup of tagging process, because size m of the relevant index is always much smaller than the size n of the whole training set.

As a result, memory complexity is equal to $T(kn)$ where n is the size of the corpus and $k = 5$ is the number of predefined decision problem classes (see Tab. I). The number of predefined classes is constant and practically should not

Table III
EXEMPLARY PROBABILITY MIXTURE MODEL PARAMETERS (WEIGHTS) FOR EACH SPACE SUBDIVISION.

$p(t_0\|\cdot)$	*to*(*it*)/SKN	*czy*(*if*)/SKN	*i*(*and*)/SKN	*frequent*/FKN	*rare*/RKN
$p(t_0\|x_0, w_{+1}, w_{+2})$	0.148	0.130	0.050	0.203	0.031
$p(t_0\|x_0, w_{-1}, w_{-2})$	0.069	0.154	0.252	0.086	0.083
$p(t_0\|x_0, w_{-1}, w_{+1})$	0.046	0.104	0.110	0.030	0.005
$p(t_0\|x_0, w_{-1}, g_{+1})$	0.056	0.010	0.004	0.075	0.005
$p(t_0\|x_0, g_{-1}, w_{+1})$	0.039	0.046	0.018	0.007	0.005
$p(t_0\|x_0, w_{-1}, t_{+1})$	0.013	0.003	0.073	0.018	0.041
$p(t_0\|x_0, t_{-1}, w_{+1})$	0.019	0.100	0.022	0.067	0.104
$p(t_0\|x_0, t_{-2}, t_{-1})$	0.046	0.013	0.018	0.037	0.005
$p(t_0\|x_0, t_{-1}, t_{+1})$	0.003	0.040	0.004	0.003	0.005
$p(t_0\|x_0, c_{-2}, c_{-1}, c_{+1})$	0.013	0.043	0.004	0.011	0.072
$p(t_0\|x_0, w_{-1}, c_{+1})$	0.135	0.090	0.041	0.060	0.083
$p(t_0\|x_0, c_{-1}, w_{+1})$	0.023	0.046	0.087	0.060	0.166
$p(t_0\|x_0, c_{-2}, c_{-1})$	0.046	0.003	0.091	0.015	0.005
$p(t_0\|x_0, c_{-1}, c_{+1})$	0.003	0.013	0.004	0.026	0.020
$p(t_0\|x_0, w_{+1})$	0.072	0.003	0.004	0.037	0.072
$p(t_0\|x_0, w_{-1})$	0.115	0.016	0.022	0.086	0.088
$p(t_0\|x_0, t_{-1}, g_{+1})$	0.003	0.003	0.073	0.011	0.020
$p(t_0\|x_0, g_{-1}, g_{+1})$	0.023	0.026	0.022	0.037	0.010
$p(t_0\|x_0, c_{+1})$	0.069	0.016	0.004	0.052	0.119
$p(t_0\|x_0, c_{-1})$	0.029	0.006	0.027	0.011	0.020
$p(t_0\|x_0, t_{+1})$	0.003	0.040	0.004	0.030	0.005
$p(t_0\|x_0, t_{-1})$	0.003	0.020	0.009	0.007	0.005
$p(t_0\|x_0, g_{+1})$	0.006	0.026	0.009	0.003	0.005
$p(t_0\|x_0, g_{-1})$	0.006	0.026	0.018	0.015	0.010
$p(t_0\|x_0)$	0.003	0.010	0.018	0.003	0.005

grow any more ($k = const$), thus memory complexity is $T(kn) = O(n)$. Computational complexity for tagging a single word is equal to $O(im)$, where i is the number of mixture components and m is the size of x_0 relevant index, where $m \ll n$.

IV. EXPERIMENTAL VERIFICATION

In this section we describe our experiments aiming at evaluation of the proposed disambiguation method (Ladder Tagger). Its results are compared to the three Polish taggers described in Sec. II.

As noted earlier, the present variant of the tagger deals only with disambiguation of the grammatical class, hence the scope of evaluation is also limited to labelling with grammatical classes, while the ability to assign full morphosyntactic tags is not assessed.

The experiments described here are made using the manually annotated part of the National Corpus of Polish [2], version 1.0. This part consists of 86 thousand sentences and 1.2 million tokens and we call it NCP in short. Each token is labelled with exactly one tag, belonging to the NCP tagset [14]. The tagset defines 35 grammatical classes (besides one special class reserved for unrecognised forms — `ign`).

A. Experimental protocol

It has been argued [15] that taggers should be evaluated as whole systems that process plain text files into tagged corpora. Such an approach provides insight into tagging errors made at every possible stage, including tokenisation, morphological analysis and disambiguation. This is a close approximation of real-life scenario of tagger application where only text is available, possibly divided into paragraphs, but with no linguistic pre-processing such as manual division into tokens.

We employ this approach here. The taggers are assessed using a metric called *accuracy lower bound* [15] that is defined as the percentage of tokens from the reference corpus (manually divided into tokens and labelled with tags) that fulfil two conditions:

1) the token is present at the tagger output (no change in tokenisation took place),
2) the tagger classified the token with the same label as in the reference corpus.

In other words, every change in tokenisation is penalised as a tagging error and tags attached to tokens that are subjected to segmentation change are not even checked.

The second condition refers to "the same label". As our evaluation is limited to tagging with grammatical classes, the label is understood as grammatical class extracted from full morphosyntactic tag.

All the experiments described here were performed using ten-fold cross-validation. Each experiment is run against the same partitioning of the data. Each run n for a tagger T consists of the following steps:

1) Training data part n is used to train tagger model M_n^T. Before training proper, the training part is subjected to morphological reanalysis as described in Sec. III-A.
2) Testing data part n ($Test_n^T$) is converted to plain text. The division into paragraphs is preserved and marked with two newline characters.
3) Testing data in plain text part n is tagged with the trained model M_n^T and its output is saved to Out_n^T.
4) Out_n^T is compared to $Test_n^T$ and value of accuracy lower

Table IV
ACCURACY LOWER BOUND MEASURED FOR **ALL** TOKENS

Split	LT	WMBT	WCRFT s2	Concraft 5.0	Rank
1	97.22%	96.74%	97.16%	97.13%	1st
2	97.19%	96.70%	97.11%	97.06%	1st
3	97.16%	96.73%	97.07%	97.00%	1st
4	97.28%	96.83%	97.24%	97.08%	1st
5	97.25%	96.73%	97.11%	97.08%	1st
6	97.18%	96.63%	97.02%	96.96%	1st
7	97.21%	96.75%	97.12%	97.05%	1st
8	97.30%	96.79%	97.22%	97.18%	1st
9	97.29%	96.79%	97.22%	97.15%	1st
10	97.23%	96.77%	97.19%	97.15%	1st
μ	97.23%	96.75%	97.15%	97.08%	1st

Table V
ACCURACY LOWER BOUND MEASURED FOR **KNOWN** TOKENS

Split	LT	WMBT	WCRFT s2	Concraft 5.0	Rank
1	97.60%	97.43%	97.55%	97.46%	1st
2	97.53%	97.37%	97.47%	97.35%	1st
3	97.54%	97.35%	97.45%	97.32%	1st
4	97.62%	97.46%	97.60%	97.41%	1st
5	97.60%	97.39%	97.49%	97.39%	1st
6	97.47%	97.21%	97.33%	97.22%	1st
7	97.54%	97.38%	97.47%	97.34%	1st
8	97.63%	97.41%	97.53%	97.46%	1st
9	97.62%	97.41%	97.55%	97.39%	1st
10	97.55%	97.42%	97.52%	97.44%	1st
μ	97.57%	97.38%	97.50%	97.38%	1st

bound (for grammatical class) is calculated.

B. Tagging quality

Table IV presents the observed values of accuracy lower bound across ten runs. The average value (μ) is given in the last row. The experiment shows that Ladder Tagger consistently outperforms all other taggers with respect to grammatical class tagging. The improvement over the second best tagger (WCRFT) corresponds to 2.9% drop in error rate.

We also took the opportunity to measure separate values of accuracy lower bound for two classes of token: known and unknown words. Results for known words (Table V) present the same trends as the overall results: Ladder Tagger consistently outperforms other taggers.

Results observed for unknown words (Table VI) are different. It is evident that Concraft is achieving best results for unknown words. This may be attributed to the advantage given to Concraft by its sophisticated model to deal with unknown words (tag guessing module). Ladder Tagger is placed second or third in the ranking.

V. SUMMARY

This paper presented a part of speech (POS) tagger. Presented POS tagger is based on a set of simple probability mixture models. A list of 25 mixture components is defined. They describe various contexts of the tagged word. Parameters of mixture models are estimated using random-restart hill climbing method. Tagging accuracy of all and known tokens is highest among all tested taggers. Tagging accuracy of unknown tokens (not recognized by tag guessing module) is ranked either second or third.

Table VI
ACCURACY LOWER BOUND MEASURED FOR **UNKNOWN** TOKENS

Split	LT	WMBT	WCRFT s2	Concraft 5.0	Rank
1	85.13%	75.19%	85.13%	86.94%	3rd
2	86.09%	74.80%	85.37%	87.52%	2nd
3	85.48%	77.56%	85.45%	87.14%	2nd
4	86.27%	76.22%	85.40%	86.44%	2nd
5	86.09%	75.64%	84.95%	87.02%	2nd
6	87.02%	76.19%	86.23%	87.76%	2nd
7	86.21%	75.98%	85.63%	87.50%	2nd
8	86.35%	76.28%	86.96%	87.72%	3rd
9	86.58%	76.90%	86.50%	89.45%	2nd
10	86.93%	76.23%	86.63%	87.98%	2nd
μ	86.22%	76.10%	85.83%	87.55%	2nd

The results obtained are encouraging enough to extend the presented approach to cover full positional tagset — and this is a priority for us at the moment. The other worthy line of research is to improve handling of unknown words.

ACKNOWLEDGMENT

This work was financed by Innovative Economy Programme project POIG.01.01.02-14-013/09 (http://www.ipipan.waw.pl/nekst/).

REFERENCES

[1] S. Sharoff, "What is at stake: a case study of Russian expressions starting with a preposition," in *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*. Association for Computational Linguistics, 2004, pp. 17–23.

[2] A. Przepiórkowski, R. L. Górski, M. Łaziński, and P. Pęzik, "Recent developments in the National Corpus of Polish," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*. Valletta, Malta: ELRA, 2010.

[3] B. Vidová-Hladká, "Czech language tagging," Ph.D. dissertation, Charles University, Faculty of Mathematics and Physics, Prague, 2000.

[4] J. Hajič, P. Krbec, P. Květoň, K. Oliva, and V. Petkevič, "Serial combination of rules and statistics: A case study in Czech tagging," in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2001. doi: 10.3115/1073012.1073047 pp. 268–275. [Online]. Available: http://dx.doi.org/10.3115/1073012.1073047

[5] M. Piasecki and G. Godlewski, "Effective architecture of the Polish tagger," in *Text, Speech and Dialogue*, vol. 4188. Brno, Czech Republic: Springer, 2006. doi: 10.1007/11846406_27 pp. 213–220. [Online]. Available: http://dx.doi.org/10.1007/11846406_27

[6] J. Hajič and B. Vidová-Hladká, "Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset," in *Proceedings of the COLING - ACL Conference*. ACL, 1998. doi: 10.3115/980845.980927 pp. 483–490. [Online]. Available: http://dx.doi.org/10.3115/980845.980927

[7] D. Tufiş, "Tiered tagging and combined language models classifiers," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, V. Matousek, P. Mautner, J. Ocelíková, and P. Sojka, Eds. Springer Berlin / Heidelberg, 1999, vol. 1692, pp. 843–843. [Online]. Available: http://dx.doi.org/10.1007/3-540-48239-3_5

[8] S. Acedański, "A morphosyntactic Brill tagger for inflectional languages," in *Advances in Natural Language Processing*, ser. Lecture Notes in Computer Science, H. Loftsson, E. Rögnvaldsson, and S. Helgadóttir, Eds. Springer Berlin / Heidelberg, 2010, vol. 6233, pp. 3–14. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-14770-8_3

[9] A. Radziszewski and T. Śniatowski, "A memory-based tagger for Polish," in *Proceedings of the 5th Language & Technology Conference, Poznań*, 2011.

[10] W. Daelemans, J. Zavrel, A. Van den Bosch, and K. van der Sloot, "MBT: Memory-Based Tagger, version 3.2." ILK, Tech. Rep. 10-04, 2010.

[11] A. Radziszewski, "A tiered CRF tagger for Polish," in *Intelligent Tools for Building a Scientific Information Platform*, ser. Studies in Computational Intelligence, R. Bembenik, Ł. Skonieczny, H. Rybiński, M. Kryszkiewicz, and M. Niezgódka, Eds. Springer Berlin Heidelberg, 2013, vol. 467, pp. 215–230. ISBN 978-3-642-35646-9. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-35647-6_16

[12] J. Waszczuk, "Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language," in *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, Mumbai, 2012, pp. 2789–2804.

[13] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, ser. ACL '96. Stroudsburg, PA, USA: Association for Computational Linguistics, 1996. doi: 10.3115/981863.981904 pp. 310–318. [Online]. Available: http://dx.doi.org/10.3115/981863.981904

[14] A. Przepiórkowski, "A comparison of two morphosyntactic tagsets of Polish," in *Representing Semantics in Digital Lexicography: Proceedings of MONDILEX Fourth Open Workshop*, V. Koseska-Toszewa, L. Dimitrova, and R. Roszko, Eds., Warsaw, 2009, pp. 138–144.

[15] A. Radziszewski and S. Acedański, "Taggers gonna tag: an argument against evaluating disambiguation capacities of morphosyntactic taggers," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, P. Sojka, A. Horák, I. Kopeček, and K. Pala, Eds. Springer Berlin Heidelberg, 2012. doi: 10.1007/978-3-642-32790-2_9 pp. 81–87. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-32790-2_9

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 171–179
DOI: 10.15439/2014F151
ACSIS, Vol. 2

A Developmental Genetic Approach to the cost/time trade-off in Resource Constrained Project Scheduling

Grzegorz Pawiński
Department of Computer Science
Kielce University of Technology
al.1000-lecia P.P. 7
25-314 Kielce, Poland
Email: g.pawinski@tu.kielce.pl

Krzysztof Sapiecha
Faculty of Electrical and Computer Engineering
Cracow University of Technology
ul. Warszawska 24
31-155 Kraków, Poland
Email: krzysztof.sapiecha@gmail.com

***Abstract*—In this paper, the use of Developmental Genetic Programming (DGP) for solving a new extension of the Resource-Constrained Project Scheduling Problem (RCPSP) is investigated. We consider a variant of the problem when resources are only partially available and a deadline is given but it is the cost of the project that should be minimized. RCPSP is a well-known NP-hard problem but in its original formulation it does not take into consideration initial resource workload and it minimises the makespan. Unlike other genetic approaches, where genotypes represent solutions, a genotype in DGP is a procedure that constructs a solution to the problem. Genotypes (the search space) and phenotypes (the solution space) are distinguished and a genotype-to-phenotype mapping (GPM) is used. Thus, genotypes are evolved without any restrictions and the whole search space is explored. The goal of the evolution is to find a procedure constructing the best solution of the problem for which the cost of the project is minimal. The paper presents genetic operators as well as GPM specified for the DGP. Experimental results showed that our approach gives significantly better results compared with methods presented in the literature.**

I. Introduction

THE Resource-Constrained Project Scheduling Problem (RCPSP), attempts to reschedule project tasks efficiently using limited renewable resources minimising the maximal completion time of all activities [Brucker et al., 1998], [Demeulemeester and Herroelen, 1997], [Demeulemeester and Herroelen, 2002]. A single project consists of m tasks which are precedence-related by finish-start relationships with zero time lags. The relationship means that all predecessors have to be finished before a task can be started. To be processed, each activity requires a limited resource R which is unique and therefore it has to perform different activities sequentially.

RCPSP is an NP-complete problem which is computationally very hard [Blazewicz et al., 1983]. The RCPSP occurs frequently, in high scale project management such as software development, power plant building, and military industry projects such as design, development and building of nuclear submarines [Pinedo and Chao, 1999]. The authors in [Möhring et al., 2003] state that it is one of the hardest problems of Operational Research. Results of the investigation in [Hartmann and Briskorn, 2010] showed that the best performing heuristics for solving the RCPSP were the Genetic algorithm (GA) of Hartmann [Hartmann, 1998] and the Tabu search (TS) procedure of Bouleimen and Lecocq [Bouleimen and Lecocq, 2003].

Various RCPSP extensions have been developed for solving practical problems [Hartmann and Briskorn, 2010]. However, there is still free room for research. In this paper we will attack the RCPSP where resources have already got their own schedule (like in a software house). Such tasks cannot be moved. Hence, the resources are available only in particular time periods. The goal is to allocate resources to the project tasks, taking into consideration the availability of resources, in order to minimise the total cost of the project and complete it before a deadline.

The resources can be used only in specified time periods what makes the problem computationally even more complex. To overcome this complexity Genetic programming (GP) [Koza, 1992] will be adopted. GP is a domain-independent method that genetically breeds a population of computer programs to solve a problem. It is an extension of the GA [Hartmann, 1998], [Goldberg, 1989], in which the structures in the population are not fixed-length character strings that encode candidate solutions to a problem, but programs that, when executed, produces the candidate solutions to the problem [Koza and Poli, 2005]. Most GP approaches do not distinguish between a genotype, i.e. a point in a search space, and its phenotype, i.e. a point in a solution space. Developmental genetic programming (DGP) [Koza et al., 2003] makes a distinction between the search space and the solution space. DGP evolves a schedule construction procedure instead of the schedule itself. Thus, genotypes are evolved without any restrictions and the whole search space is explored. The quality of the solution is evaluated on the phenotype after genotype-to-phenotype mapping (GPM). The mapping is critical to the performance of the search process [Keller and Banzhaf, 1999].

To summarize, the purpose of the paper is to introduce a new DGP approach for solving RCPSP when resources are

978-83-60810-58-3/$25.00

partially available. The next section of the paper contains a brief overview of related work. A motivation to the research is given in section 3. Section 4 presents an idea of the adaptation of developmental genetic programming for a solution of the RCPSP. In section 5, computational experiments to evaluate our approach and a comparison with other methods are given. The paper ends with conclusions.

II. Related work and preliminary remarks

Both exact and heuristic methods have been used for solving the RCPSP. Among the first ones, a depth-first branching scheme with dominance criteria and the bounding rules was proposed [Demeulemeester and Herroelen, 1997]. In [Brucker et al., 1998] a branching scheme which starts from a graph representing a set of conjunctions and disjunctions was used. Another method, a tree search algorithm was presented in [Mingozzi et al., 1998]. It is based on a mathematical formulation that uses lower bounds and dominance criteria. However, heuristics have been preferred instead of exact methods due to substantial limitations of these latter ones. In-depth study of the performance of recent RCPSP heuristics can be found in [Kolisch and Hartmann, 2006]. Heuristics described by the authors, include X-pass methods, also known as priority rule based heuristics, classical metaheuristics, such as Genetic algorithms, Tabu search, Simulated annealing (SA), and Ant Colony Optimisation (ACO). They give a performance comparison of these methods as applied to different standard instances sets, generated by ProGen in the PSPLIB [Kolish and Sprecher, 1996]. Two approaches of Tabu Search, for artificially created dataset instances, but based on real-world instances (got from Volvo IT and verified by experienced project manager), were investigated in [Skowronski et al., 2013].

One of the latest review papers on solving RCPSP by exact methods and heuristics may be found in [Deiranlou and Jolai, 2009], where a particular attention was paid to GAs. The authors introduced a new crossover operator and auto-tuning for adjusting the rates of crossover and mutation operators. In [Zamani, 2013], an effective hybrid evolutionary search method which integrates a genetic algorithm with a local search was presented. Two approaches for solving the problem with GAs and GP are given in [Frankola et al., 2008]. The authors achieved good quality results by the use of GAs. With GP, they described a methodology to evolve scheduling heuristics in a small amount of time. DGP [Keller and Banzhaf, 1999], [Koza et al., 2003] is an adaptation of GP [Koza, 1992] to the optimisation problems. The DGP is quite new (from 1999) and has never been applied to the RCPSP. However, it has already been successfully applied in the design of electronic circuits, control algorithms [Koza et al., 2003], strategy algorithms in computer games, the synthesis of embedded systems [Deniziak and Górski, 2008], etc. Many of the human-competitive results that were produced using runs of genetic programming that employed a developmental process are described in [Koza, 2010]. Reinforcement Learning (RL) is another machine learning algorithm that was used for solving the RCPSP. RL determines of how software agents ought to take actions so as to achieve one or more goals. The learning process takes place through interaction with the JABAT environment [Jedrzejowicz and Ratajczak-Ropel, 2014].

According to [Alcaraz and Maroto, 2001], the optimal solution can be achieved by exact procedures only for small projects, usually containing less than 60 tasks and not highly constrained. Moreover, exact methods may require a significant amount of computation time. Therefore heuristic approaches to the implementation of resource allocation optimization algorithms would be desired to enhance the process. For many problems, restrictions are imposed on how the structure of genotype may be created. GP algorithms handle the problems by constrained genetic operators in the manner, which makes them produce only legal individuals. The method achieved respectable results for the generation of efficient programs in different domains, e.g. mathematical calculations, robot control, text recognition, etc. In 36 cases, obtained results were as good as or even better then known solutions [Koza and Poli, 2005]. However, constrained operators create infeasible regions in the search space, also eliminating sequences of genes which may lead to high quality solutions. In the DGP approach the problem does not exists anyway. Because of separating the search space from the solution space, legal as well as illegal genotypes are evolved, while each genotype is mapped onto a legal phenotype. It is worth to notice that the evolution of an illegal genotype may lead to the legal genotype constructing the expected result. Thus, the whole search space is explored.

III. Motivation

Classical RCPSP as well as its extensions presented in the literature do not encompass some practical problems. In the classical approach the goal of optimisation is to minimise the makespan without taking into consideration the project cost. Moreover, resources are assumed to be steadily available during the execution of the whole project. In spite of that, developers in a software house or resources of an enterprise building houses, for example, may have initial workloads when starting a new project. A goal in such cases is a cost/time trade-off, i.e. to minimize the total cost while satisfying time constraints. Therefore, an extension of RCPSP where resources have already got their own schedules and cost/time trade-offs are dominant is necessary. The problems have been addressed in the literature [Ahn and Erenguc, 1998], [Drexl et al., 2000], but not combined. Together, they make the RCPSP computationally very complex. To our best knowledge there was no attempt to deal with the problem. Satisfactory results in the matter will make it possible to efficiently solve more complex real life problems faster and better, for which there currently is not any sufficient solution.

IV. Adaptation of the DGP for the RCPSP

In the DGP genotype and phenotype are distinguished. A genotype in classical GAs represents a target solution, while

in DGP the genotype comprises a procedure that construct a solution of the problem. So, if the target solution (phenotype) is a sequence of tasks with allocated resources, which is usually created by the project manager, then the construction of the solution will be a method of how the project manager selects a resource to allocate for each of the tasks. Therefore, DGP does not evolve a project schedule but the project manager itself. A genotype defines how the project manager uses resource allocation strategies to create a project schedule. During evolution only genotypes are evolved, while the genotype-to-phenotype mapping is used to create phenotypes. In that way, the quality of the phenotype may be evaluated in order to find procedures constructing the best solution.

An evolution process in DGP is similar to other genetic approaches. It starts with an initial population with POP individuals. Subsequently, pairs of individuals are randomly drawn from the population and subject to the operation of crossover and mutation. Then, the fitness of newly created genotypes is calculated. If they satisfy time constraints, they pass the *life test* and are added to the current population. Thus, the population size in each generation is at most $2 \cdot POP$. Finally, the reproduction operator is used POP times to reduce the population to its former size. It selects the best individuals for the new population that replaces the current one. This iterative process is repeated over many generations until a predefined number of generations (GEN) has been reached.

A. Genotypes and phenotypes

In our method a genotype of the project manager is represented by a binary tree that comprises resource allocation strategies and a way of applying them for the activities. The tree edges represent a division of the list of activities into two subgroups, while nodes specify a location of the division ($node_d$) and a resource allocation strategy ($node_s$) (Figure 1). The strategy, which will be assigned to each of the subgroups is specified in an appropriate child of the node. A resource may be assigned according to one of the following strategies:

1) is the fastest,
2) is the cheapest,
3) allows to start the task as soon as possible,
4) allows to finish the task as soon as possible,
5) causes the smallest increase of the project time,
6) causes the smallest increase of the project cost.

The initial population consists of individuals generated randomly by recursively creating nodes until a pre-established maximum $Tree_{height}$ is reached. An increase of the tree height causes doubling the divisions and hence the number of leafs of the tree. Therefore, in order to get all possible variations of strategies, the number of the leafs (1) should be at least the number of activities in the project. At the beginning, the genotype is a full tree, where each node has one of the strategies assigned with the same probability and a random $node_d$, which is inversely proportional to the tree height. However, it has to be verified whether nodes contain improper values of $node_d$. The dividing point cannot be bigger than the number of activities in currently considered subgroup. One of the repairing mechanisms could be a "deleting repair" that removes all children of the invalid node. The process is similar to withering of unused features in live organisms, like in intron splicing [Watson et al., 1992]. But we used a "replacing repair" that replaces the invalid node by any of its children, instead of removing the entire branch. Therefore, more genetic information will be kept in the genotype.

Fig. 1. Tree node, where $node_d$ - dividing point, $node_s$ - decision strategy

$$Leafs_{num} = 2^{Tree_{height}-1} \tag{1}$$

B. Genotype-to-phenotype mapping

A Genotype-to-phenotype mapping (GPM) is used to transform the tree structure into a sequence of decision strategies, corresponding to the project activities. The procedure is shown on Algorithm 1. GPM is done by traversing the tree in the depth-first order starting from the top node (the root). If a node has children, a $node_d$ is used for dividing currently considered project activities into two subgroups and strategies are assigned to them. The left child defines a strategy for the first subgroup (from the first to $node_d$ task in the currently considered group of activities) and the right child for the other (from $node_d$ task to the last task in the currently considered group of activities) (Figure 2a). Then, subgroups of activities are passed to offspring nodes and the process is continued (Figure 2b). As a result, we obtain a sequence of decision strategies (*strat*), each corresponding to the given project activity (Figure 2c).

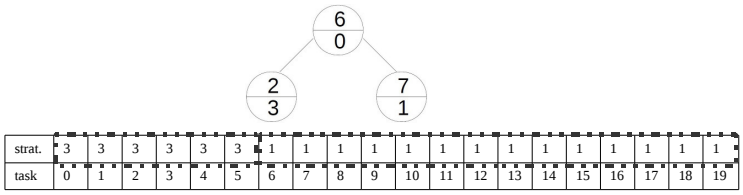
Subsequently, the decision strategies are used to create a project schedule with assigned resources (phenotype). To this end, the following steps have to be carried out:

1) search the project's activities, according to the precedence relationships, in order to find a list of ready-to-start activities,
2) assign the strategies with corresponding activities and execute the strategy to calculate a resource to allocate,
3) calculate a start time for each activity, based on the earliest precedence relationships and the feasible time of a resource,
4) repeat from step 1) until there are activities that are still unassigned.

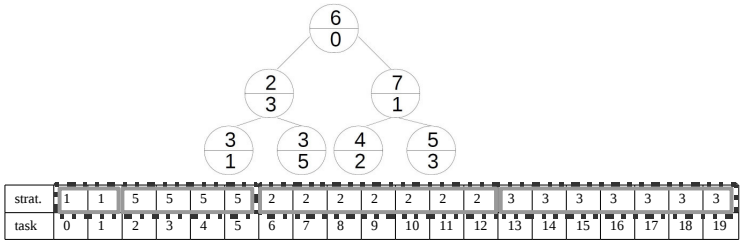
Finally, a feasible project schedule is obtained, which is used for calculating the genotype fitness. It is worth to notice that a genotype is always mapped onto a legal phenotype.

C. Fitness function

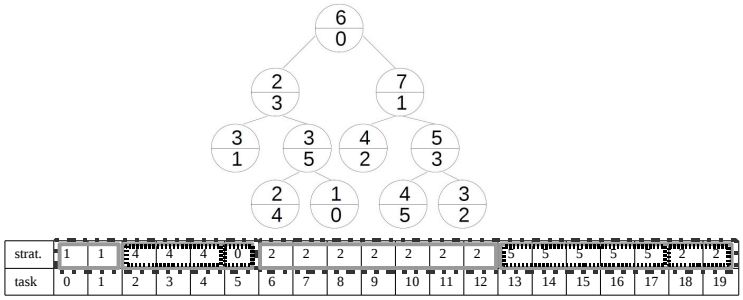
Each individual in the population is measured in order to check the quality of the solution. A numerical value, called fitness, is calculated for the project schedule obtained after



(a) II division



(b) III division



(c) IV division

Fig. 2. Genotype with 6 decision strategies (0-5) represented by a binary tree ($Tree_{height} = 4$) and a sequence of decision strategies (strat.) after the decoding, corresponding to the project activities (task)

Algorithm 1 The procedure of genotype-to-phenotype mapping, where $node$ - currently considered node, i - the index of activity in the list of activities, $size$ - the number of activities in the currently considered group of activities, $strat$ - a sequence of decision strategies, each corresponding to the given project activity

DECODETREE($root, 0, No.of activities$)
function DECODETREE($node$, i, $size$)
  if $node \neq NULL$ **then**
    DECODETREE($left\ child, i, node_d$)
    if $node = Leaf$ **then**
      $strat[i, i + size - 1] \leftarrow node_s$
    end if
    DECODETREE($right\ child, i+node_d, size-node_d$)
  end if
end function

the genotype-to-phenotype mapping. It is defined as the project cost (C), using the following equation:

$$C = \sum_{j=1}^{r} C_e(j) \cdot T_p + \sum_{j=1}^{n} \left(C_u(j) + \sum_{i=1}^{m} T(i,j) \cdot C_e(j) \right) \quad (2)$$

where $C_e(j)$ - cost of task execution per time unit by the resource R_j, T_p - the project duration, $C_u(j)$ - resource R_j unit cost, $T(i,j)$ - safe time estimate of task i being executed by the resource R_j, n - the number of resources used in the project schedule, m - the number of tasks, r - the number of resources in the resource library. The first sum corresponds to the resource maintenance cost in the project. In the second sum, the first element is the resource deployment cost and the second is the cost of the task's execution. The algorithm is set to find an individual with minimal fitness, meaning that genotypes that produce lower project cost are considered to be better ones.

D. Genetic operators

The genetic operations that are performed during the run (i.e. crossover, mutation, reproduction) are based on techniques described in [Deniziak and Wieczorek, 2012]. A crossover is applied with the probability $P_{cross} \in [0,1]$ on a randomly selected pairs of individuals. There are at most $\frac{POP}{2}$ such pairs. Next, the decision trees in pairs are pruned by removing a randomly selected edge. Then, subtrees are swapped between both parent genotypes. Similarly, a mutation is applied on each genotype in the population with the

probability $P_{mut} \in [0, 1]$. Afterwards, one of the following modifications, selected with the same probability, is done on the decision tree:

1) a randomly selected node is changed to another,
2) a randomly selected edge is pruned and the subtree is removed,
3) two random nodes are created and added to a randomly selected leaf.

Modifications are done only when the subtree contains more than one node. If the newly created tree is too high, it is pruned to the allowed height. Then, faulty nodes are removed to preserve the correct tree decoding. Implementation of genetic operators ensures that the correct genotype-tree structure is always kept. Moreover, they neither break precedence constraints nor produce infeasible schedules.

Reproduction copies the best individuals from the current generation to the next generation. We have tested several variants of reproduction such as the ranking method, proportional selection (roulette-wheel selection) and tournament selection. In the tournament selection some number of individuals (called a tournament size TS_{size}) are randomly chosen from the population and a genotype with the lowest fitness is selected. The chance of the individual's being selected in the roulette-wheel method is inversely proportional to its fitness. It is similar in ranking selection, but selection probabilities depend on an individual's position in the ranking. Each individual in the population has a numerical rank based on its fitness.

V. Experimental results

A. Test Instances

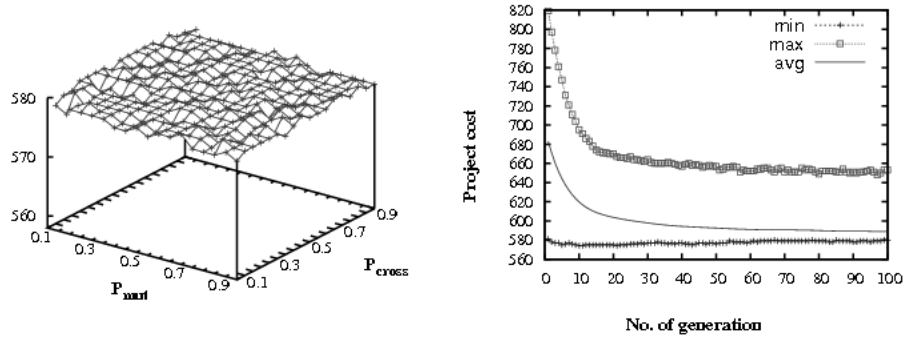
The algorithm described in the paper was tested on projects from PSPLIB, developed by [Kolish and Sprecher, 1996].The library for RCPSP contains 2040 projects with 30, 60, 90 and 120 activities for which either optimal, best-known or lower bound solutions are given. For each problem size, a set consists of 480 instances in groups of ten, which have been systematically generated by varying three parameters: network complexity, resource factor, and resource strength. The parameters have a big impact on the hardness of the project instances [Kolish and Sprecher, 1996]. The set with 30 non-dummy activities is the hardest standard set of RCPSP-instances for which all optimal solutions are currently known [Demeulemeester and Herroelen, 1997]. In our study we used project instances with 30 non-dummy activities. The renewable resources were randomly generated such that the resource development cost $C_u(j)$ and the cost of a activity's execution $C_e(j)$ might vary up to 10% from default values, which were 20 and 1 respectively. They are general purpose resources, so they may execute any of the activities. A single group of the project instances was examined, in which 10 independent runs were performed and the results were averaged. However, we considered an extension of DTCTP where resources have already got their own schedule (initial schedule). To this end, a project instance was randomly drawn from the same group in the PSPLIB. Then, activities from the project were randomly allocated to resources. Such activities cannot be moved and therefore the resources were available only in particular time periods. So even though we take the project instances from PSPLIB, our results cannot be compared with optimal because of a different problem statement. In most tests, the population size POP = 30, the number of generations (GEN) was set to 100 and the Tree height was set to 10, because for bigger values the difference of population sizes has little effect on the results.

B. Main results

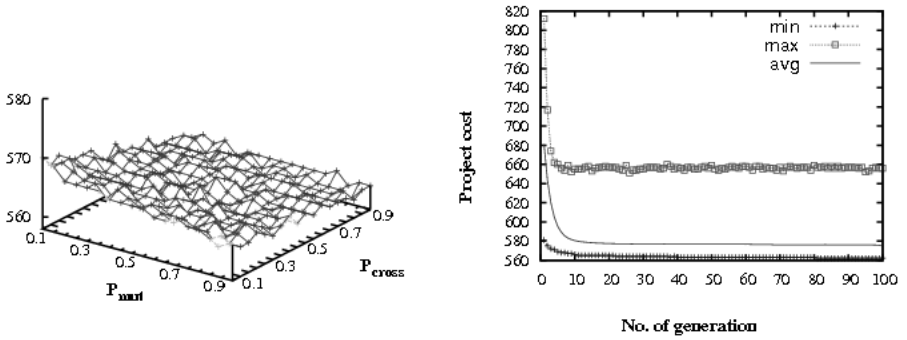
1) The selection method test: At first, we have tested three reproduction methods: ranking, roulette-wheel and tournament selection ($TS_{size} = 3$). The Figures 3 and 4 present the project cost averaged from 100 project schedules. In the roulette-wheel method (Figure 3a), the cost after 100 generations is approximately the same for various probabilities of mutation and crossover. The population contains randomly selected individuals with the probability proportional to their fitness and with no certainty that the best one will be reproduced. At the beginning a population is the most varied and its diversity lowers in further generations. But the best result in every subsequent generation is getting worse. The project cost is almost the same, along with the increase of the probability of both genetic operators, while in other methods it is significantly lower. In other methods the best results were obtained for $P_{cross} + P_{mut} > 1$. In the tournament method (Figure 2b) and ranking method (Figure 3c) there are much more good quality results in generations than bad ones. Good quality results start to dominate in the population very quickly along with the increasing number of generations and therefore the project cost decreases. The convergence of the ranking method is slightly faster than the tournament method. In the former one the rapid fall of the average cost stops after 5 generations while in the latter after 9 generations. Further improvement is very slight, but it occurs till the last generation. However, the best results were obtained for the tournament selection and therefore it was chosen for further examination.

2) Test of various probabilities of genetic operators: Next, tests were executed in order to examine how various probabilities of mutation (Figure 4a and 4b) and crossover (Figure 4c and 4d) influence the algorithm performance. The Figures show the lowest project cost in generations. Usually, the project cost becomes lower along with increasing P_{cross} as well as increasing P_{mut}, because the operators produce more new genotypes and the population is more varied. Thus, the chance of finding the optimal solution is bigger. However, only the best genotypes will be selected to the next generation. The variety of individuals may also be increased by increasing their number in generations. Generally, if the POP is bigger, the results are improved. Yet, the slope of the project cost reduction is similar.

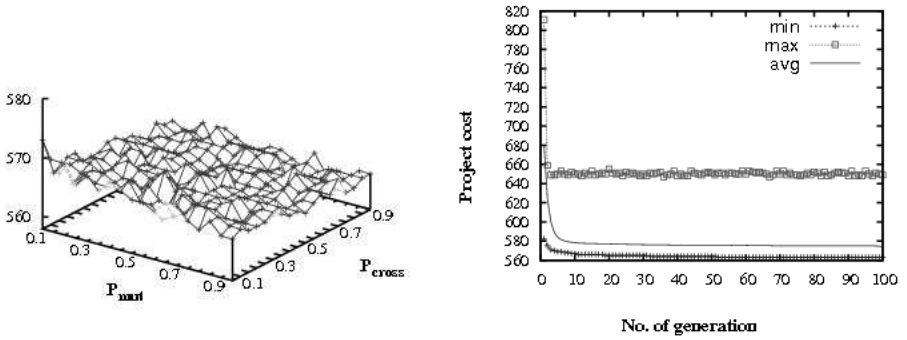
Further tests were performed to study the influence of project time constraints on the final result. Usually, the longer the project time allowed, the more genotypes were passing the *life test* so the population in each generation was bigger



(a) roulette-wheel selection ($POP_{size} = 30$, $Tree_{height} = 10$)



(b) tournament selection ($TS_{size} = 3$, $POP_{size} = 30$, $Tree_{height} = 10$)

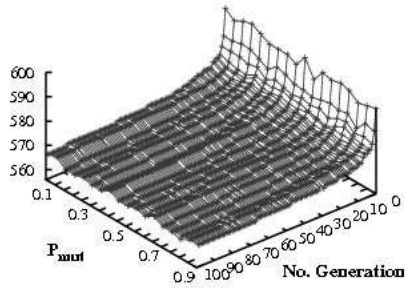


(c) ranking selection ($POP_{size} = 30$, $Tree_{height} = 10$)

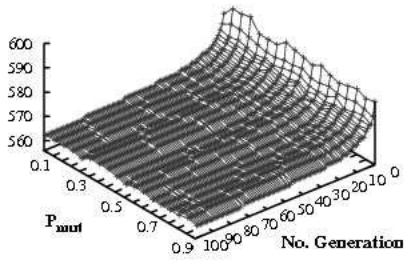
Fig. 3. Minimal cost of the project after 100 generations (left column) and the project cost in generations for $P_{cross} = 0.8$, $P_{mut} = 0.8$ (right column), where min - the lowest project cost, max - the highest project cost, avg - the average project cost, from all individuals of a given generation

and more diverse. Figure 5 shows how a reduction of the deadline affects the distribution of results in generations. The lower the deadline, the lower the number of bad individuals, which were reproduced to next generations. The percentage of newly created genotypes that passed the *life test* fell from 82 (90% of the deadline set) to 36 (60% of the deadline set). However, it did not influence the results. The average project cost, for different time constraints, is close to the best result, which indicates that most individuals in the population correspond to good quality results. Moreover, the convergence of the algorithm is very similar.

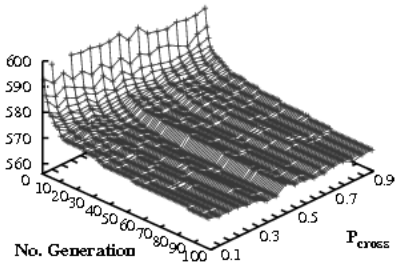
3) Comparison test: Finally, we have performed efficiency test on all 480 project instances where 10 independent runs were computed for each test case. The results were averaged and compared with other methods (Table I). Greedy procedures try to find a resource for each task, in a valid order, according to the smallest increase of the project duration ($Greedy_{time}$) or the project total cost ($Greedy_{cost}$). Another method is a heuristic based on iterative improvements driven by a metric of the gain of optimisation (MAO) [Deniziak, 2004]. It has the capacity of getting out of a local minimum. In [Pawiński and Sapiecha, 2013] the metaheuristic algorithm was adapted to take into account specific features of human resources participating in a project schedule. Their research showed high efficiency of the algorithm for resource allocation. Genetic approaches have similar evolution process



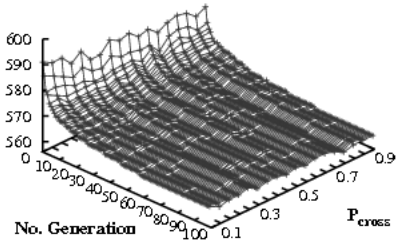
(a) $POP_{size} = 15$, $P_{cross} = 0.8$



(b) $POP_{size} = 30$, $P_{cross} = 0.8$



(c) $POP_{size} = 15$, $P_{mut} = 0.8$



(d) $POP_{size} = 30$, $P_{mut} = 0.8$

Fig. 4. Minimal cost of the project in generations for $Tree_{height} = 10$ and $P_{cross} = 0.8$ (upper row) or $P_{mut} = 0.8$ (lower row)).

TABLE I
THE METHOD COMPARISON RESULTS ($POP = 30, TS_{size} = 3, P_{cross} = 0,8, P_{mut} = 0,8, Tree_{height} = 10$)

Scheduling method	C	T_p	computation time [s]
$Greedy_{time}$	651,89	105,4	70
$Greedy_{cost}$	654,92	106,47	65
MAO	637,94	102,97	4560
GA	619,89	98,57	10062
DGP	599,9	92,69	13650

TABLE II
A COMPARISON OF THE UNCORRECTED SAMPLE STANDARD DEVIATION (S_N)

Scheduling method	S_N
GA	12,83
DGP	4,89

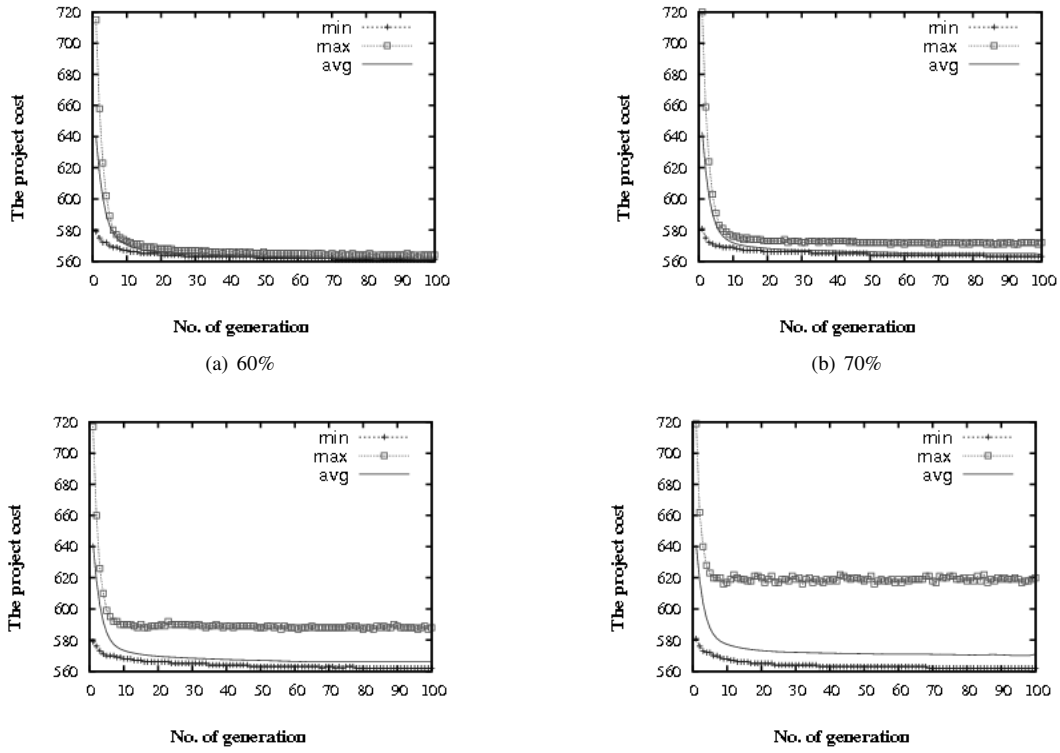
but they differ in a way of coding the genotype. In GA the genotype does not have a tree structure. Genetic operators are performed directly on a sequence of resources corresponding to project activities.

The DGP outperformed the MAO by 5.5% and the greedy methods by 8% as concerns project cost reduction, and by 6% and 12% as concerns project time reduction. Furthermore, the uncorrected sample standard deviation of the DGP was 3-times lower than the deviation of the GA (Table II). However, the DGP was 3-times slower than the MAO and 36% slower than the GA, mainly because of the high number of generations.

VI. CONCLUSIONS

The objective of this research was to introduce and evaluate a new heuristic that can efficiently solve an extension of the RCPSP. It is based on the idea of developmental genetic programming. An algorithm, which was worked out, searches for the best resource allocation strategies in a project. The method of constructing a project schedule takes the form of a decision tree that evolves, instead of evolving the solution itself. The quality of the solution is evaluated after the genotype-to-phenotype mapping.

The fitness function was defined as the project cost. Genetic operators specified for RCPSP were presented as well. To our best knowledge this is the first developmental genetic approach targeting the problem. Three reproduction methods were tested, from which the tournament method ($TS_{size} = 3$) gave the best results. The tournament reproduction ensures that only the best individuals will be selected to the next generation. Then, the influence of various probabilities of mutation and crossover on the algorithm performance was evaluated. Usually, the project cost was lower along with increasing P_{cross} as well as increasing P_{mut}. In the tournament and ranking selection the best results were obtained for $P_{cross} + P_{mut} > 1$. The project cost also decreases as the number of generations increases. Yet, 9 generations is enough



(a) 60% (b) 70%

(c) 80% (d) 90%

Fig. 5. Project cost in each generation for $P_{mut} = 0.8$, $P_{cross} = 0.8$, $POP_{size} = 30$ and with reduced time constraints, where min - the lowest project cost, max - the highest project cost, avg - the average project cost from all individuals of a given generation

to obtain good quality results. Further reduction of the project cost may be achieved by increasing the population size. The influence of the project time constraints on the final results was also tested. The number of bad individuals that were reproduced in subsequent generations was lower when the time constraints were more restricted. But though the genetic operators created 46% less genotypes that passed the *life test*, it did not influence the results. Experimental results showed that DGP is efficient and may be used for solving the extension of the RCPSP. It gives significantly better results than existing methods - it is 5% better than MAO and 8% better than greedy methods.

To our best knowledge this is the first developmental genetic approach targeting the problem. Future work will concentrate on analysing the influence of other parameters of the evolution as well as the influence of different implementations of genetic operators. We will also work on the parallel DGP model to reduce computation time.

REFERENCES

[Ahn and Erenguc, 1998] Ahn, T. and Erenguc, S. (1998). The resource constrained project scheduling problem with multiple crashable modes: A heuristic procedure. *European Journal of Operational Research*, 107(2):250–259. DOI: 10.1016/S0377-2217(97)00331-7

[Alcaraz and Maroto, 2001] Alcaraz, J. and Maroto, C. (2001). A robust genetic algorithm for resource allocation in project scheduling. *Annals of Operations Research*, 102:83-109. DOI: 10.1023/A:1010949931021

[Blazewicz et al., 1983] Blazewicz, J., Lenstra, J. K., and Kan, A. H. G. R. (1983). Scheduling subject to resource constraints: Classification and complexity. *Discrete Applied Mathematics*, 5:11-24.

[Bouleimen and Lecocq, 2003] Bouleimen, K. and Lecocq, H. (2003). A new efficient simulated annealing algorithm for the resource-constrained project scheduling problem and its multiple mode version. *European Journal of Operational Research*, 149(2):268–281. DOI: 10.1016/S0377-2217(02)00761-0

[Brucker et al., 1998] Brucker, P., Knust, S., Schoo, A., and Thiele, O. (1998). A branch-and-bound algorithm for the resource-constrained project scheduling problem. *European Journal of Operational Research*, 107:272-288. DOI: 10.1016/S0377-2217(97)00335-4

[Deiranlou and Jolai, 2009] Deiranlou, M. and Jolai, F. (2009). A new efficient genetic algorithm for project scheduling under resource constrains. *World Applied Sciences Journal*, 7(8):987-997. DOI: 10.1002/nav.10029

[Demeulemeester and Herroelen, 1997] Demeulemeester, E. L. and Herroelen, W. S. (1997). New benchmark results for the resource-constrained project scheduling problem. *Management Science*, 43:1485-1492. DOI: 10.1287/mnsc.43.11.1485

[Demeulemeester and Herroelen, 2002] Demeulemeester, E. L. and Herroelen, W. S. (2002). *Project scheduling - A research handbook*. International Series in Operations Research, Management Science, Boston, MA, USA.

[Deniziak, 2004] Deniziak, S. (2004). Cost-efficient synthesis of multiprocessor heterogeneous systems. *Control and Cybernetics*, 33:341-355.

[Deniziak and Górski, 2008] Deniziak, S. and Górski, A. (2008). Kosynteza systemów soc metoda rozwojowego programowania genetycznego. *Wydawnictwo Politechniki Krakowskiej, in polish*, 105(1-I):19-32.

[Deniziak and Wieczorek, 2012] Deniziak, S. and Wieczorek, K. (2012). Evolutionary optimization of decomposition strategies for logical functions. In *Proceedings of 11th International Conference on Artificial Intelligence and Soft Computing*, volume 7269, page 182-189. Lecture Notes in Computer Science. DOI: 10.1007/978-3-642-29353-5_21

[Drexl et al., 2000] Drexl, A., Patterson, J. H., and Salewski, F. (2000). ProGen/px An instance generator for resource-constrained project scheduling

problems with partially renewable resources and further extensions. *European Journal of Operational Research*, 125:59–72. DOI: 10.1016/S0377-2217(99)00205-2

[Frankola et al., 2008] Frankola, T., Golub, M., and Jakobovic, D. (2008). Evolutionary algorithms for the resource constrained scheduling problem. In *Proceedings of 30th International Conference on Information Technology Interfaces*, volume 7269, page 715-722. Information Technology Interfaces.

[Goldberg, 1989] Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co. and Inc., Boston, MA, USA.

[Hartmann, 1998] Hartmann, S. (1998). A competitive genetic algorithm for resource-constrained project scheduling. *Naval Research Logistics*, 45:733-750. DOI: 10.1002/(SICI)1520-6750(199810)45:7¡733::AID-NAV5¿3.0.CO;2-C

[Hartmann and Briskorn, 2010] Hartmann, S. and Briskorn, D. (2010). Survey of variants and extensions of the resource-constrained project scheduling problem. *European journal of operational research*, 207:1-15. DOI: 10.1016/j.ejor.2009.11.005

[Jedrzejowicz and Ratajczak-Ropel, 2014] Jedrzejowicz, P. and Ratajczak-Ropel, E. (2014). Reinforcement Learning Strategy for Solving the Resource-Constrained Project Scheduling Problem by a Team of A-Teams. *Intelligent Information and Database Systems*, page 197–206. DOI 10.1007/978-3-642-40495-5_46

[Keller and Banzhaf, 1999] Keller, R. E. and Banzhaf, W. (1999). The evolution of genetic code in genetic programing. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 1999)*, page 1077-1082. Information Technology Interfaces.

[Kolisch and Hartmann, 2006] Kolisch, R. and Hartmann, S. (2006). Experimental investigation of heuristics for resource-constrained project scheduling: An update. *European journal of operational research*, 174:23-37. DOI: 10.1016/j.ejor.2005.01.065

[Kolish and Sprecher, 1996] Kolish, R. and Sprecher, A. (1996). Psplib - a project scheduling library. *European journal of operational research*, 96:205-216.

[Koza et al., 2003] Koza, J., Keane, M. A., Streeter, M. J., Mydlowec, W., Yu, J., and Lanza, G. (2003). *Genetic Programming IV: Routine Human-Competitive Machine Intelligence*. Kluwer Academic Publishers.

[Koza, 1992] Koza, J. R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, USA.

[Koza, 2010] Koza, J. R. (2010). Human-competitive results produced by genetic programming. *Genetic Programming and Evolvable Machines*, 11:251-284. DOI 10.1007/s10710-010-9112-3

[Koza and Poli, 2005] Koza, J. R. and Poli, R. (2005). *Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques*. Springer, New York, NY, USA.

[Mingozzi et al., 1998] Mingozzi, A., Maniezzo, V., Ricciardelli, S., and Bianco, L. (1998). An exact algorithm for the resource-constrained project scheduling problem based on a new mathematical formulation. *Management Science*, 44:714-729.

[Möhring et al., 2003] Möhring, R. H., Shulz, A. S., Stork, F., and Utez, M. (2003). Solving project scheduling problems by minimum cut computations. *Management Science*, 49(3):330-350. DOI: 10.1287/mnsc.49.3.330.12737

[Pawiński and Sapiecha, 2013] Pawiński, G. and Sapiecha, K. (2013). Cost-efficient project management based on distributed processing model. In *Proceedings of the 21th International Euromicro Conference on Parallel, Distributed and Network-Based Processing (PDP 2013)*, page 157-163. IEEE Computer Society. DOI: 10.1109/PDP.2013.30

[Pinedo and Chao, 1999] Pinedo, M. and Chao, X. (1999). *Operations Scheduling with applications in Manufacturing*. Irwin/McGraw-Hill, Boston, New York, NY, USA, 2nd edition.

[Skowronski et al., 2013] Skowronski, M. E., Myszkowski, P., Adamski, M. and Kwiatek, P. (2013) Tabu search approach for Multi-Skill Resource-Constrained Project Scheduling Problem. In *Federated Conference on Computer Science and Information Systems (FedCSIS 2013)*, page 153–158. IEEE Computer Society.

[Watson et al., 1992] Watson, J. D., Hopkins, N. H., Roberts, J. W., Steitz, J. A., and Weiner, A. M. (1992). *Molecular Biology of the Gene*. Benjamin Cummings, Menlo Park, CA.

[Zamani, 2013] Zamani, R. (2013). Integrating iterative crossover capability in orthogonal neighborhoods for scheduling resource-constrained projects. *Evolutionary Computation*, 21(2):341–360. DOI: i:10.1162/EVCO_a_00085

International Workshop on Artificial Intelligence in Medical Applications

THE workshop on Artificial Intelligence in Medical Applications – AIMA'2014 - provides an interdisciplinary forum for researchers and developers to present and discuss latest advances in research work as well as prototyped or fielded systems of applications of Artificial Intelligence in the wide and heterogenious field of medicine, health care and surgery. The workshop covers the whole range of theoretical and practical aspects, technologies and systems based on Artificial Intelligence in the medical domain and aims to bring together specialists for exchanging ideas and promote fruitful discussions.

Topics

The topics of interest include, but are not limited to:

- Artificial Intelligence Techniques in Health Sciences
- Knowledge Management of Medical Data
- Data Mining and Knowledge Discovery in Medicine
- Health Care Information Systems
- Clinical Information Systems
- Agent Oriented Techniques in Medicine
- Medical Image Processing and Techniques
- Medical Expert Systems
- Diagnoses and Therapy Support Systems
- Biomedical Applications
- Applications of AI in Health Care and Surgery Systems
- Machine Learning-based Medical Systems
- Medical Data- and Knowledge Bases
- Neural Networks in Medicine
- Ontology and Medical Information
- Social Aspects of AI in Medicine
- Medical Signal and Image Processing and Techniques
- Ambient Intelligence and Pervasive Computing in Medicine and Health Care

Event Chairs

Pancerz, Krzysztof, University of Information Technology and Management in Rzeszów, Poland

Piątek, Łukasz, University of Information Technology and Management in Rzeszów, Poland

Program Committee

Andrushevich, Aliaksei, Lucerne University of Applied Sciences, Switzerland

Bazan, Jan, University of Rzeszów, Poland

Cardoso, Jaime, University of Porto, Portugal

Drahansky, Martin, Brno University of Technology, Czech Republic

Grzymala-Busse, Jerzy, University of Kansas, United States

Hassanien, Aboul Ella, Cairo University, Egypt

Hiroyasu, Tomoyuki, Doshisha University, Japan

Iantovics, Barna, Petru Maior University, Romania

Kountchev, Roumen, Technical Univerity of Sofia, Bulgaria

Krawczyk, Bartosz, Wroclaw University of Technology, Poland

Kumar, Sajeesh, University of Tennessee, Health Science Center, United States

Marchenko, Dmitro, Volodymyr Dahl East Ukrainian National University, Ukraine

Min, Fan, Zhangzhou Normal University, China

Mohyuddin, Mohyuddin, King Abdullah International Medical Research Center, Saudi Arabia

Olszewska, Joanna Isabelle, University of Gloucestershire, United Kingdom

Sawada, Hideyuki, Kagawa University, Japan

Shulgin, Sergiy Kostyantynovych, Volodymyr Dahl East Ukrainian National University, Ukraine

Slezak, Dominik, University of Warsaw & Infobright Inc., Poland

Strzelecki, Michal, Lodz University of Technology, Poland

Wei, Wei, School of Computer science and engineering, Xi'an University of Technology, China

Wysocki, Marian, Rzeszow University of Technology, Poland

Yanushkevich, Svetlana, University of Calgary, Canada

Zaitseva, Elena, University of Zilina, Slovakia

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 183–190
DOI: 10.15439/2014F203
ACSIS, Vol. 2

FUZZY VIKOR APPROACH: EVALUATING QUALITY OF INTERNET HEALTH INFORMATION

Eric Afful-Dadzie
Faculty of Applied Informatics
Tomas Bata University in Zlin
Czech Republic
Email: afful@fai.utb.cz

Stephen Nabareseh
Faculty of Management and Economics.
Tomas Bata University in Zlin.
Czech Republic
Email: nabareseh@fame.utb.cz

Zuzana Komínková Oplatková
Faculty of Applied Informatics
Tomas Bata University in Zlin
Czech Republic
Email: oplatkova@fai.utb.cz

Abstract—Proliferation of health related information on the internet is both welcoming and a concern. For instance when solicited information goes wrong, it tends to have dire consequences on the general public. Assessing the quality of internet health information is often difficult but a rational and systematic approach can be useful in evaluating the quality of the services they render to the public. The paper proposes a fuzzy VIKOR framework for evaluating and ranking internet health information providers under a fuzzy environment where uncertainties and subjectivities are catered for with linguistic variables. Linguistic variables with triangular fuzzy numbers (TFN) are used to evaluate weights of the evaluation criteria and the rankings of each internet health information provider. A numerical example is demonstrated using HIV/AIDS online information providers in the most adult prevalent country in the world. The proposed method is compared with TOPSIS and can be applied in evaluating the quality of other specific internet health information providers.

I. Introduction

HEALTH information, until the advent of the internet, was the exclusive preserve of medical professionals. Today, with high-speed broadband, smart mobile devices and wireless networks, more people rely on the internet for a range of health information support [1], [2]. Users often read about specific medical conditions, communicate in real-time with health care providers via chat rooms and answer health assessment questionnaires online [3], [4]. Majority of the people who search online for health advice do so to be better informed and prepared when consulting their physicians or just to reassure themselves of the status of their health. However, while most of the internet health information comes from authoritative sources such as governmental agencies, research institutions, product vendors, medical centres and individual professionals [45], a lot more of them also come from sources who though well-intentioned, tends to misinform and mislead users. This phenomenon breeds mistrust and presents issues of credibility regarding the source or the websites from which information is sought.

In Korea and China where online health information assists the aged in particular to take good care of themselves by adhering to personal care practices and avoiding illnesses [5], [6] misinformation can be fatal to their health. In the US, there are increasing numbers of citizens managing their health mainly from the information they seek online [7], [8] especially those who are unable to access certain health insurance supports. Such people are vulnerable to misleading information.

The growth in the number of people searching for health related information online has seen a corresponding increase with unregulated sites offering unprofessional advice. Additionally, a study in [9] found that health anxious individuals often do not care about the credibility of an online health information forum provided the information is reassuring and allays their fears. Health anxiety [10], [11], raises fears and often misconception about potential severity of ones' illnesses. In another study on changes occurring in the use of e-health services [12], two thirds of the respondents never checked for assurance of privacy of websites visited and 23% could not recollect the specific name of the site used. Whiles this is frightening, more worrying is that majority of the authors of online health information are not health professionals nor trained to author health information [13], [14], [15].

Subsequently, a number of studies have come out with models and frameworks for assessing the quality of online health information. Some of the notable criteria used in evaluating the quality of internet health information are accuracy, authority, currency, disclaimer, design, and security among others. This study makes a contribution by using fuzzy mathematics and VIKOR multi-criteria decision making (MCDM) technique to demonstrate how online health information providers could be ranked on a number of established criteria. The purpose is to guide users in their choice of websites for health related information. The concept and steps in fuzzy VIKOR are explained and a numerical example is performed using the websites of top 4 HIV/AIDS support organizations in Swaziland to show the usefulness of the technique in ranking health related information providers in any topical area.

II. Fuzzy MCDM

Multi-criteria decision making (MCDM) as a modelling and methodological tool is used to deal with complex decision making problems. MCDM has over the years become one of the most well-known branches of decision making [16], [17] applied in many disciplines. Fuzzy logic has proven to be a

978-83-60810-58-3/$25.00

useful and efficient way in approaching MCDM in situations of imprecise or subjective data in our natural language expression of thoughts and judgements. Since Bellman and Zadeh [18] proposed decision making in fuzzy environment, many extended theories and applications have been carried out to tackle various forms of MCDM. Among few of the Fuzzy MCDM applications are [19] where fuzzy Entropy and t-norm based fuzzy compromise programming is used in locating nuclear power plants in Turkey. In [20], a fuzzy linear programming MCDM model is used in allocating orders to suppliers in a supply chain under uncertainty environment, [21] employed fuzzy MCDM to measure the possibility of successful knowledge management. A hybrid fuzzy MCDM approach based on DEMATEL, ANP and TOPSIS is proposed by [22] to evaluate green suppliers and in [23] a conjunctive MCDM approach also based on DEMATEL, fuzzy ANP, and TOPSIS is modelled as an innovation support system for Taiwanese higher education.

Fuzzy logic has been extended to almost all other MCDM techniques such as Analytic Hierarchy Process (AHP), Analytic Network Process (ANP), ELimination and Choice Expressing REality (ELECTRE), Grey Relational Analysis (GRA), Preference Ranking Organization Method for Enrichment Evaluation (PROMETHEE), Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), Weighted Product Model and VIsekriterijumska optimizacija i KOmpromisno Resenje (VIKOR).

III. FUZZY VIKOR METHOD

VIKOR is a compromise ranking method introduced by Opricovic [24]. The VIKOR method first establishes (1) a compromise ranking-list, (2) a compromise solution, and (3) the weight stability intervals for the compromise solution [24], [25]. It then determines the positive-ideal solution and the negative-ideal solution to aid in ranking and selection [26]. The underlying principle of the VIKOR MCDM method is to deal with ranking and selection of alternatives which have multi-conflicting or non-commensurable criteria [27].

As is usual of most MCDM techniques, the VIKOR method was also extended to accommodate subjectivity and imprecise data under fuzzy environment [28]. A number of applications from various disciplines have been carried out using the fuzzy VIKOR method. In [29], fuzzy VIKOR is used in selecting insurance companies in a group decision making process whiles [30] employed fuzzy VIKOR to resolve multi-criteria decision-making problems. The method is used by [31], [32] for supplier selection problems. In [32], however, the method is modified using entropy measure for objective weighting. In [33] fuzzy VIKOR is utilized for optimized partners' choice in IS/IT outsourcing projects. In [34] the compromise method is used to select renewable energy project in Spain. Similarly in [35] an integrated fuzzy VIKOR and AHP methodology is used to plan renewable energy in Istanbul. In [36] a combined form of fuzzy VIKOR and GRA techniques is utilized to evaluate service quality of airports, [37] applied fuzzy VIKOR for material selection and [38] used fuzzy VIKOR in a robot selection. Again in [39], fuzzy VIKOR based on DEMATEL and ANP is utilized in assessing information security risk control. The literature reviewed portrays the underlying principle of the VIKOR method for selecting and ranking problems but seldom applied in evaluation of service quality.

IV. FUZZY SET THEORY

The human language is filled with imprecision, subjectivities and vagueness when used to judge, describe and communicate information. In view of this, Zadeh [24] introduced the fuzzy set theory to model human judgements. The following are some useful definitions of the fuzzy set theory.

Definition 1: Fuzzy Set. Let X be a nonempty set, the universe of discourse $X=\{x_1, x_2, ..., x_n\}$. A fuzzy set A of X is a set of ordered pairs: $\{(x_1, f_A(x_1)),(x_2, f_A(x_2)),...,(x_n, f_A(x_n))\}$, characterized by a membership function $f_A(x)$ that maps each element x in X to a real number in the interval $[0,1]$. The function value $f_A(x)$ stands for the membership degree of x in A. To capture the vagueness and variations in the subjective ratings of a decision maker, a fuzzy number is used. A Fuzzy number is an expression of membership functions of a linguistic term and ascribe a rating set between the interval [0, 1] for subjective ratings. The two most popular fuzzy numbers are the trapezoidal and triangular fuzzy numbers. In this paper we use the Triangular Fuzzy Number (TFN).

Definition 2: Triangular fuzzy number. A triangular fuzzy number (TFN) is expressed as a triplet (a,b,c). The membership function $f_A(x)$ of a triangular fuzzy number is as defined in Eqn. 1

$$f_A(x)=\begin{cases}0 & x<a, x>b\\ \frac{x-a}{b-a}, & a\le x\le b\\ \frac{c-x}{c-b}, & b\le x\le c\end{cases} \quad (1)$$

Fuzzy models that use TFNs prove to be effective for solving decision-making problems where the available information is subjective and vague [19, 20].

Definition 3: Basic TFN Operations: Assuming $A=(a,b,c)$ and $B=(a_1,b_1,c_1)$ are two TFNs, the basic operations on these two fuzzy triangular numbers are as follows:

$$A\oplus B=(a,b,c)+(a_1,b_1,c_1)=(a+a_1,b+b_1,c+c_1) \quad (2)$$

$$A-B=(a,b,c)-(a_1,b_1,c_1)=(a-c_1,b-b_1,c-a_1) \quad (3)$$

$$A\times B=(a,b,c)\times(a_1,b_1,c_1)=(aa_1,bb_1,cc_1) \quad (4)$$

$$A\div B=(a,b,c)\div(a_1,b_1,c_1)=\left(\frac{a}{c_1},\frac{b}{b_1},\frac{c}{a_1}\right) \quad (5)$$

V. EVALUATING QUALITY OF INTERNET HEALTH INFORMATION

The growing interests and efforts at assessing the quality of health information on the Internet have generated several sets of criteria from a number of sources with little research work on

standardizing such criteria. This study first proposes a new set of criteria for evaluating quality of internet health information culled from several sources [40], [41], [42], [43], [44], [45], [46], [47], [48], [49] as shown in Fig. 1. Secondly, the criteria are used to construct a framework for evaluating the quality of internet health information using fuzzy VIKOR method. We propose that the decision makers be composed of consumer health information experts, self-help group representatives, clinical specialists, general practitioners, lay medical publishers, community health Association representatives, health journalists and information security experts. The criteria used in this study are grouped into four main clusters namely: (a) credibility (b) content (c) design and (d) security. With each cluster having a set of sub-criteria, the total criteria used in this study are fifteen (15). The rationale for selecting the four clusters and their sub-criteria are explained below.

A. Credibility

This cluster examines users' trust in online health information [42], [49]. There are four indicators to measure the credibility of a website providing health information. These are the source, context, relevance and disclosure. The most important criterion for judging the credibility of an online health information provider is the source since it helps to defuse user doubts about the credibility of the information accessed.

B. Content

The content of a website providing health related information is equally deemed important for winning users trust. The sub-criteria are accuracy, currency, disclaimer and authority [43], [45]. Accuracy is often regarded the most important criteria for evaluating "content" and seeks for the scientific validity of the information provided. Users expect proven solutions that are rooted in scientific theory [49].

C. Design

Design defines the quality features and the ease of use of a health information website [43]. Though design does not contribute directly to the quality of information on a website, it is a necessary requirement to ensure frequent delivery of information to users. This is made possible through logical organization of the website information for user understanding [45]. The sub-criteria are accessibility, attractiveness and links.

D. Security

Security is essential in a website providing health related information because of the sensitive and confidential information shared in real-time interactions [49]. Some websites provide chat rooms where users seek advice on a range of issues. It is incumbent on the internet health information provider to assure users of their confidentiality. In this proposed framework, security is measured using caveat together with the CIA triad of confidentiality, integrity and availability. Caveat looks at a website's ability to assure consumers through statements that personal information would not be transferred to third parties or even stored [45]. CIA triad [46], [47] is a widely applied model designed to guide and evaluates information systems security policies. The most obvious element of the CIA triad is confidentiality which ensures that data or an information system is accessed only by authorized persons. Confidentiality is achieved through the protection of user Id's and passwords and other policy based security measures [45].

VI. PROPOSED FUZZY FRAMEWORK

The fuzzy VIKOR approach used in this study is organized in the following order. First, the importance weights of the evaluation criteria are determined and then the performance rating matrix is constructed. Second is the computation of the fuzzy best and worst values of the criteria. Normalized fuzzy difference and the separation values are also computed. Lastly, the triangular fuzzy numbers are defuzzified into crisp values to determine rankings of the alternatives and consequently a compromise solution is proposed.

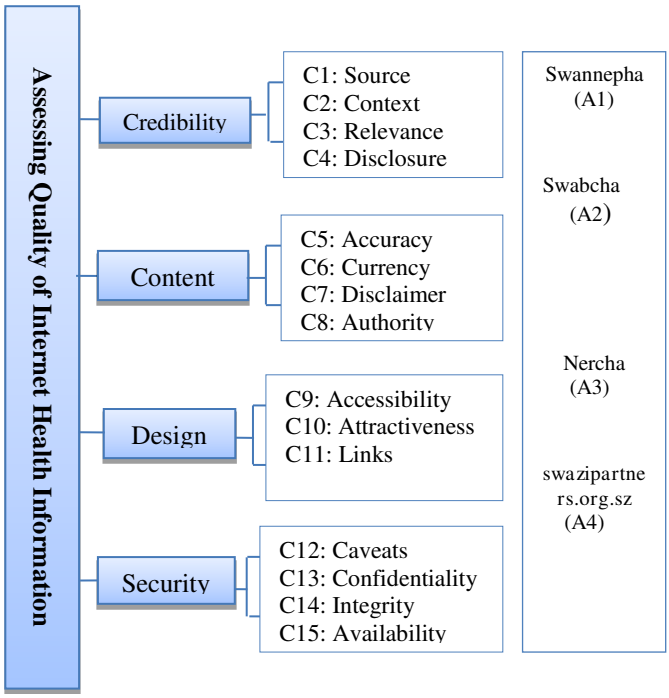


Fig. 1: A framework for evaluating quality of internet Health Information

Step 1: Determining linguistic Variables

The first step in the fuzzy VIKOR method is to determine the linguistic variables; the criteria for evaluating the quality of internet health information. Linguistic terms transformed into fuzzy numbers are used by the experts to rate each linguistic variable. Linguistic terms are qualitative words or phrases of a natural language that reflect the subjective view of an expert about the criteria per each alternative under consideration [50]. In this study, triangular fuzzy numbers are used as shown in Table I and Table II respectively to capture the ratings of the criteria and alternatives on a scale of 0-1.

Table I. Linguistic Scale for the importance of criteria

Linguistic terms	Triangular fuzzy number
Very Low (VL)	(0.0,0.1,0.3)
Low (L)	(0.1,0.3,0.5)
Medium(M)	(0.3,0.5,0.7)
High (H)	(0.5,0.7,0.9)
Very High (VH)	(0.7,0.9,1.0)

Table III. Linguistic scale for ratings of alternatives

Linguistic terms	Triangular fuzzy number
Very Poor (VP)	(0.0,0.0,0.2)
Poor (P)	(0.0,0.2,0.4)
Fair (F)	(0.2,0.4,0.6)
Good (G)	(0.4,0.6,0.8)
Very Good (VG)	(0.6,0.8,1.0)
Excellent (E)	(0.8,0.1,1.0)

Step 2: Determining importance weight of criteria

The evaluation criteria for determining the quality of internet health information providers are supposed to have different importance weights. To determine the importance weight of each criterion, the decision makers rate each criterion using the linguistic terms in Table I. This is expressed in Eq. 6 as vector

$$\tilde{W}: \tilde{W}=\left[\tilde{w}_1,\tilde{w}_2,...,\tilde{w}_n\right] j=1,2,...,n \quad (6)$$

where $\tilde{w}_j$ represents the weight of the jth criterion based on the linguistic preference assigned by a decision maker. Each weight $\tilde{w}_j^k=(w_{j1}^k,w_{j2}^k,w_{j3}^k)$ is expressed as a TFN. These preferences signify the importance attributed to a criterion by a decision maker. The study uses the graded mean integration method [51] to aggregate the decision makers' opinions. The fuzzy importance weight $\tilde{w}_j$ for criterion C_j is computed as:

$$\tilde{w}_j^k=\left(w_{j1},w_{j2},w_{j3}\right) \text{ where, } w_{j1}=\min_k\left\{w_{jk1}\right\},\ w_{j2}=\frac{1}{k}\sum_{k=1}^{k}w_{jk2},$$

$$w_{j3}=\max_k\left\{w_{jk3}\right\} \text{ for } i=1,2,...,m;\ j=1,2,...,n \quad (7)$$

Step 3: Constructing the fuzzy decision matrix

Consider a group of k decision-makers $(D_1,D_2,...,D_k)$ presented with m alternatives $(A_1,A_2,...,A_m)$ against n set of criteria $(C_1,C_2,...,C_n)$ in a typical MCDM problem. A fuzzy multi-criteria decision-making is formally expressed as:

$$\tilde{D}=\begin{array}{c} \\ A_1\\ A_2\\ \vdots\\ A_m\end{array}\begin{array}{c}\begin{array}{cccc} C_1 & C_2 & \dots & C_n\end{array}\\ \begin{bmatrix}\tilde{x}_{11} & \tilde{x}_{12} & \dots & \tilde{x}_{1n}\\ \tilde{x}_{21} & \tilde{x}_{22} & \dots & \tilde{x}_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ \tilde{x}_{m1} & \tilde{x}_{m2} & \dots & \tilde{x}_{mn}\end{bmatrix}\end{array} i=1,2,...,m;\ j=1,2,...,n \quad (8)$$

where, $\tilde{x}_{mn}$ is the rating of alternative A_m with respect to criterion C_j. Note that for a decision maker k, $\tilde{x}_{ij}^k=(a_{ij}^k,b_{ij}^k,c_{ij}^k)$ is a TFN. Similarly as in step 2, the graded mean integration method is used to aggregate the opinions of the decision makers concerning the ratings of the alternatives (websites).This is formally expressed as $\tilde{x}_{ij}^k=\left(a_{ij}^k,b_{ij}^k,c_{ij}^k\right)$ where,

$$a_{ij}=\min_k\left\{a_{ij}^k\right\},b_{ij}=1/k\sum_{k=1}^{k}b_{ij}^k,c_{ij}=\max_k\left\{c_{ij}^k\right\}\ i=1,2,...,m;\ j=1,2,...,n \quad (9)$$

Step 4: Fuzzy best value f_i^* and fuzzy worst value f_i°

The fuzzy best value $f_i^*=\left(a_i^*,b_i^*,c_i^*\right)$ and the fuzzy worst values $f_i^\circ=\left(a_i^\circ,b_i^\circ,c_i^\circ\right)$ are computed respectively using Eq. 10 and 11 below [25], [30].

$$\tilde{f}_i^*=\max_j\tilde{f}_{ij},\ \tilde{f}_i^\circ=\min_j\tilde{a}_{ij}, \text{ for } i\in B \quad (10)$$

$$\tilde{f}_i^*=\min_j\tilde{f}_{ij},\ \tilde{f}_i^\circ=\max_j\tilde{a}_{ij}, \text{ for } i\in C \quad (11)$$

where B is the benefit criteria and C, the cost criteria.

Step 5: Normalized fuzzy difference $\tilde{d}_{ij}$

To obtain the fuzzy difference $\tilde{d}_{ij}$, it is computed as below:

$$\tilde{d}_{ij}=(\tilde{f}_i^*-\tilde{x}_{ij})/(c_i^*-a_i^\circ) \quad \text{for } i\in B \quad (12)$$

$$\tilde{d}_{ij}=(\tilde{x}_{ij}-\tilde{f}_i^*)/(c_i^\circ-a_i^*) \quad \text{for } i\in C \quad (13)$$

where B is the benefit criteria and C, the cost criteria

Step 6: Computing separation Measures $\tilde{S}_j$ and $\tilde{R}_j$

The next step computes the separation $\tilde{S}_j$ of alternative A_j from the fuzzy best value f_i^*. Similarly, the separation of $\tilde{R}_j$ of alternative A_j from the fuzzy worst value f_i° is also computed. These are respectively measured using Eq. 14 and 15:

$$\tilde{S}_j=\sum_{i=1}^{n}(\tilde{w}_j\otimes\tilde{d}_{ij}) \quad (14)$$

$$\tilde{R}_j=\max_i(\tilde{w}_j\otimes\tilde{d}_{ij}) \quad (15)$$

where $\tilde{S}_j=(S_j^a,S_j^b,S_j^c)$ is a fuzzy weighted sum of the separation measure of A_j from the best value f_i^* [27]. Similarly, $\tilde{R}_j=(R_j^a,R_j^b,R_j^c)$ is a fuzzy MAX which refers to the separation measure of A_j from the worst value f_i° where w_j is the importance weight of criterion C_j.

Step 7: Computing the value of $\tilde{Q}_j$

The value $\tilde{Q}_j=(a_j,b_j,c_j)$ expressed in a triangular fuzzy number is computed as following:

$$\tilde{Q}_j=v(\tilde{S}_j-\tilde{S}^*)/(S^{\circ c}-S^{*a})\oplus(1-v)(\tilde{R}_j-\tilde{R}^*)/(R^{\circ c}-R^{*a}) \quad (16)$$

where $\tilde{S}^*=MIN_j\tilde{S}_j,\ S^{\circ c}=MAX_jS_j^c,\ \tilde{R}^*=MIN_j\tilde{R}_j$

$R^{\circ c}=MAX_jR_j^c$ and $v(v=n+1/2n)$ is taken as a weight for the strategy of "majority criteria" (or "maximum utility"), where $1-v$ represents the weight of the individual regret [28]. The best values of S and R are respectively $\tilde{S}^*$ and $\tilde{R}^*$.

Step 8: Defuzifying $\tilde{S}_j$, $\tilde{R}_j$ and $\tilde{Q}_j$

In fuzzy logic, defuzzification is the process of converting the fuzzy numbers into crisp values [50]. The defuzzification is computed by locating the Best Non fuzzy Performance (BNP). A range of defuzzification methods such as Center Of Area (COA), mean of maximum and weighted average method [53] can be used. This paper uses the defuzzification method of COA for ranking fuzzy numbers by [52, 53]. The defuzzification process converts $\tilde{S}_j$, $\tilde{R}_j$ and $\tilde{Q}_j$ into crisp values S, R and Q.

Step 9: Ranking the alternatives

This step ranks the alternatives by sorting the values of S, R and Q in descending order resulting in three ranking lists $\{A\}_S$, $\{R\}_S$ and $\{Q\}_S$ respectively. The index Q_i is the separation measure of A_i from the best alternative. Consequently, the smaller Q_i, the better the alternative.

Step 10: Proposing a Compromise solution

A compromised solution is proposed at this stage where alternative $(A^{(1)})$ is the best ranked by the measure Q (minimum) if the following two conditions are satisfied:

[Condition 1]: Acceptable advantage:

$$Q(A^{(2)}) - Q(A^{(1)}) \geq DQ \quad (17)$$

where $A^{(2)}$ represents the alternative with second position in the ranking list $\{A\}_Q$. Additionally, the threshold $DQ = 1/(n-1)$ where n indicates the number of feasible alternatives.

[Condition 2]: Acceptable stability in decision-making:

The alternative $A^{(1)}$ must be the best ranked by S or/and R. Here if one of these conditions is not satisfied, then a set of compromise solution is proposed consisting of:

1. Alternatives $A^{(1)}$ and $A^{(2)}$ if only condition 2 is not satisfied, or
2. Alternatives $A^{(1)}, A^{(2)}, \ldots, A^{(M)}$ if condition 1 is not satisfied; $A^{(M)}$ is determined by the relation $Q(A^{(M)}) - Q(A^{(1)}) \leq DQ$ for maximum M (the positions of these alternatives are in "closeness").

VII. Numerical Example

This section demonstrates how the fuzzy VIKOR method can be used to evaluate and rank online health information providers. The numerical example in this paper assumes an 8-member decision making team evaluating and ranking the websites of four HIV/AIDS organizations in Swaziland. Swaziland has the world's highest HIV/AIDS adult prevalent rates [54]. The rise in internet rate in Africa leads to reliance on internet for information. In view of this, the quality of information provided by a website on health is crucial. The four websites used in this demonstration are shown in Fig 1. In the following steps, Fuzzy VIKOR is used to demonstrate how to arrive at decision-makers' preferable compromise solution or alternative. The computational illustration of this numerical example is shown as follows:

Step 1: Determining linguistic Variables

The linguistic variables and the alternatives are as shown in Fig. 1. The linguistic terms for the importance weight criteria and the ratings for the alternatives per each criterion used in this paper are as subsequently shown in Table I and Table II.

Step 2: Determining importance weight of criteria

The evaluation is organized into four main clusters comprising 15 sub-criteria for the evaluation of the quality of online health information as shown in Fig 1. This second step in the fuzzy VIKOR MCDM process offers evaluators the chance to choose by rating the most important criteria for the evaluation guided by the linguistic terms in Table I. The linguistic preferences for our assumed eight decision makers concerning the importance attached to each criterion is as shown in Table III below.

Table III. Importance weight of criteria

	D1	D2	D3	D4	D5	D6	D7	D8
C1	VL	M	H	VL	L	L	M	L
C2	M	H	M	M	M	H	H	H
C3	VH	M	H	H	H	VH	M	H
C4	VH	M	H	M	H	M	M	VH
C5	VH	H	VH	VH	VH	H	VH	H
C6	VH	M	H	VH	VH	M	VH	VH
C7	M	M	VH	H	VH	VH	H	VH
C8	L	L	M	H	M	VL	L	L
C9	M	M	H	H	M	L	M	VL
C10	H	M	M	H	H	M	L	H
C11	H	M	H	M	M	H	H	L
C12	H	M	VH	H	VH	VH	M	VH
C13	M	M	M	H	H	H	M	VH
C14	M	H	H	VH	M	L	M	VH
C15	VH	M	M	VH	VH	M	H	H

The graded mean integration method defined in Eq. 7 is used to aggregate the decision makers' opinions regarding the importance weightings of each criterion. The result of such aggregation is shown in Table IV. To determine the importance of each criterion by ranking, the fuzzy numbers are defuzzified. The paper uses the COA (center of area) method in computing the Best Non-Fuzzy Performance value (BNP) to rank the order of importance of each criterion. The BNP value of the fuzzy number $W_k = (L_{wk}, M_{wk}, U_{wk})$ is calculated using the expression in Eq. 18.

$$BNP_{wk} = L_{wk} + [(U_{wk} - L_{wk}) + (M_{wk} - L_{wk})]/3 \quad (18)$$

Table IV. Fuzzy aggregated weights of criterion

	TFN	BNP	Rank
C1	(0.0, 0.35, 0.9)	0.417	15
C2	(0.3, 0.60, 0.9)	0.600	10
C3	(0.3, 0.70, 1.0)	0.667	5
C4	(0.3, 0.65, 1.0)	0.650	7
C5	(0.5, 0.83, 1.0)	0.775	1
C6	(0.3, 0.78, 1.0)	0.692	2
C7	(0.3, 0.74, 1.0)	0.679	4
C8	(0.0, 0.38, 0.9)	0.425	14
C9	(0.0, 0.48, 0.9)	0.458	13
C10	(0.1, 0.58, 0.9)	0.525	11
C11	(0.1, 0.58, 0.9)	0.525	11
C12	(0.3, 0.75, 1.0)	0.683	3
C13	(0.3, 0.63, 1.0)	0.642	8
C14	(0.1, 0.75, 1.0)	0.617	9
C15	(0.3, 0.70, 1.0)	0.667	5

For example, the BNP value for criteria 1 (C1) is computed as follows:

$$=0.0 + [(0.90-0.0) + (0.35-0.0)]/3 = 0.417 \quad (19)$$

By the BNP value computation, the major influential criteria out of the 15 are C5 with a rank of 1 and (C6, C12 and C7) with a rank of 2, 3 and 4 respectively. The least important criterion would be C1 with a rank of 15.

Step 3: Constructing the fuzzy decision matrix

Similarly as in step 2, the decision makers rate the various online health information providers using linguistic terms in Table II. These linguistic judgments would represent the opinions of the evaluators in rating and ranking the four HIV/AIDS organizations. Table V demonstrates assumed ratings of evaluators which have been aggregated using Eq. 9.

Table V. Aggregated fuzzy decision matrix

	A1	A2	A3	A4
C1	(3.18,5.18,7.18)	(4.01,6.01,8.01)	(4.48,6.68,8.68)	(2.46,4.46,6.46)
C2	(3.26,5.18,7.18)	(4.51,6.51,8.43)	(5.34,7.34,9.18)	(2.68,4.68,6.68)
C3	(2.19,4.19,6.19)	(3.85,5.85,7.78)	(4.19,6.19,8.19)	(3.02,5.02,7.02)
C4	(3.21,5.21,7.21)	(4.52,6.50,8.42)	(4.10,6.02,7.85)	(2.19,4.02,6.02)
C5	(3.19,5.19,7.19)	(4.85,6.85,8.69)	(4.53,6.53,8.44)	(3.02,5.02,7.02)
C6	(3.02,5.02,7.02)	(4.69,6.69,8.52)	(5.19,7.19,9.02)	(1.85,3.69,5.69)
C7	(3.10,5.02,7.02)	(4.69,6.69,8.60)	(4.35,6.35,8.35)	(2.85,4.85,6.85)
C8	(4.69,6.69,8.44)	(3.51,5.34,7.34)	(2.51,4.34,6.34)	(2.09,4.01,6.01)
C9	(4.70,6.70,8.61)	(4.36,6.36,8.28)	(3.03,5.03,7.03)	(1.70,3.53,5.53)
C10	(3.08,5.10,7.08)	(4.91,6.91,8.75)	(4.59,6.59,8.50)	(3.25,5.25,7.25)
C11	(3.12,5.12,7.12)	(4.79,6.79,8.62)	(5.29,7.29,9.12)	(1.95,3.79,5.79)
C12	(3.14,5.06,7.06)	(4.73,6.73,8.71)	(4.39,6.39,8.39)	(2.89,4.89,6.89)
C13	(2.58,4.42,6.42)	(2.17,4.09,6.09)	(4.78,6.78,8.52)	(3.58,5.42,7.42)
C14	(3.09,5.09,7.09)	(4.42,6.42,8.34)	(4.76,6.76,8.67)	(1.76,3.59,5.59)
C15	(2.73,5.49,7.34)	(4.45,7.12,8.36)	(2.53,3.78,5.59)	(4.79,6.79,8.68)

Step 4: Fuzzy best value f_i^* and fuzzy worst value f_i°

The study utilizes Eqs. 10 and 11 to determine the fuzzy best and fuzzy worst values for the evaluation criteria. The result of this process is shown in Table VI.

Table VI. Fuzzy best value f_i^* and fuzzy worst value f_i°

Criteria	f_i^*	f_i°
C1	(4.48,6.68,8.68)	(2.46,4.46,6.46)
C2	(5.34,7.34,9.18)	(2.68,4.68,6.68)
C3	(4.19,6.19,8.19)	(2.19,4.19,6.19)
C4	(4.52,6.50,8.42)	(2.19,4.02,6.02)
C5	(4.53,6.53,8.44)	(3.02,5.02,7.02)
C6	(5.19,7.19,9.02)	(1.85,3.69,5.69)
C7	(4.69,6.69,8.60)	(2.85,4.85,6.85)
C8	(4.69,6.69,8.44)	(2.09,4.01,6.01)
C9	(4.70,6.70,8.61)	(1.70,3.53,5.53)
C10	(4.91,6.91,8.75)	(3.08,5.10,7.08)
C11	(5.29,7.29,9.12)	(1.95,3.79,5.79)
C12	(4.73,6.73,8.71)	(2.89,4.89,6.89)
C13	(4.78,6.78,8.52)	(2.58,4.42,6.42)
C14	(4.76,6.76,8.67)	(1.76,3.59,5.59)
C15	(4.79,6.79,8.68)	(2.53,3.78,5.59)

Step 5: Normalized fuzzy difference $\tilde{d}_{ij}$

In this step, the normalized fuzzy difference $\tilde{d}_{ij}$ is computed using Eqs. 12 and 13. For example, $\tilde{d}_{A1}$ is computed as below.

$$\tilde{d}_{A1} = \frac{[(4.48,6.68,8.68)-(3.18,5.18,7.18)]}{8.68\text{-}2.46} \quad (20)$$

$$= \frac{[(4.48\text{-}7.18),(6.68\text{-}5.18),(8.68\text{-}3.18)]}{6.22} = (-0.434, 0.241, 0.884)$$

The rest of the normalized fuzzy differences are calculated in the same manner.

Step 6: Computing separation Measures $\tilde{S}_j$ and $\tilde{R}_j$

The separation measures of $\tilde{S}_j$ and $\tilde{R}_j$ of alternative A_j from the fuzzy best and worst values respectively are computed using Eqs. 14 and 15. The resulting Table VII is as shown below:

Table VII. Index $\tilde{S}_j$ and $\tilde{R}_j$

	A1	A2	A3	A4
C1	(0,0.084,0.795)	(0.0,0.037,0.67)	(0.0,0.0,0.61)	(0.0,0.12,0.90)
C2	(-0.085,1.99,1.0)	(-0.14,0.08,0.65)	(-0.177,0.0,0.53)	(-0.062,0.24,0.9)
C3	(-1.0,2.33,1.0)	(-0.18,0.04,0.72)	(-0.2,0.0,0.67)	(-0.14,0.136,0.86)
C4	(-0.13,0.13,0.84)	(-0.19,0.0,0.63)	(-0.16,0.05,0.69)	(-0.072,0.26,1.0)
C5	(-0.25,0.21,0.97)	(-0.38,-0.05,0.66)	(-0.36,0.0,0.72)	(-0.23,0.23,1.0)
C6	(-0.08,0.24,0.84)	(-0.14,0.05,0.6)	(-0.16,0.0,0.53)	(-0.021,0.38,1.0)
C7	(-0.12,0.21,0.96)	(-0.2,0.0,0.68)	(-0.19,0.04,0.74)	(-0.11,0.24,1.0)
C8	(0.0,0.0,0.53)	(-0.0,0.08,0.7)	(0.0,0.14,0.84)	(0.0,0.16,0.90)
C9	(0.0,0.0,0.51)	(-0.0,0.024,0.55)	(0.0,0.116,0.73)	(0.0,0.22,0.90)
C10	(-0.04,0.185,0.9)	(-0.068,0.0,0.61)	(-0.06,0.03,0.66)	(-0.04,0.17,0.87)
C11	(-0.30,0.17,0.75)	(-0.05,0.04,0.54)	(-0.05,0.0,0.48)	(-0.006,0.28,0.9)
C12	(-0.12,0.21,0.96)	(-0.20,0.0,0.68)	(-0.188,0.04,0.74)	(-0.11,0.288,1.0)
C13	(-0.08,0.25,1.0)	(-0.066,0.28,1.06)	(-0.188,0.0,0.63)	(-0.13,0.14,0.83)
C14	(-0.03,0.18,0.81)	(-0.05,0.04,0.62)	(-0.056,0.0,0.56)	(-0.012,0.34,1.0)
C15	(-0.12,0.14,0.97)	(-0.17,-0.03,0.68)	(-0.039,0.34,1.0)	(-0.189,0.0,0.63)
$\tilde{S}_j$	(-1.185,2.46,12.6)	(-1.84,0.58,10.07)	(-1.84,0.77,10.14)	(-1.13,3.22,13.70)
$\tilde{R}_j$	(0.0,0.25,1.0)	(0.0,0.285,1.07)	(0.0,0.34,1.0)	(0.0,0.38,1.0)

Step 7: Computing the value of $\tilde{Q}_j$

$\tilde{S}^* = (-1.850, 0.588, 10.078)$; $\tilde{R}^* = (0.00, 0.25, 1.00)$, $S^{oc} = 13.69885$; $R^{oc} = 1.069024$.

For example $\tilde{Q}_{jA1}$ is computed using Eq.16 as shown below:

$$\tilde{Q}_{jA1} = \{0.5[(-1.18-10.08, 2.46-0.59, 12.64+1.84)]/(13.69+1.84)\}$$
$$+\{1-0.5[0-1, 0.25-0.25, 1-0]/(1.06-0)\}$$
$$= (-0.82995, 0.06028, 0.93366)$$

By same calculation, the values of the other alternatives are

$\tilde{Q}_{jA2}$=(–0.85130,0.01637,0.88359), $\tilde{Q}_{jA3}$=(–0.85116,0.04898,0.85327)

$\tilde{Q}_{jA4}$=(–0.82831,0.14579,0.96772)

Step 8: Defuzifying $\tilde{S}_j$, $\tilde{R}_j$ and $\tilde{Q}_j$

The defuzzification process converts $\tilde{S}_j$, $\tilde{R}_j$ and $\tilde{Q}_j$ into crisp numbers S, R and Q. The results are shown in Table VIII.

Table VIII. Defuzzified values of S, R and Q

	A1	**A2**	**A3**	**A4**
Q	0.54661	0.016218	0.017031	0.095065
S	4.63946	2.939804	3.021934	5.263415
R	0.416768	0.451442	0.447534	0.460251

Step 9: Ranking the alternatives

The crisp value of the alternatives for Q is ranked from the smallest value to the highest value. The alternatives are ranked as shown in Table IX below.

Table IX. Rank for alternatives

	A1	A2	A3	A4
Q	0.54661	0.016218	0.017031	0.095065
Rank	3	1	2	4

Step 10: Proposing a Compromise solution

In Table IX, the best ranked alternative is A2 which happens to be the best compromise solution. According to the values of Q_j and S_j as shown in Table VIII, the ascending rank of the four HIV/AIDS online information providers in Swaziland is $Q_{A2} \succ Q_{A3} \succ Q_{AI} \succ Q_{A4}$ and $S_{A2} \succ S_{A3} \succ S_{AI} \succ S_{A4}$

Now by the ascending rank order, the HIV/AIDS support organization known as Swabcha (A2), which had the minimum of Q_j and S_j, would be said to have the best quality in terms of provision of online HIV/AIDS information in Swaziland.

VIII. COMPARISON WITH FUZZY TOPSIS

This stage compares the fuzzy VIKOR results from the study with another popular MCDM method called the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS). Fuzzy VIKOR and TOPSIS are both widely used for various selection and ranking solutions. The TOPSIS technique was proposed by [55] but extended to fuzzy TOPSIS by [56]. The technique introduces the shortest distance from the Fuzzy Positive Ideal Solution (FPIS) and the farthest distance from the Fuzzy Negative Ideal Solution (FNIS) simultaneously for the best rank.

In view of this, Fuzzy TOPSIS technique was found ideal in comparison with fuzzy VIKOR since both methods arrive at a scalar (crisp) value in their ranking that considers the best and worst fuzzy values in calculation [56],[57]. They are both found also to be theoretically robust [56]. In Table X, the ranking of the alternatives for both the fuzzy TOPSIS method and the fuzzy VIKOR are presented. The results show that both methods yielded the same order of ranking of the alternatives based on the same data used. Note that in TOPSIS unlike VIKOR, the bigger the value of the relative closeness coefficient, the better the alternative.

Table X: Compared ranking of fuzzy VIKOR and fuzzy TOPSIS results

Alternatives	Fuzzy VIKOR		Fuzzy TOPSIS	
	Results (Q)	Rank	Results (CC_i)	Rank
A1	0.05466	3	0.4600	3
A2	0.0162	1	0.4928	1
A3	0.0170	2	0.4719	2
A4	0.0951	4	0.4085	4

IX. IMPLICATIONS

The growth of the internet means an increase in consumers of online information for a range of purposes. One critical use of the internet is seeking for health information which hitherto was the exclusive preserve of health professionals. Health delivery challenges and shortage of medical professionals in some parts of the world especially in Africa could let people become overly dependent on online health information. To ensure that users access quality online information for improved health, providers of such health related information must be evaluated regularly. To lead in this direction, the proposed fuzzy VIKOR framework could prove handy in ranking health information providers to among other things (1) help users or self-help groups know which websites have the mandate and the competence to educate the public on topical health issues (2) aid health information consumer groups and associations in their resolve to ensuring quality of health information on the internet (3) create competition among specific area health information providers. For example, the evaluation and ranking could introduce competition among diabetes online health information providers or malaria information providers to improve upon their website content and design.

X. CONCLUSION

In this paper, a fuzzy VIKOR framework is proposed for evaluating and ranking internet health information providers. To demonstrate how the framework can be used, a numerical example is carried out using HIV/AIDS organizations in Swaziland who provide internet information related to HIV/AIDS for the Swazis. The organizations used in the study are real organizations providing HIV/AIDS support information and care in Swaziland but the results of the ranking in this paper is just for demonstration purposes.

The study first proposes a new set of criteria for evaluating quality of internet health information. A fuzzy VIKOR framework is then used to demonstrate how this can be carried out experimentally. The results show a methodology that can prove effective in evaluating online health information on any topic. The outcome of results compare favorably to the fuzzy TOPSIS technique justifying its reliability.

ACKNOWLEDGMENT

This work was supported by Internal Grant Agency of Tomas Bata University IGA/FAI/2014/037, IGA/FaME/2014/007 and by the European Regional Development Fund under the project CEBIA-Tech No. CZ.1.05/2.1.00/03.0089.

REFERENCES

[1] A. Suziedelyte, "How does searching for health information on the Internet affect individuals' demand for health care services?." Social science & medicine, vol. 75, no. 10, pp. 1828-1835, Nov. 2012. DOI: 10.1016/j.socscimed.2012.07.022

[2] K. M AlGhamdi, and N.A. Moussa. "Internet use by the public to search for health-related information." International journal of medical informatics, vol. 81, no. 6, pp. 363-373, June 2012. DOI: 10.1016/j.ijmedinf.2011.12.004

[3] S. Fox, and M. Duggan. "Health online 2013." Health, Jan. 2013

[4] Rozmovits, Linda, and Sue Ziebland. "What do patients with prostate or breast cancer want from an Internet site? A qualitative study of information needs." Patient education and counselling, vol. 53, no. 1 pp. 57-64, April 2004. DOI: 10.1016/S0738-3991(03)00116-2

[5] S.J Chang, and I. E.O Im. "A path analysis of Internet health information seeking behaviors among older adults." Geriatric Nursing, Nov.2013. DOI: 10.1016/j.gerinurse.2013.11.005

[6] A. Leung, P. Ko, K.S. Chan, I. Chi, and N. Chow. "Searching health information via the web: Hong Kong Chinese older adults' experience." Public Health Nursing, vol. 24, no. 2, pp. 169-175, Feb. 2007. DOI: 10.1111/j.1525-1446.2007.00621.x

[7] R.A Cohen and B. Stussman. "Health information technology use among men and women aged 18–64: early release of estimates from the National Health Interview Survey, January–June 2009." National Center for Health Statistics, 2010.

[8] Centers for Disease Control and Prevention. "Health Information Technology Use among US Adults", retrieved from: http://www.cdc.gov/features/dshealthinfo/index.html, Oct. 2010.

[9] S.E Baumgartner and T. Hartmann. “The role of health anxiety in online health information search.” Cyberpsychology, Behavior, and Social Networking, vol. 14, no. 10, pp. 613-618, Oct. 2011. DOI: 10.1089/cyber.2010.0425

[10] G. J. G Asmundson, S. Taylor, and B. J. Cox. Health anxiety: Clinical and research perspectives on hypochondriasis and related conditions. Chichester, 2001.

[11] P.M Salkovskis, K.A. Rimes, H. M. C. Warwick, and D. M. Clark. “The Health Anxiety Inventory: development and validation of scales for the measurement of health anxiety and hypochondriasis.” Psychological medicine, vol. 32, no. 5, pp. 843-853, July, 2002. http://dx.doi.org/10.1017/S0033291702005822

[12] E. Sillence, P. Briggs, P. Harris, and L. Fishwick. “Changes in online health usage over the last 5 years.” In CHI'06 Extended Abstracts on Human Factors in Computing Systems, pp. 1331-1336. ACM, 2006. DOI:10.1145/1125451.1125698

[13] M.S Eastin. “Credibility assessments of online health information: The effects of source expertise and knowledge of content. *“Journal* of Computer-Mediated Communication, vol. 6, no. 4, pp. 0-0, Jun, 2001. DOI: 10.1111/j.1083-6101.2001.tb00126.x

[14] Eachus, P. “Health information on the Internet: is quality a problem?.” International journal of health promotion and education, vol. 37, no. 1, pp. 30-33, 1999. DOI:10.1080/14635240.1999.10806089.

[15] L. Theodosiou, and J. Green. “Emerging challenges in using health information from the internet. “Advances in Psychiatric treatment, vol. 9, no. 5, pp. 387-396, 2003. DOI: 10.1192/apt.9.5.387

[16] C. Kahraman. Fuzzy multi-criteria decision making: theory and applications with recent developments. Vol. 16. Springer, 2008.

[17] J. Lu, G. Zhang, and D. Ruan. Multi-objective group decision making: methods, software and applications with fuzzy set techniques. Imperial College Press, 2007.

[18] R.E. Bellman and L.A. Zadeh. “Decision-making in a fuzzy environment.” Management science, vol. 17, no. 4, pp. B-141, Dec. 1970. http://dx.doi.org/10.1287/mnsc.17.4.B141.

[19] İ. Erol, S. Sencer, A. Özmen, and C. Searcy. “Fuzzy MCDM framework for locating a nuclear power plant in Turkey.” Energy Policy, vol. 67, pp. 186-197, Apr. 2013. DOI: 10.1016/j.enpol.2013.11.056.

[20] Haleh, H., and A. Hamidi. "A fuzzy MCDM model for allocating orders to suppliers in a supply chain under uncertainty over a multi-period time horizon." Expert Systems with Applications, vol. 38, no. 8, pp. 9076-9083, Aug. 2011. DOI: 10.1016/j.eswa.2010.11.064.

[21] T.H Chang, and T.C. Wang. “Using the fuzzy multi-criteria decision making approach for measuring the possibility of successful knowledge management.” Information Sciences, vol. 179, no. 4, pp. 355-370, Feb. 2009. DOI: 10.1016/j.ins.2008.10.012.

[22] G. Büyüközkan, and G. Çifçi. “A novel hybrid MCDM approach based on fuzzy DEMATEL, fuzzy ANP and fuzzy TOPSIS to evaluate green suppliers.” Expert Systems with Applications, vol. 39, no. 3, pp. 3000-3011, Feb. 2012. DOI: 10.1016/j.eswa.2011.08.162.

[23] J.K. Chen, and I. Chen. “Using a novel conjunctive MCDM approach based on DEMATEL, fuzzy ANP, and TOPSIS as an innovation support system for Taiwanese higher education.” Expert Systems with Applications, vol. 37, no. 3, pp. 1981-1990, Mar. 2010. DOI: 10.1016/j.eswa.2009.06.079.

[24] S. Opricovic. “Multicriteria optimization of civil engineering systems.” Faculty of Civil Engineering, Belgrade, vol. 2, no. 1, pp. 5-21, 1998.

[25] S. Opricovic, and G.H. Tzeng. “Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS.” European Journal of Operational Research, vol. 156, no. 2, pp. 445-455, Jul. 2004. DOI: 10.1016/S0377-2217(03)00020-1.

[26] M. Wu, and Z. Liu. “The supplier selection application based on two methods: VIKOR algorithm with entropy method and Fuzzy TOPSIS with vague sets method.” International Journal of Management Science and Engineering Management, vol. 6, no. 2, pp. 109-115, May, 2013. DOI:10.1080/17509653.2011.10671152.

[27] T.H. Chang. “Fuzzy VIKOR method: A case study of the hospital service evaluation in Taiwan.” Information Sciences, 2014. DOI: 10.1016/j.ins.2014.02.118.

[28] S. Opricovic. “Fuzzy VIKOR with an application to water resources planning.” Expert Systems with Applications, vol. 38, no. 10, pp. 12983-12990, Sept. 2011. DOI: 10.1016/j.eswa.2011.04.097.

[29] G. N. Yücenur, and N.C. Demirel. “Group decision making process for insurance company selection problem with extended VIKOR method under fuzzy environment.” Expert Systems with Applications, vol. 39, no. 3, pp. 3702-3707, Feb. 2012. DOI: 10.1016/j.eswa.2011.09.065.

[30] T.C. Wang and T. H. Chang. “Fuzzy VIKOR as a resolution for multicriteria group decision-making.” In The 11th International Conference on Industrial Engineering and Engineering Management, pp. 352-356. 2005.

[31] A.S. Sanayei, F. Mousavi, and A. Yazdankhah. “Group decision making process for supplier selection with VIKOR under fuzzy environment.” Expert Systems with Applications, vol. 37, no. 1, pp. 24-30, Jan. 2010. DOI: 10.1016/j.eswa.2009.04.063.

[32] A. Shemshadi, H. Shirazi, M. Toreihi, and M. J. Tarokh. “A fuzzy VIKOR method for supplier selection based on entropy measure for objective weighting.” Expert Systems with Applications, vol. 38, no. 10, pp. 12160-12167, Sept. 2011. DOI: 10.1016/j.eswa.2011.03.027.

[33] L.Y. Chen, and T.C. Wang. “Optimizing partners’ choice in IS/IT outsourcing projects: The strategic decision of fuzzy VIKOR.” International Journal of Production Economics, vol. 120, no. 1, pp. 233-242, Jul. 2009. DOI: 10.1016/j.ijpe.2008.07.022.

[34] J.R. San Cristóbal. “Multi-criteria decision-making in the selection of a renewable energy project in Spain: the Vikor method.” Renewable energy vol. 36, no. 2, pp. 498-502, Feb. 2011. DOI: 10.1016/j.renene.2010.07.031.

[35] T. Kaya, and C. Kahraman. “Multicriteria renewable energy planning using an integrated fuzzy VIKOR & AHP methodology: The case of Istanbul.” Energy, vol. 35, no. 6, pp. 2517-2527, Jun. 2010. DOI: 10.1016/j.energy.2010.02.051.

[36] M.S. Kuo, and G.S. Liang. “Combining VIKOR with GRA techniques to evaluate service quality of airports under fuzzy environment.” Expert Systems with Applications, vol. 38, no. 3, pp. 1304-1312, Mar. 2011. DOI: 10.1016/j.eswa.2010.07.003.

[37] A. Jahan, F. Mustapha, M.Y. Ismail, S. M. Sapuan, and M. Bahraminasab. “A comprehensive VIKOR method for material selection.” Materials & Design, vol. 32, no. 3, pp. 1215-1221, Mar. 2011. DOI: 10.1016/j.matdes.2010.10.015.

[38] K. Devi. “Extension of VIKOR method in intuitionistic fuzzy environment for robot selection.” Expert Systems with Applications, vol. 38, no. 11, pp. 14163-14168, Oct. 2011. DOI: 10.1016/j.eswa.2011.04.227.

[39] Y.P Ou Yang, H.M. Shieh, and G.H. Tzeng. “A VIKOR technique based on DEMATEL and ANP for information security risk control assessment.” Information Sciences, vol. 232, pp. 482-500, May. 2013. DOI: 10.1016/j.ins.2011.09.012.

[40] P. Kim, T.R. Eng, M. J. Deering, and A. Maxfield. “Published criteria for evaluating health related web sites: review.” BMJ: British Medical Journal, vol. 318, no. 7184, pp. 647, Mar. 1999. DOI:/10.1136/bmj.318.7184.647

[41] G. Eysenbach, J. Powell, O. Kuss, and E.R Sa. “Empirical studies assessing the quality of health information for consumers on the World Wide Web: a systematic review.” Jama, vol. 287, no. 20, pp. 2691-2700, May, 2002. DOI:10.1001/jama.287.20.2691.

[42] A.R Jadad, and A. Gagliardi. “Rating health information on the Internet: navigating to knowledge or to Babel?.” Jama, vol. 279, no. 8, pp. 611-614, Feb. 1998. DOI:10.1001/jama.279.8.611

[43] Healthcare Research and Quality (HRQ).“Assessing the Quality of Internet Health Information”. U.S. Department of Health & Human Services. 1999.

[44] American Public Health Association. “Criteria for assessing the quality of health information on the Internet.” American Journal of Public Health, vol. 91, no. 3, pp. 513, Mar. 2001.

[45] J. Ambre et al. criteria for assessing the quality of health information on the Internet. Working draft white paper, Oct. 1997.

[46] D.E.R Denning. Information warfare and security. Vol. 4. Reading MA: Addison-Wesley, 1999.

[47] R. Von Solms. “Information security management: why standards are important.” Information Management & Computer Security, vol. 7, no. 1, pp. 50-58, 1999. DOI:10.1108/09685229910255223

[48] Leonard M. Miller School of Medicine. Confidentiality, Integrity and Availability (CIA). University of Miami. Retrieved from : http://it.med.miami.edu/x904.xml

[49] J.M. Moreno, J. M. Cadenas, S. Alonso, and E. Herrera-Viedma. “An Evaluation Methodology of Quality for Health Web Sites based on Fuzzy Linguistic Modelling.” In Proceedings of IPMU, vol. 8, p. 1091.

[50] Klir, George J., and Bo Yuan. Fuzzy sets and fuzzy logic. Vol. 4. New Jersey: Prentice Hall, 1995.

[51] C. Chou. “The canonical representation of multiplication operation on triangular fuzzy numbers”, Computers & Mathematics with Applications, vol. 45, no.10, pp. 1601-1610, May. 2003. DOI: 10.1016/S0898-1221(03)00139-1.

[52] S. Opricovic, G.H. Tzeng. “Defuzzification within a multicriteria decision model”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 11, no.5, Oct. 2003.

[53] R. Zhao, R. Govind, Algebraic characteristics of extended fuzzy numbers, Information Sciences, 54(1) (1991), 103-130.

[54] Central Intelligence Agency (CIA). “Country Comparison of HIV/AIDS - adult prevalence rates” CIA Library, 2013.

[55] C. L. Hwang, K. Yoon. "Multiple Attribute Decision Making: Methods and Applications, A State of the Art Survey. 1981." Sprinnger-Verlag, New York, NY.

[56] H. Deng, Y. Chung-Hsing, R. J. Willis. "Inter-company comparison using modified TOPSIS with objective weights." Computers & Operations Research 27.10 (2000): 963-973.

[57] H.S. Shih, H.J. Shyur, E. S. Lee. "An extension of TOPSIS for group decision making." Mathematical and Computer Modelling 45.7 (2007): 801-813.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 191–199
DOI: 10.15439/2014F363
ACSIS, Vol. 2

Construction of Healthcare System Structure for Reliability Analysis

Miroslav Kvassay, Elena Zaitseva
University of Zilina,
Faculty of Management Science and Informatics
Zilina, Slovakia
Email: {miroslav.kvassay, elena.zaitseva}@fri.uniza.sk

***Abstract*—The principal goal of application of information technologies in medicine is improvement of medical care. Modern healthcare systems have to guarantee perfect care of patient. Therefore, the healthcare has to be characterized by high reliability first of all and reliability analysis of such system is an important problem. Most of the reliability analysis methods suppose the investigation of the structure of a system. But the healthcare system consists of heterogeneous components, such as software, hardware, human factor and etc. The structure definition for such system is a complex problem. We propose original approach for the construction of a healthcare system structure based on the monitoring of the system behavior. The monitoring examples are interpreted as Direct Partial Logic Derivatives of the structure function of the healthcare system. The structure function defines the correlation between system components states and system performance level unambiguously.**

I. Introduction

THE applications of information technologies, artificial intelligent methods and systems allow to enhance the quality of the healthcare system. But it is not to obliterate medical errors. It is well-known that about 4–7% people die due to medical errors [1], [2]. The development of a healthcare system with high reliability is one of the ways for solving this problem. In general, the basic reliability concept is defined as the probability that a system will perform its intended function during a period of running time without any failure [1]. A fault is an erroneous state of the system. Although the definitions of a fault are different for different systems and in different situations, a fault is always an existing part of the system and it can be removed by correcting the erroneous part of the system. Therefore, the determination of the fault probability is an important problem that can be decided based on system structure analysis [1]–[3].

A healthcare system consists of some principal components from the point of view of reliability analysis. Two of them have been defined in paper [1]: equipment/device and human factor. Detail structure of human factor and human errors for healthcare system is presented in works [2] and [4]. The healthcare system structure includes three components (technical, human and administrative) in [5]. The technical component consists of two parts as hardware and software. The hardware includes two types of medical devices/equipment that are based on special and standards-based technologies according to [1]. For example, the first type includes the medical decision support system, system for integration electronic medical records or picture archiving communication systems. The second type consists of special medical devices and equipment that can be used for special operation only (as magnetic resonance imaging scanners, for example). The human component of the healthcare system models medical errors. The organization component of the system unites management aspects and maintenance of the healthcare system.

Therefore, healthcare systems consist of heterogeneous components that complicate the definition of the mathematical model of their structure.

Two types of mathematical models are used in reliability analysis: a Binary-State System (BSS) and a Multi-State System (MSS). A BSS allows defining only two states in system/component performance – perfect functioning (presented as state 1) and failed (represented by number 0). The mathematical model based on a BSS permits to describe and investigate two types of system behavior that are system failure and system repair [6]. However, real systems have not only two performance levels and, therefore, one of the main problems of a BSS is to define the boundary between situations in which the system can be regarded as working and situations when it is considered failed [7], [8]. A MSS allows avoiding this problem, because it makes possible defines more than only two states in system/component behavior. Therefore, MSSs permit to investigate the system degradation and improvement. This mathematical model can be used for the analysis and determination of conditions that cause the system fault.

The healthcare system is a complex system and its overall performance can have different levels. Therefore, MSSs are appropriate mathematical model for the representation and estimation of the healthcare system. Most methods for MSS

This work was supported by the grant of 7th RTD Framework Program No 610425 (RASimAs) and grant of Slovak Research and Development Agency SK-PL-0023-12.

reliability analysis suppose the structure function as one of initial data [8], [9]. This function defines the system performance level depending on its components states. But the definition of the system structure function with heterogeneous component is a complex problem [5], [8]. One of possible ways for the definition of such system structure function is monitoring the system behavior as the set of the system performance level changes depending on system component state change (Fig. 1).

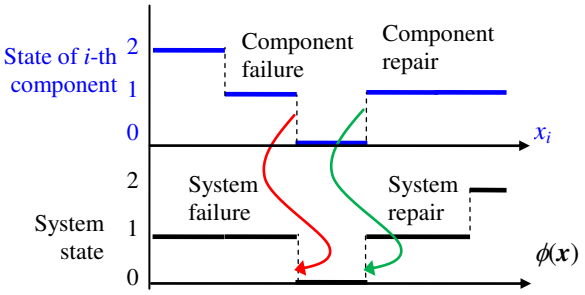


Fig. 1 System reliability interpretations for a MSS

In paper [10] such MSS performance level change depending on the component state change has been defined as Direct Partial Logic Derivative (DPLD) from mathematical point of view. The definition of the DPLD has been proposed in [11] for logic functions. According to this definition, a DPLD with respect to variable x_i allows investigation of the influence of the i-th variable value change on the change of the function value. The application of DPLDs in reliability analysis has been considered in papers [10], [12]. In this paper, we propose a new method for the construction of a MSS structure in form of the structure function by DPLDs. Application and approbation of the method is considered by the example of the typical healthcare system that has been defined in [2].

II. Mathematical Background

A. Structure Function

The correlation between system performance level and states of system components is defined by the structure function. Consider a system that consists of n components and the i-th system component can be in one of m_i states from 0 to m_i -1. Assume m performance levels for the system. Then the structure function of the system is defined as [7], [8]:

$$\phi(\boldsymbol{x}):\{0,\ldots,m_1-1\}\times\ldots\times\{0,\ldots,m_n-1\}\to\{0,\ldots,m-1\}, \quad (1)$$

where $\boldsymbol{x} = (x_1,\ldots, x_n)$ is a vector of system components states (state vector) and state 0 represent the total failure of the system/component while state m -1 (m_i -1) corresponds to perfect functioning of the system (the i-th component).

For a BSS, $m_1 = \ldots = m_n = m = 2$ in (1):

$$\phi(x_1,\ldots,x_n)=\phi(\boldsymbol{x}):\{0,1\}^n\to\{0,1\}. \quad (2)$$

According to definition (2), the BSS structure function can be interpreted as a Boolean function [6].

Every component of a MSS/BSS is characterized by probabilities of its states:

$$p_{i,s}=\Pr\{x_i=s\}, \quad s\in\{0,\ldots,m_i-1\}. \quad (3)$$

Equations (1) and (3), and (2) and (3) constitute the overall definition of a steady-state MSS and BSS.

The typical property of many real systems is their coherency. A system is coherent if it meets the following assumptions [6]–[8]:

a) the system structure function is monotone, i.e. the failure (degradation) of any system component cannot causes the repair (improvement) of system performance level;
b) system components are s-independent, i.e. there is no correlation between states of different components.

Below, we consider only the coherent system.

B. Basic Reliability Characteristics of a System

The basic reliability characteristics of a system are availability and unavailability. The availability is the probability of the system functioning in the fixed time [6]. The unavailability is defined as the probability that the system is failed [6]. Using (2) and (3), the availability of a steady-state BSS can be computed as follows [6]:

$$A=\Pr\{\phi(\boldsymbol{x})=1\}, \quad (4)$$

and its unavailability in the following way [6]:

$$U=\Pr\{\phi(\boldsymbol{x})=0\}=1-A. \quad (5)$$

For a MSS, the availability is defined as probability that the system performance does not fall below given level and, therefore, it should be defined with respect to system performance level [8]:

$$A(j)=\Pr\{\phi(\boldsymbol{x})\geq j\}, \quad j\in\{1,\ldots,m\}. \quad (6)$$

The system unavailability can be defined in the same manner as for a BSS [8]:

$$U=\Pr\{\phi(\boldsymbol{x})=0\}=1-A(1). \quad (7)$$

If we know the mean time between two consecutive breakdowns of a system, i.e. the Mean Time Between Failures (MTBF), then there can be derived other reliability characteristics using system availability and unavailability – the Mean Time To Failure (MTTF) and the Mean Time To Repair (MTTR).

For a BSS, these characteristics can be calculated as the following products [6], [13]:

$$MTTF = A\,MTBF, \quad (8)$$

$$MTTR = U\,MTBF. \quad (9)$$

Other application of the structure function in reliability estimation is importance analysis that can be used to identify the influence of system components on the system performance/availability. There has been proposed a lot of

measures that can be used for this task. These measures are known as importance measures [8], [10], [14] and some of the most commonly used are the Structural Importance (SI), Birnbaum's Importance (BI), Criticality Importance (CI) and Fussell-Vesely Importance (FVI) (Table I).

TABLE I.
IMPORTANCE MEASURES

Importance measures	Meaning
SI	SI concentrates only on the topological structure of the system. It is defined as the relative number of situations in which a given component is critical for the system activity.
BI	BI of a given component is defined as the probability that the component is critical for the system work.
CI	CI of a given component is calculated as the probability that the system failure (degradation) has been caused by the component failure (deterioration) given that the system is failed (not perfect functional).
FVI	FVI of a given component is defined as the probability that the component contributes to the system failure (degradation) probability.

C. Logical Differential Calculus

Logical differential calculus is a special tool that has been developed to analyze dynamic properties of logic functions [11], [15]. In papers [10], [12], there was considered its using in reliability analysis of BSSs and MSSs, respectively. Direct Partial Logic Derivatives (DPLDs) are one of several instruments of logical differential calculus. They allow identifying situations in which the change of a logic variable value coincides with the change of analyzed logic function. In terms of reliability analysis, a DPLD allows finding correlation between component failure/repair and system failure/repair.

For a BSS, we define a DPLD as follows [12]:

$$\partial\phi(j \to \bar{j})/\partial x_i(s \to \bar{s}) = \begin{cases} 1, \text{ if } \phi(s_i, \boldsymbol{x}) = j \text{ and } \phi(\bar{s}_i, \boldsymbol{x}) = \bar{j} \\ 0, \text{ other} \end{cases}, \quad (10)$$

where $\phi(a_i, \boldsymbol{x}) = \phi(x_1,\ldots, x_{i-1}, a, x_{i+1},\ldots, x_n)$ for $a \in \{s, \bar{s}\}$ and $s, j \in \{0, 1\}$.

There exist four DPLDs (10) with respect to the i-th variable and they have the following properties [10], [15]:

$$\begin{aligned} \partial\phi(1 \to 0)/\partial x_i(1 \to 0) &= \partial\phi(0 \to 1)/\partial x_i(0 \to 1), \\ \partial\phi(1 \to 0)/\partial x_i(0 \to 1) &= \partial\phi(0 \to 1)/\partial x_i(1 \to 0). \end{aligned} \quad (11)$$

In the reliability analysis, DPLDs $\partial\phi(1 \to 0)/\partial x_i(1 \to 0)$ and $\partial\phi(0 \to 1)/\partial x_i(0 \to 1)$ can be used to discover situations in which the failure of a given component coincides with the system failure and situations when the repair of the i-th component leads into the repair of the system, respectively. Other two DPLDs make it possible to find situations in which the system failure correlates with the component repair ($\partial\phi(1 \to 0)/\partial x_i(0 \to 1)$) or when the system repair is caused by the component failure ($\partial\phi(0 \to 1)/\partial x_i(1 \to 0)$). However, in this paper only coherent systems are taken into account and therefore there exist no situation in which the component failure can cause system repair or vice versa, and, therefore, these two DPLDs have only zero values for a coherent BSS, which means that they are irrelevant from the reliability point of view [12], [13].

For a MSS, a DPLD with respect to variable i is defined in [10] as follows:

$$\partial\phi(j \to h)/\partial x_i(s \to r) = \begin{cases} 1, \text{ if } \phi(s_i, \boldsymbol{x}) = j \text{ and } \phi(r_i, \boldsymbol{x}) = h \\ 0, \text{ other} \end{cases}, \quad (12)$$

where $\phi(a_i, \boldsymbol{x}) = \phi(x_1,\ldots, x_{i-1}, a, x_{i+1},\ldots, x_n)$ for $a \in \{s, r\}$; s, $r \in \{0,\ldots, m_i - 1\}$, $s \neq r$ and $j, h \in \{0,\ldots, m - 1\}$, $j \neq h$.

From the reliability point of view, the nonzero elements of DPLD (12) identifies situations in which the change of the state of the i-th component from value s to value r causes the change of the system performance level from j to h. Clearly, when $j > h$ and $s > r$ then DPLDs (12) can be used to find correlation between system degradation and component degradation, while for $j < h$ and $s < r$, DPLDs (12) discover situations in which component improvement results system improvement. Specially, when $j > h$ and $s < r$ or $j < h$ and $s > r$, DPLDs (12) discover coincidence between system degradation and component improvement or vice versa.

According to property a) of a coherent system, the degradation of any system component cannot cause system improvement and, therefore, there exist no DPLD (12) that would have nonzero values for $j < h$ and $s > r$ [16]. Property a) of a coherent system can also be interpreted as a statement that the improvement of any system component cannot cause system degradation and this implies that all DPLDs (12) for which $j > h$ and $s < r$ have only zero values [16]. So, the result is that only DPLDs (12) for which $j > h$ and $s > r$ or $j < h$ and $s < r$ can contain nonzero values and therefore only these two types of DPLDs (12) are important from the reliability point of view.

Another point is that we assume that a component of a MSS degrades gradually, i.e. step by step, which means that only the following DPLDs have to be investigated to find correlation between the system deterioration and component degradation [16]:

$$\partial\phi(j \to h)/\partial x_i(s \to s-1), \quad s > 0 \text{ and } j > h. \quad (13)$$

In paper [16], there has been proposed another type of a DPLD for a MSS that is named as DPLD union and that can be defined in the following way:

$$\partial\phi(\downarrow j \downarrow)/\partial x_i(s \to s-1) = \bigcup_{l=j}^{m-1}\left(\bigcup_{h=0}^{j-1}\partial\phi(l \to h)/\partial x_i(s \to s-1)\right) = \begin{cases} 1 & \text{if } \phi(s_i, \boldsymbol{x}) \geq j \text{ and } \phi((s-1)_i, \boldsymbol{x}) < j \\ 0 & \text{other} \end{cases}. \quad (14)$$

This DPLD allows identifying the total influence of the degradation of the *i*-th system component on the system performance level, because it reveals situations in which the analyzed degradation of the component causes the deterioration of the system below specified level *j*.

However, in terms of maintenance, there exist more strategies on how to perform system improvement. Two basic approaches are minor improvement (by one state) and major improvement (by more than one state) [8]. A special type of major improvement is the fully improvement when the component is replaced by a totally new one. The consequences of the minor improvement can be modelled by the following DPLDs:

$$\partial\phi(j \to h)/\partial x_i(s \to s+1) \;\; s < m_i - 1 \text{ and } j < h, \quad (15)$$

and results of the fully improvement as follows:

$$\partial\phi(j \to h)/\partial x_i(s \to m_i - 1), \;\; s < m_i - 1 \text{ and } j < h. \quad (16)$$

In paper [16], the concept of DPLD union (14) was originally developed for the modelling of the system improvement caused by minor improvement of the *i*-th system component:

$$\begin{aligned} &\partial\phi(\uparrow j \uparrow)/\partial x_i(s \to s+1) = \\ &= \bigcup_{l=0}^{j-1}\left(\bigcup_{h=j}^{m-1}\partial\phi(l \to h)/\partial x_i(s \to s+1)\right) = \\ &= \begin{cases} 1 & \text{if } \phi(s_i, \boldsymbol{x}) < j \text{ and } \phi((s+1)_i, \boldsymbol{x}) \geq j \\ 0 & \text{other} \end{cases}. \end{aligned} \quad (17)$$

The union (17) of DPLDs has more informative value than DPLD (15), because it analyzes the total influence of the *i*-th component minor improvement on the system performance level *j*.

D. Minimal Path Vectors and Minimal Cut Vectors

Minimal Path Vectors (MPVs) and Minimal Cut Vectors (MCVs) are special types of state vectors. Firstly, consider two arbitrary state vectors $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$. Then using notation $\boldsymbol{x} < \boldsymbol{y}$ means that $x_i \leq y_i$ for any $i \in \{1, \ldots, n\}$ and there exists at least one i such that $x_i < y_i$.

For a BSS, a MPV represents such situation in which failure of any working component results system failure and a MCV correlates with a situation when the repair of any failed component leads into the system repair. So, a state vector $\boldsymbol{x}$ is a MPV if $\phi(\boldsymbol{x}) = 1$ and $\phi(\boldsymbol{y}) = 0$ for any $\boldsymbol{y} < \boldsymbol{x}$. Similarly, a state vector $\boldsymbol{x}$ is a MCV if $\phi(\boldsymbol{x}) = 0$ and $\phi(\boldsymbol{y}) = 1$ for any $\boldsymbol{y} > \boldsymbol{x}$.

In the case of MSSs, MPVs and MCVs have to be defined with regard to system performance level [8]. A MPV for a given performance level j (for $j = 1, \ldots, m$ -1) of a MSS defines such situation in which degradation of any not-failed component by one state causes degradation of system performance to value less than the specified performance level j, or more formally, a state vector $\boldsymbol{x}$ is a MPV for system performance level j if $\phi(\boldsymbol{x}) \geq j$ and $\phi(\boldsymbol{y}) < j$ for any $\boldsymbol{y} < \boldsymbol{x}$ [8], [16]. In the contrast to a MPV, a MCV for a given performance level of a MSS represents such case when improvement of any not perfect working component by one state results system improvement at least to performance level j (for $j = 1, \ldots, m$ -1), i.e. a state vector $\boldsymbol{x}$ is a MCV for system performance level j if $\phi(\boldsymbol{x}) < j$ and $\phi(\boldsymbol{y}) \geq j$ for any $\boldsymbol{y} > \boldsymbol{x}$ [8], [16].

In papers [16], [17], the calculation of MCVs based on DPLDs has been proposed. For a BSS, there was shown that all MCVs can be computed as the special intersection of all modified (extended) DPLDs $\partial_e\phi(0 \to 1)/\partial_e x_i(0 \to 1)$, i.e. for all $i \in \{1, \ldots, n\}$. The extended DPLD can be derived from DPLD (10) as follows:

$$\begin{aligned} &\partial_e\phi(j \to \bar{j})/\partial_e x_i(s \to \bar{s}) = \\ &= \begin{cases} 1 & \text{if } x_i = s \text{ and } \phi(s_i, \boldsymbol{x}) = j \text{ and } \phi(\bar{s}_i, \boldsymbol{x}) = \bar{j} \\ 0 & \text{if } x_i = s \text{ and } \phi(s_i, \boldsymbol{x}) = \phi(\bar{s}_i, \boldsymbol{x}) \\ * & \text{if } x_i \neq s \end{cases}, \end{aligned} \quad (18)$$

where symbol "*" denotes situations for which DPLD (10) is not defined, i.e. situation when the studied component is in state $\bar{s}$.

The intersection of two modified DPLDs defines situation in which the state change of at least one component causes the required change of the system state. This intersection is defined in Table II.

TABLE II.
THE INTERSECTION OF TWO MODIFIED DPLDS

		$\partial_e\phi(j \to \bar{j})/\partial_e x_i(s \to \bar{s})$		
		*	0	1
$\partial_e\phi(j \to \bar{j})/\partial_e x_k(s \to \bar{s})$	*	*	0	1
	0	0	0	0
	1	1	0	1

MPVs of a BSS can be computed similarly, but we use the extended DPLDs $\partial_e\phi(1 \to 0)/\partial_e x_i(1 \to 0)$ instead of the previous one. Then value 1 in the intersection of all considered extended DPLDs identifies state vectors that are MPVs.

The relation between MCVs and DPLDs, defined in paper [17], was generalized for MSSs in paper [16]. This generalization is based on the using of other types of DPLDs that has been named as the extended union of DPLDs and the merge of extended unions. Using union (17) of DPLDs, the merge of extended unions is defined as follows:

$$\begin{aligned} &\partial_e\phi(\uparrow j \uparrow)/\partial_e x_i = \\ &= \begin{cases} \partial\phi(\uparrow j \uparrow)/\partial x_i(s \to s+1) & \text{if } x_i < m_i - 1 \\ * & \text{if } x_i = m_i - 1 \end{cases}, \end{aligned} \quad (19)$$

and it identifies all situations in which the change of system component state by one value causes the transition of the system performance from level less than j to level greater than or equal to j. Therefore, the intersection of merges (19)

for given system performance level j and for all components identifies all MCVs for system performance level j. The intersection of two merges (19) is defined same as the intersection of two extended DPLDs (Table II).

MPVs for given system performance level j can be computed same as its MCVs, but, instead of merge (19), we use the merge of extended unions that is based on DPLDs (13) and (14) and we denote it as follows:

$$\partial_e \phi(\downarrow j \downarrow)/\partial_e x_i = \begin{cases} \partial\phi(\downarrow j \downarrow)/\partial x_i (s \to s-1) & \text{if } x_i > 0 \\ * & \text{if } x_i = 0 \end{cases}. \quad (20)$$

Value 1 in the intersection of all merges (20) for given system performance level j identifies all MPVs for the considered system performance level.

III. Revelation of the System Structure from Direct Partial Logic Derivatives

The structure function is very important in reliability analysis. However, there exist situations that the structure function of the system is unknown and only the consequences of the system components failure/degradation or repair/improvement on the system performance level are known. In these cases, the structure function of the system has to be discovered to find some reliability characteristics of the system. When we know the results of individual system components changes on the system work, then it means that individual DPLDs are known for the system. Therefore, in the next part, we assume that DPLDs are recognized for the analyzed system and we want to identify its structure function from them.

A. Discovering the Structure Function of a Binary-State System

The structure function (2) of a BSS is formally identical with the definition of a Boolean function. Moreover, the structure function of a coherent BSS can be interpreted as a monotonic Boolean function. Therefore, some concepts of Boolean algebra can be used to find the structure function of a BSS. One of these concepts is the idea that every Boolean function can be expressed unambiguously in the form of minimal disjunctive (conjunctive) normal form. Moreover, for a monotonic Boolean function, there exists only one minimal disjunctive (conjunctive) normal form [18]. When we want to find minimal disjunctive (conjunctive) normal form of a monotonic Boolean function, then all prime implicants (implicates) of this function have to be found.

An implicant of a monotonic Boolean function is a set of logic variables whose simultaneous true values imply that the monotonic Boolean function has true value. A prime implicant is an implicant from which no variable can be removed without losing its status as implicant [18]–[20]. In terms of reliability analysis, an implicant can be interpreted as a set of components whose simultaneous work ensures that the system will work. Also, in terms of reliability analysis, an implicant is known as a path set and prime implicant as a Minimal Path Set (MPS) [21].

An implicate of a monotonic Boolean function is a set of logic variables whose simultaneous false values imply that the considered function has false value [18]. In reliability analysis, the definition of implicate corresponds to cut set and the definition of prime implicate coincides with the definition of Minimal Cut Set (MCS). So, a cut set is a set of components whose simultaneous failure causes the failure of the system and it is minimal if no component can be removed from it without losing its status as a cut set [21].

According to the previous paragraphs, the revelation of MPSs or MCSs of the analyzed BSS implies discovering the system structure function. However, in this paper, we assume that the structure function is defined in terms of vectors and not in terms of set theory. Therefore, MPSs (MCSs) have to be transformed into the form of state vectors. This can be done by using the relation between prime implicants (implicates) and MPSs (MCSs). Every prime implicant (implicate) can be transformed in the vector form that is known as minimal true vector for prime implicant and maximal false vector for prime implicate, respectively [22]. In the terms of reliability engineering, a minimal true vector corresponds to MPV while maximal false vector agrees with a MCV. Therefore, based on paper [22], there exists one-to-one correspondence between MPSs (MCSs) and MPVs (MCVs), i.e. a MPV (MCV) corresponds to a MPS (MCS) in terms of state vectors. This implies that the structure function of any coherent BSS can be defined via MPVs or MCVs unambiguously. So, if DPLDs $\partial\phi(1 \to 0)/\partial x_i(1 \to 0)$ for all $i \in \{1,\ldots,n\}$ are given, then we can formulate the next algorithm for finding the structure function of a BSS:

1. Derive the extended DPLDs $\partial_e\phi(1 \to 0)/\partial_e x_i(1 \to 0)$ from $\partial\phi(1 \to 0)/\partial x_i(1 \to 0)$ for every component, i.e. for $i \in \{1,\ldots,n\}$ according to (18).
2. Compute the intersection of all extended DPLDs from the previous step based on the rules in Table II.
3. Define MPVs that agree to value 1 in the intersection, calculated in the previous step.
4. According to the definition of a MPV (section II.D), define the structure function of the system as follows:
   a. if an arbitrary state vector $\boldsymbol{y}$ meets the condition $\boldsymbol{y} \geq \boldsymbol{x}$ at least for one MPV $\boldsymbol{x}$, then $\phi(\boldsymbol{y}) = 1$;
   b. if an arbitrary state vector $\boldsymbol{y}$ meets the condition $\boldsymbol{y} < \boldsymbol{x}$ at least for one MPV $\boldsymbol{x}$, then $\phi(\boldsymbol{y}) = 0$.

Another approach is based on MCVs. In this case, we assume DPLDs $\partial\phi(0 \to 1)/\partial x_i(0 \to 1)$ for all $i \in \{1,\ldots,n\}$ are known (by the way, according to (11), they are identical to $\partial\phi(1 \to 0)/\partial x_i(1 \to 0)$) and then we can reveal the structure function of the coherent BSS in the following way:

1. Derive the extended DPLDs $\partial_e\phi(0 \to 1)/\partial_e x_i(0 \to 1)$ from $\partial\phi(0 \to 1)/\partial x_i(0 \to 1)$ for every component, i.e. for $i \in \{1,\ldots,n\}$ according to (18).

2. According to Table II, compute the intersection of all extended DPLDs from the previous step.
3. Define MCVs that agree to value 1 in the intersection, calculated in the previous step.
4. According to the definition of a MCV (section II.D), define the structure function of the system as follows:
   a. if an arbitrary state vector $\boldsymbol{y}$ meets the condition $\boldsymbol{y} \leq \boldsymbol{x}$ at least for one MCV $\boldsymbol{x}$, then $\phi(\boldsymbol{y}) = 0$;
   b. if an arbitrary state vector $\boldsymbol{y}$ meets the condition $\boldsymbol{y} > \boldsymbol{x}$ at least for one MCV $\boldsymbol{x}$, then $\phi(\boldsymbol{y}) = 1$.

B. Discovering the Structure Function of a Multi-State System

For MSSs, the structure function (1) is a multiple-valued function. Specially, when $m_1 = \ldots = m_n = m$ in definition (1), then the structure function can be interpreted as a Multiple-Valued Logic (MVL) function. In a case of MVL functions and also multiple-valued functions, there exist some normal forms that are equivalent to minimal disjunctive (conjunctive) normal form of Boolean functions. These forms are known as minimal Sum-of-Products (SoP) and minimal Products-of-Sum (PoS). By the way, a minimal disjunctive (conjunctive) normal form of a Boolean function is sometimes denoted as minimal SoP (PoS), however, in this paper, we used term minimal disjunctive (conjunctive) normal form when we deal with Boolean functions and minimal SoP (PoS) in a case of multiple-valued functions.

There exist several approaches for definition of SoP (PoS) [23], but definitions based on vector approach are the best ones for our work. According to this approach, there exist two important types of vectors for monotone multiple-valued functions: lower vectors (or lower 9 (boundary) points) and upper vectors (or upper (boundary) points) [24].

Lower points for value j of a monotone multiple-valued function of m values are minimal vectors for which the function has value j, for $j = 1,\ldots, m$ -1. When all lower points for every value of a monotone multiple-valued function are known, then the function has value j for an arbitrary vector $\boldsymbol{x}$ if there exists at least one lower point for level j that is lower than or equal to $\boldsymbol{x}$ and at least one lower point for level $j+1$ (given that $j < m$ -1) that is greater than $\boldsymbol{x}$. (A vector $\boldsymbol{x}$ is lower than a vector $\boldsymbol{y}$ if $\boldsymbol{x} < \boldsymbol{y}$ (see section II.D) and greater than $\boldsymbol{y}$ if $\boldsymbol{x} > \boldsymbol{y}$.)

Upper points for value j of a monotone multiple-valued function of m values are maximal vectors for which the function has value less than or equal to j, for $j = 0,\ldots, m$ -2. Therefore, if we know all upper points for every value of the considered function, then the function has value j for an arbitrary vector $\boldsymbol{x}$ if there exists at least one upper point for level j that is greater than or equal to $\boldsymbol{x}$ and at least one upper point for level j -1 (given that $j > 0$) that is lower than $\boldsymbol{x}$.

According to the previous paragraphs, every monotonic multiple-valued function can be defined by using all its upper points or all its lower points. A lower point for value j (for $j = 1,\ldots, m$ -1) of a monotonic multiple-valued function $\phi(\boldsymbol{x})$ is defined as vector $\boldsymbol{x}$ that meets the following conditions [15], [24], [25]:

a) $\phi(\boldsymbol{x}) \geq j$,

b) $\phi(\boldsymbol{y}) < j$ for any $\boldsymbol{y} < \boldsymbol{x}$.

This definition is same as the definition of a MPV of a coherent MSS (section II.D). This fact implies that the structure function of any coherent MSS can be defined via MPVs unambiguously.

An upper point for value j (for $j = 0,\ldots, m$ -2) of a monotonic multiple-valued function $\phi(\boldsymbol{x})$ is defined as vector $\boldsymbol{x}$ that meets the following conditions [15], [24], [25]:

a) $\phi(\boldsymbol{x}) \leq j$,

b) $\phi(\boldsymbol{y}) > j$ for any $\boldsymbol{y} > \boldsymbol{x}$.

The definition of an upper point is not equal to the MCV definition of a coherent MSS (section II.D). This is caused by the fact that upper points are defined for values $j = 0,\ldots, m$ -2, while MCVs are defined for $j = 1,\ldots, m$ -1. Therefore, assume that we do substitution $j+1 = l$ in the definition of the upper point. Due to the substitution, the upper point $\boldsymbol{x}$ for $l = 1,\ldots, m$ -1 has the following properties:

a) $\phi(\boldsymbol{x}) < l$,

b) $\phi(\boldsymbol{y}) \geq l$ for any $\boldsymbol{y} > \boldsymbol{x}$.

These properties are same as the properties of a MCV for level l of a MSS defined by the considered monotonic multiple-valued function. This implies that an upper point for value j of a monotonic multiple-valued function corresponds to a MCV for value $j+1$ of the MSS defined by the considered function. Therefore, there exists one-to-one relation between upper points for level j and MCVs for level $j+1$ of a MSS. So, the knowledge of all MCVs allows defining the structure function of the considered MSS.

According to the aforementioned text, we only need to find all MPVs (MCVs) of the analyzed MSS to reveal the system structure. So, if the consequences of degradation of any component on the system performance are known, i.e. DPLDs $\partial\phi(\downarrow j\downarrow)/\partial x_i(s \to s-1)$ for all $i \in \{1,\ldots, n\}$, $s \in \{1,\ldots, m_i - 1\}$ and $j \in \{1,\ldots, m - 1\}$ are defined, then we can formulate the following algorithm for discovering the system structure function:

1. Derive the merge of extended unions $\partial_e\phi(\downarrow j\downarrow)/\partial_e x_i$ from $\partial\phi(\downarrow j\downarrow)/\partial x_i(s \to s-1)$ for fixed performance level j and for every system component, i.e. for $i \in \{1,\ldots, n\}$ according to (20).
2. According to Table II, compute the intersection of all merges (20) of extended unions from the previous step.
3. Define MPVs for given system performance level j that agree to value 1 in the intersection, calculated in the previous step.
4. Repeat steps 1. – 3. for all relevant system performance levels, i.e. for $j \in \{1,\ldots, m - 1\}$.
5. According to the definition of a MPV of a MSS (section II.D), define the value of the system structure function for an arbitrary state vector $\boldsymbol{y}$ as follows:

- if the state vector $\boldsymbol{y}$ meets the condition $\boldsymbol{y} \geq \boldsymbol{x}$ at least for one MPV $\boldsymbol{x}$ for system performance level m -1, then $\phi(\boldsymbol{y}) = m$ -1;
- else if the state vector $\boldsymbol{y}$ meets the condition $\boldsymbol{y} \geq \boldsymbol{x}$ at least for one MPV $\boldsymbol{x}$ for system performance level m -2, then $\phi(\boldsymbol{y}) = m$ -2;
  …
- else if the state vector $\boldsymbol{y}$ meets the condition $\boldsymbol{y} \geq \boldsymbol{x}$ at least for one MPV $\boldsymbol{x}$ for system performance level 1, then $\phi(\boldsymbol{y}) = 1$;
- else $\phi(\boldsymbol{y}) = 0$.

Similarly as in the case of BSSs, the previous algorithm can be reformulated in the terms of MCVs. In this situation, we assume that all DPLDs $\partial\phi(\uparrow j \uparrow)/\partial x_i(s \to s+1)$ for $i \in \{1,\ldots, n\}$, $s \in \{0,\ldots, m_i - 2\}$ and $j \in \{1,\ldots, m-1\}$ are defined, i.e. the influence of minor improvement of any component on system performance level is known. In this case, the following algorithm can be formulated:

1. Derive the merge of extended unions $\partial_e\phi(\uparrow j \uparrow)/\partial_e x_i$ from $\partial\phi(\uparrow j \uparrow)/\partial x_i(s \to s+1)$ for fixed performance level j and for every component, i.e. for $i \in \{1,\ldots, n\}$ according to (19).
2. According to Table II, compute the intersection of all merges (19) of extended unions that were gained in the previous step.
3. Define MCVs for given system performance level j that agree to value 1 in the intersection, calculated in the previous step.
4. Repeat steps 1. – 3. for all relevant system performance levels, i.e. for $j \in \{1,\ldots, m-1\}$.
5. According to the definition of a MCV of a MSS (section II.D), define the value of the system structure function for an arbitrary state vector $\boldsymbol{y}$ as follows:
   - if there is a MCV $\boldsymbol{x}$ for system performance level 1 that meets the condition $\boldsymbol{y} \leq \boldsymbol{x}$, then $\phi(\boldsymbol{y}) = 0$;
   - else if there is a MCV $\boldsymbol{x}$ for system performance level 2 that meets the condition $\boldsymbol{y} \leq \boldsymbol{x}$, then $\phi(\boldsymbol{y}) = 1$;
     …
   - else if there is a MCV $\boldsymbol{x}$ for system performance level m -1 that meets the condition $\boldsymbol{y} \leq \boldsymbol{x}$, then $\phi(\boldsymbol{y}) = m$ -2;
   - else $\phi(\boldsymbol{y}) = m$ -1.

IV. Reliability Analysis of a Healthcare System with Unknown Structure

Consider the human module of the health care system from book [2]. This module is formed by two persons – a doctor and a nurse, and it defines the consequences of the wrong doctor and nurse behavior on a patient health. In the terms of reliability analysis, the nurse and doctor can be interpreted as two independents modules of the analyzed system. The nurse can perform three types of errors and doctor can also make three types of bad decisions. The wrong decisions are caused by facts that are defined in Table III. These decisions can be interpreted as independent components of the human module and their occurrence can caused the human module degradation that is defined as the deterioration of the patient health.

TABLE III.
COMPONENTS OF THE HUMAN MODULE OF THE HEALTH CARE SYSTEM

System modules	System components	$p_{i,0}$	$p_{i,1}$
Nurse	Correct interpretation of doctor's instructions, x_1	0.01	0.99
	Good work environment, x_2	0.02	0.98
	Not-haste, x_3	0.03	0.97
Doctor	Correct diagnosis, x_4	0.04	0.96
	Good Surroundings, x_5	0.06	0.94
	Not-haste, x_6	0.05	0.95

According to Table III, every component of the human module has two performance levels – failed (an error has occurred) and functioning (a problem has not occurred).

In Table IV, there are defined performance levels of the human module.

TABLE IV.
HUMAN MODULE PERFORMANCE LEVELS

System performance levels	Interpretation
0	Patient received an inadequate amount of wrong medication
1	Patient received an inadequate amount of correct medication
2	Patient received a safe amount of incorrect medication
3	Patient received an adequate amount of correct medication

The results of failures of individual components of the human module on the patient health are defined by DPLDs $\partial\phi(\downarrow j \downarrow)/\partial x_i(1 \to 0)$, for $i \in \{1,\ldots, 6\}$ and $j \in \{1, 2, 3\}$, in Table V. For example, the nonzero element (1,0,0,0,0,0) of DPLD $\partial\phi(\downarrow 1 \downarrow)/\partial x_1(1 \to 0)$ means that the failure of the first component causes the total failure of the analyzed system in situation when all other components are failed.

Now, we want to find the structure function of the human module. According to the previous section, we need to find all MPVs of the considered module. For this task, the first algorithm from section III.B can be used.

In the first step, the merges $\partial_e\phi(\downarrow j \downarrow)/\partial_e x_i$ of extended unions have to be derived from DPLDs defined in Table V. For example, the merges $\partial_e\phi(\downarrow 1 \downarrow)/\partial_e x_i(1 \to 0)$ are calculated in Table IX (white columns). In the next step, their intersection has to be computed to identify MPVs for system performance level 1 (gray columns in Table IX). This procedure has to be repeated for other relevant performance levels of the system, i.e. for $j = 2, 3$. After that, all MPVs of the system are known (Table VI) and we can reveal the system structure function according to rules defined in the

last step of the used algorithm. The discovered structure function of the system is presented in Table VII.

TABLE V.
THE CONSEQUENCES OF COMPONENTS FAILURES ON THE PATIENT HEALTH DEFINED BY DPLDS

Component (i)	**System performance level** (j)	**Nonzero elements of DPLD** $\partial\phi(\downarrow j \downarrow)/\partial x_i(1 \to 0)$
1	1	(1,0,0,0,0,0)
	2	(1,0,0,0,1,1) (1,0,0,1,0,1) (1,0,0,1,1,0) (1,0,0,1,1,1) (1,1,0,1,0,0) (1,1,1,0,0,0) (1,1,1,0,0,1) (1,1,1,0,1,0) (1,1,1,1,0,0)
	3	(1,0,1,1,1,1) (1,1,0,1,1,1) (1,1,1,0,1,1) (1,1,1,1,0,1) (1,1,1,1,1,0)
2	1	(0,1,0,0,0,0)
	2	(0,1,0,0,1,1) (0,1,0,1,1,0) (0,1,0,1,1,1) (0,1,0,1,1,1) (1,1,0,1,0,0) (1,1,1,0,0,0) (1,1,1,0,0,1) (1,1,1,0,1,0) (1,1,1,1,0,0)
	3	(0,1,1,1,1,1) (1,1,0,1,1,1) (1,1,1,0,1,1) (1,1,1,1,0,1) (1,1,1,1,1,0)
3	1	(0,0,1,0,0,0)
	2	(0,0,1,0,1,1) (0,0,1,1,0,1) (0,0,1,1,1,0) (0,0,1,1,1,1) (1,1,1,0,0,0) (1,1,1,0,0,1) (1,1,1,0,1,0)
	3	(0,1,1,1,1,1) (1,0,1,1,1,1) (1,1,1,0,1,1) (1,1,1,1,0,1) (1,1,1,1,1,0)
4	1	(0,0,0,1,0,0)
	2	(0,0,1,1,0,1) (0,0,1,1,1,0) (0,1,0,1,0,1) (0,1,0,1,1,0) (0,1,1,1,0,1) (0,1,1,1,1,0) (1,0,0,1,0,1) (1,0,0,1,1,0) (1,0,1,1,0,1) (1,0,1,1,1,0) (1,1,0,1,0,0) (1,1,0,1,0,1) (1,1,0,1,1,0)
	3	(0,1,1,1,1,1) (1,0,1,1,1,1) (1,1,0,1,1,1) (1,1,1,1,0,1) (1,1,1,1,1,0)
5	1	(0,0,0,0,1,0)
	2	(0,0,1,0,1,1) (0,0,1,1,1,0) (0,1,0,0,1,1) (0,1,0,1,1,0) (0,1,1,0,1,1) (0,1,1,1,1,0) (1,0,0,0,1,1) (1,0,0,1,1,0) (1,0,1,0,1,1) (1,0,1,1,1,0) (1,1,0,0,1,1)
	3	(0,1,1,1,1,1) (1,0,1,1,1,1) (1,1,0,1,1,1) (1,1,1,0,1,1) (1,1,1,1,1,0)
6	1	(0,0,0,0,0,1)
	2	(0,0,1,0,1,1) (0,0,1,1,0,1) (0,1,0,0,1,1) (0,1,0,1,0,1) (0,1,1,0,1,1) (0,1,1,1,1,0) (1,0,0,0,1,0) (1,0,0,1,0,1) (1,0,1,0,1,1) (1,0,1,1,0,0) (1,1,0,0,1,1)
	3	(0,1,1,1,1,1) (1,0,1,1,1,1) (1,1,0,1,1,1) (1,1,1,0,1,1) (1,1,1,1,0,1)

TABLE VI.
THE MPVS OF THE CONSIDERED HUMAN MODULE

System performance level	**MPVs**
1	(0,0,0,0,0,1) (0,0,0,0,1,0) (0,0,0,1,0,0) (0,0,1,0,0,0) (0,1,0,0,0,0) (1,0,0,0,0,0)
2	(0,0,1,0,1,1) (0,0,1,1,0,1) (0,0,1,1,1,0) (0,1,0,0,1,1) (0,1,0,1,0,1) (0,1,0,1,1,0) (0,1,1,0,1,1) (0,1,1,1,1,0) (1,0,0,0,1,1) (1,0,0,1,0,1) (1,0,0,1,1,0) (1,0,1,0,1,1) (1,0,1,1,1,0) (1,1,0,1,0,0) (1,1,1,0,0,0)
3	(0,1,1,1,1,1) (1,0,1,1,1,1) (1,1,0,1,1,1) (1,1,1,0,1,1) (1,1,1,1,0,1) (1,1,1,1,1,0)

Now, we can use the revealed structure function with combination of data from Table III for computation of system availability (6) and unavailability (7). The final values are in Table VIII. According to results in Table VIII, there is very little probability that the considered human module of a health care system will failed.

TABLE VII.
THE STRUCTURE FUNCTION OF THE CONSIDERED HUMAN MODULE

	$x_1\,x_2\,x_3$							
$x_4\,x_5\,x_6$	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
0 0 0	0	1	1	1	1	1	1	2
0 0 1	1	1	1	1	1	1	1	2
0 1 0	1	1	1	1	1	1	1	2
0 1 1	1	2	2	2	2	2	2	3
1 0 0	1	1	1	1	1	1	2	2
1 0 1	1	2	2	2	2	2	2	3
1 1 0	1	2	2	2	2	2	2	3
1 1 1	1	2	2	3	2	3	3	3

TABLE VIII.
AVAILABILITY AND UNAVAILABILITY OF THE CONSIDERED HUMAN MODULE

System performance level (j)	$A(j)$	U
0	-	7.2e-10
1	9.9999e-1	-
2	9.9966e-1	-
3	9.8392e-1	-

V. CONCLUSION

In this paper, a new method for the construction of the structure function based on the DPLDs is considered. This method can be used for the design of the system mathematical model in a case when the initial system has complex structure and correlation between components is not clear defined. The monitoring result of the initial system behavior is interpreted as the set of system performance changes depending on the changes of fixed system component states and these sets are collected for all system components. According to the definition of a DPLD, such sets are interpreted as DPLDs of the structure function. According to the proposed method based on DPLDs, the structure function (1) or (2) of MSS or BSS can be constructed. Then, numerous reliability indices and measures presented in section II.B can be calculated to investigate the initial system.

REFERENCES

[1] B. S. Dhillon, *Medical Device Reliability and Associated Areas*. Boca Raton FLA: CRC Press, 2000, 240 p.
[2] B. S. Dhillon, *Human Reliability and Error in Medical System*. Singapore: World Scientific, 2003, 232 p.
[3] E. Zio, "Reliability engineering: Old problems and new challenges," *Reliability Engineering & System Safety*, vol. 94, no. 2, pp. 125–141, Feb. 2009, http://dx.doi.org/10.1016/j.ress.2008.06.002.
[4] M. Lyons, S. Adams, M. Woloshynowych and Ch. Vincent, "Human reliability analysis in healthcare: A review of techniques," *International Journal of Risk & Safety in Medicine*, vol. 16, no. 4, pp. 223–237, Jan. 2004.

[5] E. Zaitseva and M. Rusin, "Healthcare system representation and estimation based on viewpoint of reliability analysis," *Journal of Medical Imaging and Health Informatics*, vol. 2, no. 1, pp. 80–86, March 2012, http://dx.doi.org/10.1166/jmihi.2012.1067.

[6] M. Xie, Y.-S. Dai and K.-L. Poh, *Computing System Reliability. Models and Analysis.* New York, NY: Kluwer Academic Publishers, 2004, 293 p.

[7] B. Natvig, *Multistate Systems Reliability Theory with Applications*. New York, NY: Wiley, 2011, 262 p., http://dx.doi.org/10.1002/9780470977088.

[8] A. Lisnianski and G. Levitin, *Multi-state System Reliability. Assessment, Optimization and Applications*. Singapore: World Scientific, 2003, 376 p.

[9] A. Lisnianski, I. Frenkel and Y. Ding, *Multi-state System Reliability Analysis and Optimization for Engineers and Industrial Managers*. London, UK: Springer-Verlag London Ltd., 2010, 393 p., http://dx.doi.org/10.1007/978-1-84996-320-6.

[10] E. Zaitseva and V. Levashenko, "Multiple-Valued Logic mathematical approaches for multi-state system reliability analysis," *Journal of Applied Logic*, vol. 11, no. 3, pp. 350–362, Special Issue, 2013, http://dx.doi.org/10.1016/j.jal.2013.05.005.

[11] M. A. Tapia, T. A. Guima and A. Katbab, "Calculus for a multivalued-logic algebraic system," *Applied Mathematics & Computation*, vol. 42, no. 3, pp. 255–285, April 1991, http://dx.doi.org/10.1016/0096-3003(91)90004-7.

[12] E. N. Zaitseva and V. G. Levashenko, "Importance analysis by logical differential calculus," *Automation and Remote Control*, vol. 74, no. 2, pp. 171–182, Feb. 2013, http://dx.doi.org/10.1134/S000511791302001X.

[13] W. G. Schneeweiss, "A short Boolean derivation of mean failure frequency for any (also non-coherent) system," *Reliability Engineering & System Safety*, vol. 94, no. 8, pp. 1363–1367, Aug. 2009, http://dx.doi.org/10.1016/j.ress.2008.12.001.

[14] W. Kuo and X. Zhu, *Importance Measures in Reliability, Risk, and Optimization*. Chichester, UK: John Wiley & Sons, Ltd, 2012, 472 p., http://dx.doi.org/10.1002/9781118314593.

[15] S. N. Yanushkevich, D. M. Miller, V. P. Shmerko and R. S. Stankovic, *Decision Diagram Techniques for Micro- and Nanoelectronic Design. Handbook*. Boca Raton, FL: CRC Press, 2006, 952 p.

[16] M. Kvassay, E. Zaitseva, V. Levashenko and J. Kostolny, "Minimal cut vectors and logical differential calculus," in *Proc. IEEE 44th International Symposium on Multiple-Valued Logic (ISMVL) 2014*, pp. 167–172, http://dx.doi.org/10.1109/ISMVL.2014.37.

[17] E. Zaitseva, J. Kostolny, M. Kvassay, V. Levashenko and K. Pancerz, "Failure analysis and estimation of the healthcare system," in *Proc. Federated Conference on Computer Science and Information Systems (FedCSIS) 2013,* pp. 235–240.

[18] T. Eiter, K. Makino and G. Gottlob, "Computational aspects of monotone dualization: A brief survey," *Discrete Applied Mathematics*, vol. 156, no. 11, pp. 2035–2049, June 2008, http://dx.doi.org/10.1016/j.dam.2007.04.017.

[19] R. B. Cutler and S. Muroga, "Derivation of minimal sums for completely specified functions," *IEEE Transactions on Computers*, vol. C-36, no. 3, pp. 277–292, March 1987, http://dx.doi.org/10.1109/TC.1987.1676900.

[20] P. Jain and G. Gopalakrishnan, "Efficient symbolic simulation-based verification using the parametric form of boolean expressions," *IEEE Transactions On Computer-Aidede Design of Integrated Circuits and Sysytem*, vol. 13, no. 8, pp. 1005–1015, Aug. 1994, http://dx.doi.org/10.1109/43.298036.

[21] M. Rausand and A. Høyland, *System Reliability Theory: Models, Statistical Methods, and Applications*. Haboken, NJ: John Wiley & Sons, Inc., 2004, 664 p.

[22] V. Gurvich and L. Khachiyan, "On generating the irredundant conjunctive and disjunctive normal forms of monotone Boolean functions," *Discrete Applied Mathematics*, vol. 96–97, pp. 363–373, , Oct. 1999, http://dx.doi.org/10.1016/S0166-218X(99)00099-2.

[23] K. Nakashima, Y. Nakamura and N. Takagi, "Logic expressions of monotonic multiple-valued functions," in *Proc. IEEE 26th International Symposium on Multiple-Valued Logic (ISMVL) 1996*, pp. 290–295, http://dx.doi.org/10.1109/ISMVL.1996.508370.

[24] R. A. Boedigheimer and K. C. Kapur, "Customer-driven reliability models for multistate coherent systems," *IEEE Transactions on Reliability,* vol. 43, no. 1, pp. 46–50, March 1994, http://dx.doi.org/10.1109/24.285107.

[25] J. C. Hudson and K. C. Kapur, "Modules in coherent multistate systems," *IEEE Transactions on Reliability,* vol. R-32, no. 2, pp. 183–185, June 1983, http://dx.doi.org/10.1109/TR.1983.5221522.

TABLE IX
THE MERGE (20) OF EXTENDED DPLDS UNIONS FOR LEVEL 1 OF THE CONSIDERED HUMAN MODULE

	$x_1\,x_2\,x_3$															
$x_4\,x_5\,x_6$	0 0 0		0 0 1		0 1 0		0 1 1		1 0 0		1 0 1		1 1 0		1 1 1	
0 0 0	* * * * * *	*	* * 1 * * *	**1**	* 1 * * * *	**1**	* 0 0 * * *	0	1 * * * * *	**1**	0 * 0 * * *	0	0 0 * * * *	0	0 0 0 * * *	0
0 0 1	* * * * * 1	**1**	* * 0 * * 0	0	* 0 * * * 0	0	* 0 0 * * 0	0	0 * * * * 0	0	0 * 0 * * 0	0	0 0 * * * 0	0	0 0 0 * * 0	0
0 1 0	* * * * 1 *	**1**	* * 0 * 0 *	0	* 0 * * 0 *	0	* 0 0 * 0 *	0	0 * * * 0 *	0	0 * 0 * 0 *	0	0 0 * * 0 *	0	0 0 0 * 0 *	0
0 1 1	* * * * 0 0	0	* * 0 * 0 0	0	* 0 * * 0 0	0	* 0 0 * 0 0	0	0 * * * 0 0	0	0 * 0 * 0 0	0	0 0 * * 0 0	0	0 0 0 * 0 0	0
1 0 0	* * * 1 * *	**1**	* * 0 0 * *	0	* 0 * 0 * *	0	* 0 0 0 * *	0	0 * * 0 * *	0	0 * 0 0 * *	0	0 0 * 0 * *	0	0 0 0 0 * *	0
1 0 1	* * * 0 * 0	0	* * 0 0 * 0	0	* 0 * 0 * 0	0	* 0 0 0 * 0	0	0 * * 0 * 0	0	0 * 0 0 * 0	0	0 0 * 0 * 0	0	0 0 0 0 * 0	0
1 1 0	* * * 0 0 *	0	* * 0 0 0 *	0	* 0 * 0 0 *	0	* 0 0 0 0 *	0	0 * * 0 0 *	0	0 * 0 0 0 *	0	0 0 * 0 0 *	0	0 0 0 0 0 *	0
1 1 1	* * * 0 0 0	0	* * 0 0 0 0	0	* 0 * 0 0 0	0	* 0 0 0 0 0	0	0 * * 0 0 0	0	0 * 0 0 0 0	0	0 0 * 0 0 0	0	0 0 0 0 0 0	0

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 201–208
DOI: 10.15439/2014F111
ACSIS, Vol. 2

Bronchopulmonary Dysplasia Prediction Using Support Vector Machine and LIBSVM

Marcin Ochab
AGH University of Science and Technology,
30 Mickiewicza 30-059 Kraków, Poland
Email: marcin.ochab@labor.it.pl

Wiesław Wajs
AGH University of Science and Technology,
30 Mickiewicza 30-059 Kraków, Poland
Email: wwa@agh.edu.pl

***Abstract*—The paper presents BPD (Bronchopulmonary Dysplasia) prediction for extremely premature infants after their first week of life. SVM (Support Vector Machine) algorithm implemented in LIBSVM[1] was used as classifier. Results are compared to others gathered in previous work [2] where LR (Logit Regression) and Matlab environment SVM implementation were used. Fourteen different risk factor parameters were considered and due to the high computational complexity only 3375 random combinations were analysed. Classifier based on eight feature model provides the highest accuracy which was 82.60%. The most promising 5-feature model which gathered 82.23% was reasonably immune to random data changes and consistent with LR results. The main conclusion is that unlike Matlab SVM[2] implementation, LIBSVM can be successfully used in considered problem, but it is less stable than LR. In addition, the article discusses influence of the model parameters selection on prediction quality.**

I. Introduction

BRONCHOPULMONARY dysplasia (BPD) is a chronic pulmonary morbidity which affects premature infants [3], [4]. It is most common among children who received prolonged mechanical ventilation to treat respiratory distress syndrome [5], [6] and those with low birth weight. Almost a third of infants with birth weight lower than 1000g [7] are affected. Due to the fact that the disease is poorly understood, many projects are focused on identifying its factors of risk. Since it can not be diagnosed until a 28th day of life [8], it is very important to predict such a result after the end of the first week, which would enable an early prevention of the disease[9]. Therefore, an intensive work has been done to define a classifier, based on static parameters (gathered after birth) and dynamic ones (collected during the first week of life), which would be able to predict the diagnosis. Although several prediction models of BPD [10], [11], [12], [13], [14], [15], [16], [17], [18], [19] used in research have been reported, none of them could be used in common clinical practice due to the variety of reasons and none use SVM.

II. Related works

As mentioned before, there are numerous works related to BPD, its risk factors and prediction [20], [21], [22], [23]. The most popular one is the analysis of static data whose main features are gestational age and birth weight. The other factors considered are admission of surfactant, presence of patent ductus arteriosis (PDA), or respiratory support. In addition, dynamical data (which is much harder to obtain) is analysed in more sophisticated models. Most of such parameters are: arterial blood gas variables like fraction of inspired oxygen (FiO_2) or alveolar-arterial ratio (AA) [24] (which is respiratory distress degree measure); blood gas levels like oxygen saturation of arterial hemoglobin (SpO_2) and its standard deviation, mean value etc. [25] or even time series analysis [26]; heart beat and its derivatives.

$$AA = \frac{pO_2}{p_{ATM} \cdot FiO_2 - pCO_2}, \tag{1}$$

where pO_2— oxygen partial pressure, p_{ATM}—atmospheric pressure, pCO_2—carbon dioxide partial pressure, FiO_2—fraction of inspired oxygen.

Some of the papers introduced race and ethnicity or sex as factors which seem to be promising but require a very big set of data. It should be indicated that the vast majority of studies uses logit regression (LR) in prediction. Best LR models gain about 73% to 82% of accuracy. Many of authors mention use of support vector machine(SVM)[27] in future works, however it is difficult to find them. In our previous work[2] we compared SVM with LR classifiers. Unfortunately, due to internal Matlab SVM library usage results were highly unsatisfactory. The highest accuracy gained was only 79.39%. Moreover, the bigger features set was used the worse results we got. In general, only three and four feature models were able to gain accuracy higher than 70%. That is why we decided to use LIBSVM implementation instead, which gave us a very wide scope of parameters tuning. Although additional parameters highly increased computational complexity of optimal model search, even limited random parameters space exploration gave us quite good results, comparable to LR.

III. General ideas of used methods

A. Logistic regression

Probability of the dependent variable equalling a BPD positive diagnosis ($y_k = 1$), on condition that explanatory variables (features of specific case k) equals $X_k = (x_{1,k}, x_{2,k}, .., x_{n,k})$, we define as:

$$p_k = P(y_k = 1|X_k) = \frac{e^{a_0 + \sum_{i=1}^{n} a_i x_i}}{1 + e^{a_0 + \sum_{i=1}^{n} a_i x_i}}, \tag{2}$$

978-83-60810-58-3/$25.00

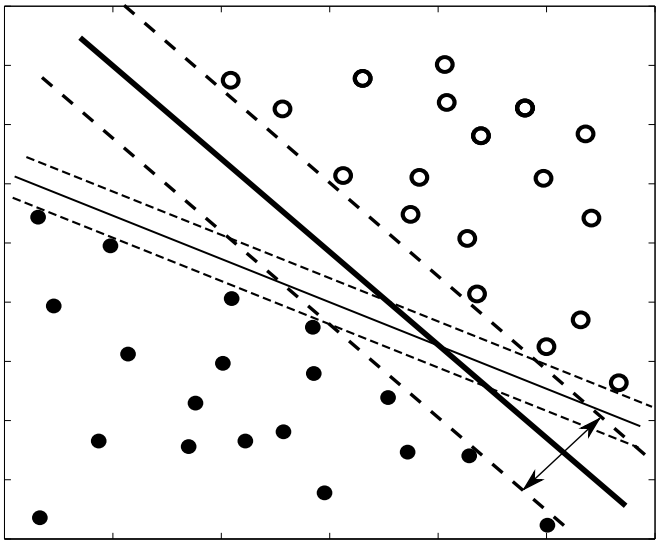


Fig. 1. Maximizing margin in SVM method.

where $x_{i,k}$—explanatory variables (feature values for case k), a_i—regression coefficients, n—number of features.

In contrast to linear regression, where we assume normal distribution of the independent variables and because explanatory variables variance are not equal, we can not use the method of least squares to obtain regression coefficients. Thus they are usually calculated using maximum likelihood estimation, maximizing likelihood function (L) or minimizing its negative logarithm using learning data:

$$L = \prod_{y_k=1} p_k \prod_{y_k=0} (1 - p_k), \tag{3}$$

$$ln(L) = \sum_{k=1}^{m} [y_k \cdot ln(p_k) + (1 - y_k) \cdot ln(1 - p_k)], \tag{4}$$

where k—observation number (learning case), y_k —diagnosis for case k, m—number of observations.

Having a_i regression coefficients, we can easily predict BPD positive diagnosis probability of case X using Eq. 2.

B. Support vector machine

We define the learning data D divided in two classes y as:

$$D = \{(X_k, y_k) | X_k \in R^n, y_k \in \{1, -1\}\}_{k=1}^{m} \tag{5}$$

We are looking for hyperplane

$$W \bullet X + b = 0 \tag{6}$$

which separates classes and provides maximum margin as on Fig. 1, which is the same as the problem of minimizing L :

$$L(W) = \frac{||W||^2}{2} + c \cdot \sum_{k=1}^{m} \varepsilon_k, \tag{7}$$

with conditions:

$$y_k(W \cdot \phi(X_k) + b) \geq 1 - \varepsilon_k \tag{8}$$

where $\varepsilon \geq 0$—slack variable, $c > 0$—penalty parameter for each point wrongly classified, ϕ—kernel function.

Thanks to the kernel functions for non linear separable problems, we can transform original data from n dimensional space to p dimensional ($p > n$, as on Fig. 2), in which there is much higher likelihood that they will be linear separable.

IV. Data and Methods

Data was collected thanks to the Neonatal Intensive Care Unit of The Department of Pediatrics at Jagiellonian University Medical College using our own software. It includes 109 patients born prematurely with birth weight less than or equal to 1500g admitted, no later than on the second day of life. For 46 of them BPD have been diagnosed after fourth week of life.

To build a suitable model 14 different features mentioned in literature were considered:

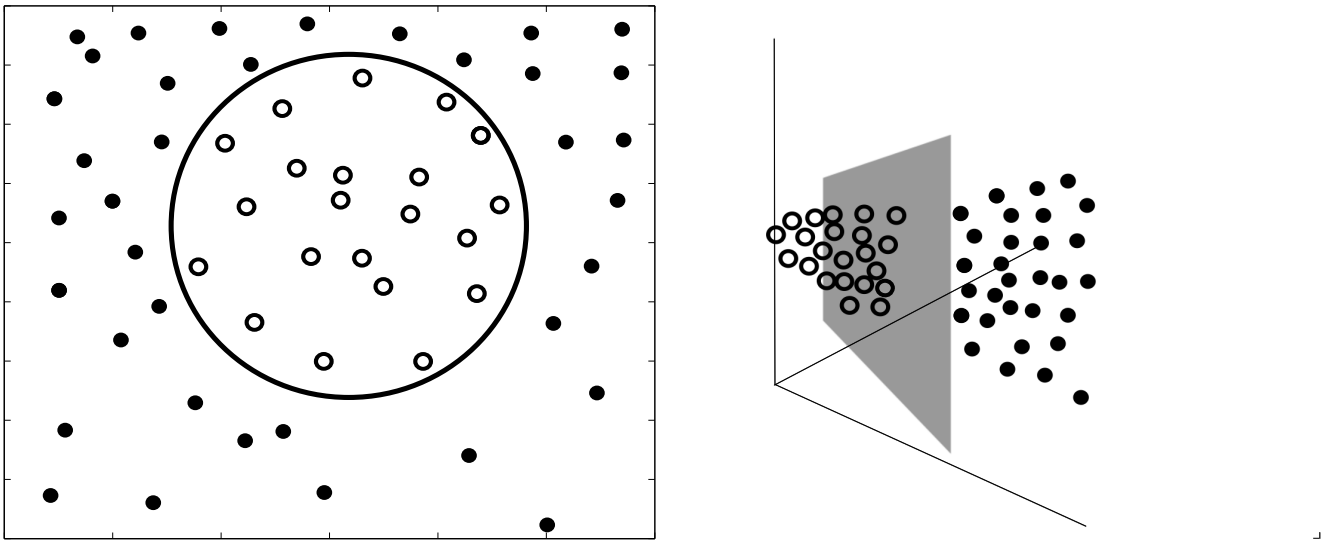
- Binary such as:
  - presence of patent ductus arteriosus (PDA) [28] ,
  - use of a respirator ($RESPIMV$) during the first week of life,
  - administration of surfactant ($SURFACT$) [29] in the same period.
- Real-Valued (values range in parentheses) such as:
  - birth weight ($BWEIGHT$) (550-1500g),
  - gestational age ($GAGE$) (22-34 weeks),
  - alveolar-arterial ratio (AA) (0.05-1) measured during patient admission,
  - a percentage of the time during first week for which the oxygen saturation of hemoglobin was less than 85% ($LOW85$) (0.03%-12.45%) or higher than 94% ($HIGH94$) (14.56%-99.02%),
  - average number of heartbeats per minute ($BPMMEAN$)[5](124.69-161.42 bpm),
  - mean and standard deviation of oxygen saturation ($SPO2MEAN$, $SPO2DEV$) (accordingly 89.89%-98.99% and 1.19-7.98) and their trends (first day to first week ratio: $BPMMEAN\_TR$, $SPO2MEAN\_TR$, $SPO2DEV\_TR$) (accordingly 0.8-1.18 , 0.96-1.07 and 0.51-2.36).

Accuracy (ACC) defined as below was considered as preliminary result measure. The sensitivity(TPR) and specificity(SPC) were also obtained:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \tag{9}$$

$$TPR = \frac{TP}{TP + FN}, \tag{10}$$

$$SPC = \frac{TN}{TN + FP}, \tag{11}$$



(a) data in R^2 space

(b) data transformed to R^3 space

Fig. 2. Making data linear separable with dimensional transformation.

where TP—True Positives, FP—False Positives, FN—False Negatives, TN—True Negatives.

As mentioned before in SVM computation version 3.17 of LIBSVM library was used. Based on a few arbitrary chosen and tested models it has been found that for specified problem C-SVC method is more effective than nu-SVC (range of parameter c (Eq. 7) is from zero to infinity, rather than as in nu always between [0,1]). It has also been discovered that sigmoid kernel function gives better results and is much faster in finding the separating hyperplane than radial basis (RBF) one.

$$K(X_i, X_j) \equiv \phi(X_i)^T \phi(X_j), \tag{12}$$

$$RBF : K(X_i, X_j) = e^{-\gamma ||X_i - X_j||^2}, \gamma > 0, \tag{13}$$

$$Sigmoid : K(X_i, X_j) = tanh(\gamma X_i^T X_j + r), \tag{14}$$

where γ, r—kernel parameters.

As suggested in library documentation, data have been normalized to [-1,1] range and optimization of parameters γ ($\gamma = 2^{-15}, 2^{-14}, \ldots, 2^3$) and c ($c = 2^{-5}, 2^{-4}, \ldots, 2^{15}$) was performed for each model (Fig. 3). Unlike suggested, in presented results we did not use *grid.py* script which provides cross-validation and parameters optimization. This method gives very promising results for considered problem, achieving easily up to 83-86% of ACC, but models found with this method turned out to be very unstable. Any random change of data (removing or adding samples) significantly decreased its accuracy.

It is very important to find a model the most possibly independent on specific learning data. Therefore, we decided to use a method similar to Jacknife [30]: for each pair c and γ parameters calculations were repeated 30 times, each time randomly excluding 30 samples of data and using cross-validation procedure (each patient was treated as a test sample while all other data was learning set) on the rest of it. This way deviation and mean value of accuracy, sensitivity and specificity were obtained, which gives an estimate on the model 'sensitivity' to data structure (it might be important when calculating on such a little data set as 109 patients). It should be mentioned that each time we refer in this paper to ACC, TPR or SPC values we mean average computed as above. The test was repeated once again for the best results, excluding the data of only 10 random patients - it shows wether extension of learning data increases or decreases accuracy and overfitting occurs. Due to the high computational complexity of proposed optimisation procedure only 3375 random models containing 2 to 14 parameters were analyzed - it took more than a week for modern 8-core Intel Core i7 based PC. Nevertheless, such little number of experiments gave satisfactory results.

To compare with LR algorithm we used data from previous article[2], where we reviewed all of the 2^{14} possible combinations of models with exactly same Jacknife and cross-validation procedure.

V. Results

The most essential results are presented in Table II . To compare, in each presented model mean value of ACC, TPR and SPC were obtained with different Jacknife parameters, using both methods: LR and sigmoid SVM with LIBSVM. Where applicable we added RBF Matlab SVM implementation results (as M. SVM) from [2].

According to the assumptions in previous section the highest mean value of accuracy among SVM results gained eight-parameter model with 82.60%. Unfortunately standard deviation of ACC was 5.15, which in contrast to others was

TABLE I
SVM MODELS WITH BOTH TPR AND SPC HIGHER THAN 80%

Params count	BWEIGHT	GAGE	RESPIMV	AA	PDA	SURFACT	SPO2MEAN	SPO2DEV	LOW85	HIGH94	BPMMEAN	BPMMEAN_TR	SPO2DEV_TR	SPO2MEAN_TR	Items excluded with Jacknife	ACC	TPR	SPC	Method	ACC dev
																mean value				
8		•	•	•	•	•		•	•	•					10	82.60%	80.10%	84.37%	LIBSVM	5.15
															30	79.57%	75.84%	82.17%	LIBSVM	4.95
															10	77.49%	78.62%	76.43%	LR	1.51
															30	76.42%	76.72%	76.09%	LR	2.75
5		•	•			•					•	•			10	82.18%	83.68%	83.68%	LIBSVM	2.95
															10	81.62%	87.21%	77.47%	LR	1.51
															30	80.73%	85.99%	76.56%	LR	2.27
															30	79.79%	78.02%	81.03%	LIBSVM	3.42
11	•	•		•	•		•	•	•		•	•	•	•	10	81.41%	80.65%	82.01%	LIBSVM	3.33
															30	80.81%	75.82%	84.28%	LIBSVM	3.78
															10	79.83%	83.83%	76.90%	LR	2.22
															30	78.29%	80.05%	76.87%	LR	3.02
11		•	•	•	•	•	•	•	•	•		•	•		10	80.50%	80.17%	80.68%	LIBSVM	2.73
															30	77.26%	77.54%	76.91%	LIBSVM	3.07
															10	76.99%	80.30%	74.53%	LR	1.21
															30	76.00%	79.87%	72.71%	LR	3.00
9		•	•	•	•	•			•			•	•	•	10	80.38%	80.46%	80.31%	LIBSVM	2.65
															10	78.12%	82.49%	75.02%	LR	1.26
															30	77.58%	77.15%	77.67%	LIBSVM	4.29
															30	76.83%	83.04%	71.87%	LR	2.93

very high (STD values are presented in Table III). The TPR value was 80.10% and SPC 84.37%, which is a very good result in comparison with the fact that only five of examined SVM models achieved both parameters higher than 80% (see Table I).

Next we present generally the best ACC model which was six-parameter LR which gained 82.79% with TPR=84.20%, SPC=81.73% and standard deviation of ACC as low as 1.11. Both above models give quite good results, but only using specific algorithm (LR or SVM).

In contrast to the above fact, we present next two best five- and six-parameter models which give accuracy higher than 81% for both methods. Particularly noteworthy model is the five-parameter one - it gives respectively 82.23% for SVM and 82.59% for LR of ACC which was the best of examined five-parameter models for both methods. Although the TPR result for SVM was only 77.13%, this model seems to be a very reasonsonable choice with SPC=85.87% for SVM and TPR=84.65%, SPC=81.07% for LR. Even in that case, where standard deviation of ACC for SVM was 3.04, it was about twice higher than 1.49 for LR. An interesting similar 5-parameter model with little lower accuracy is presented in Table I.

Afterwards we made a review of the best accuracy results for each method and each model size:

- 4-parameter models - for LR we succeeded to obtain 82.01% of ACC and only 80.85% for SVM. However, this second model (bolded in Table III) draws attention due to its simplicity, low ACC standard deviation (about 1.4) and ACC higher than 80% for both methods.
- 3-parameter models - 81.29% for LR and 80.57% for SVM are good results just as standard deviation of ACC lower than 1.5, though none of models exceeds the psychological barrier of 80% of ACC for both methods at a time.
- 2-parameter models - SVM results are unsatisfactory due to 76.01% of highest ACC and its standard deviation as high as 5.81. However, LR was able to gain 80.30% of ACC, which is quite interesting.

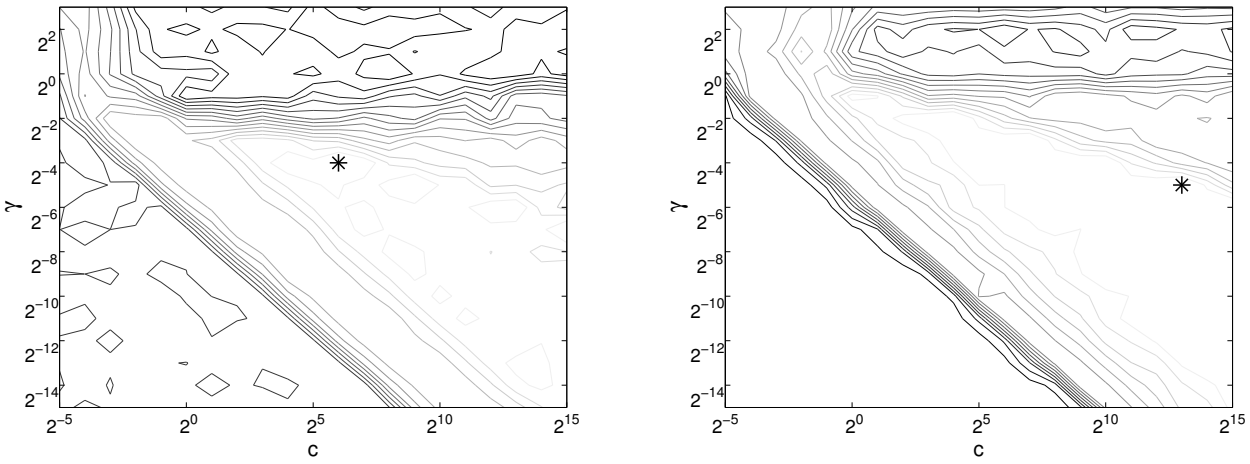
Lastly, it must be noted that full 14-parameter SVM model gained 76.86% and LR 77.55% of ACC. As a graphic example two-parameter model result was presented on Fig. 4.

VI. DISCUSSION

The first conclusion is that SVM classification algorithm can be almost as accurate as LR and if its parameters are properly chosen it gives rewarding results, even for a complicated multi-parameter model of BPD. The best choice for such a prediction is to use the LIBSVM instead of Matlab's implementation, which gives less control on computation process. Most likely that was the reason why the bigger parameter set we used the worse results we got using Matlab[2]. Although we did not test all possible 2^{14} combinations of parameters (only 3375 random models), nonetheless looking on Table III it can be concluded that standard deviation of accuracy for SVM is much higher than for logit regression. Using bigger learning set (only 10 samples excluded) for best results it reaches even

TABLE II
MODELS AND METHODS ACCURACY COMPARISON

Params count	BWEIGHT	GAGE	RESPIMV	AA	PDA	SURFACT	SPO2MEAN	SPO2DEV	LOW85	HIGH94	BPMMEAN	BPMMEAN_TR	SPO2DEV_TR	SPO2MEAN_TR	Items excluded with Jacknife	ACC	TPR	SPC	Method	Model comment
																mean value				
8		•	•	•	•	•		•	•	•					10	**82.60%**	**80.10%**	**84.37%**	**LIBSVM**	best LIBSVM
															30	79.57%	75.84%	82.17%	LIBSVM	
															10	77.49%	78.62%	76.43%	LR	
															30	76.42%	76.72%	76.09%	LR	
6		•		•	•			•			•	•			10	**82.79%**	**84.20%**	**81.73%**	**LR**	best LR
															30	80.85%	82.72%	79.36%	LR	
															30	78.49%	72.35%	82.78%	LIBSVM	
															10	78.22%	73.38%	81.77%	LIBSVM	
															10	70.01%	64.56%	74.01%	M.SVM	
															30	69,00%	61.91%	74.00%	M.SVM	
6	•	•			•	•			•		•				10	82.67%	87.50%	79.09%	LR	6-feature both methods ACC > 81%
															30	81.12%	85.78%	77.62%	LR	
															10	81.50%	77.98%	84.04%	LIBSVM	
															30	78.22%	72.38%	82.30%	LIBSVM	
															10	70.48%	65.77%	73.83%	M.SVM	
															30	69.90%	64.58%	73.69%	M.SVM	
5		•			•				•		•	•			10	**82.59%**	**84.65%**	**81.07%**	**LR**	5-feature best LR &SVM
															10	**82.23%**	**77.13%**	**85.87%**	**LIBSVM**	
															30	81.46%	81.53%	81.30%	LR	
															30	80.46%	74.31%	84.82%	LIBSVM	
															10	72.10%	64.12%	77,88%	M.SVM	
															30	71.40%	63.94%	76.69%	M.SVM	
4		•			•				•		•				10	82.01%	84.38%	80.34%	LR	4-feature best LR
															30	80.64%	80.08%	81.02%	LR	
															10	78.78%	72.96%	83.07%	LIBSVM	
															30	77.89%	72.25%	81.97%	LIBSVM	
															10	75.38%	71.32%	78.31%	M.SVM	
															30	74.68%	69.58%	78.16%	M.SVM	
4		•							•		•	•			10	80.85%	71.95%	87.32%	LIBSVM	4-feature best LIBSVM
															30	80.35%	71.11%	86.96%	LIBSVM	
															10	80.19%	73.27%	85.07%	LR	
															30	79.90%	74.31%	83.94%	LR	
3		•				•			•						10	81.29%	88.70%	75.90%	LR	3-feature best LR, best M.SVM
															30	81.05%	88.52%	75.34%	LR	
															10	80.15%	78.73%	81.17%	M.SVM	
															10	79.83%	78.51%	80.77%	LIBSVM	
															30	79.19%	79.84%	78.71%	M.SVM	
															30	79.02%	77.73%	79.68%	LIBSVM	
3		•				•				•					10	80.57%	78.83%	81.82%	LIBSVM	3-feature best LIBSVM
															10	79.66%	78.94%	80.19%	M.SVM	
															30	79.47%	79.70%	79.25%	LIBSVM	
															30	79.43%	79.08%	79.54%	M.SVM	
															10	79.40%	87.20%	73.57%	LR	
															30	79.32%	86.36%	73.98%	LR	
2		•				•									30	81.06%	84.39%	78.62%	LR	2-feature best LR
															10	80.30%	83.48%	78.01%	LR	
															10	73.57%	57.54%	85.22%	LIBSVM	
															30	73.22%	55.87%	85.28%	LIBSVM	
															30	71.38%	52.67%	84.98%	M.SVM	
															10	71.02%	51.68%	84.90%	M.SVM	
2		•						•							30	77.35%	72.50%	80.41%	LR	2-feature best LIBSVM
															10	76.85%	69.73%	81.95%	LR	
															10	76.01%	56.68%	90.09%	LIBSVM	
															30	73.04%	58.38%	83.77%	LIBSVM	



(a) eight-parameter model (b) five-parameter model

Fig. 3. Parameter γ and c optimization for eight-parameter and five-parameter models bolded in Table II.

TABLE III
STANDARD DEVIATION OF MODELS ACCURACY

Parameters count	Method	Jacknife excluded count			
		10		30	
		ACC	ACC STD	ACC	ACC STD
8	LIBSVM	82.60%	5.15	79.57%	4.95
	LR	77.49%	1.51	76.42%	2.75
6	LR	82.79%	1.11	80.85%	2.52
	LIBSVM	78.22%	1.96	78.49%	2.74
6	LR	82.67%	1.19	81.12%	2.56
	LIBSVM	81.50%	4.89	78.22%	5.06
5	LR	82.59%	1.49	81.46%	2.51
	LIBSVM	82.23%	3.04	80.46%	2.44
4	LR	82.01%	1.73	80.64%	2.60
	LIBSVM	78.78%	2.68	77.89%	4.36
4	LIBSVM	**80.85%**	**1.39**	80.35%	2.28
	LR	**80.19%**	**1.43**	79.90%	2.89
3	LR	81.29%	1.74	81.05%	2.36
	LIBSVM	79.83%	1.73	79.02%	3.34
3	LIBSVM	80.57%	1.48	79.47%	3.06
	LR	79.40%	1.28	79.32%	2.30
2	LR	80.30%	1.62	81.06%	2.38
	LIBSVM	73.57%	2.40	73.22%	3.01
2	LR	76.85%	1.14	77.35%	2.41
	LIBSVM	76.01%	5.81	73.04%	2.70

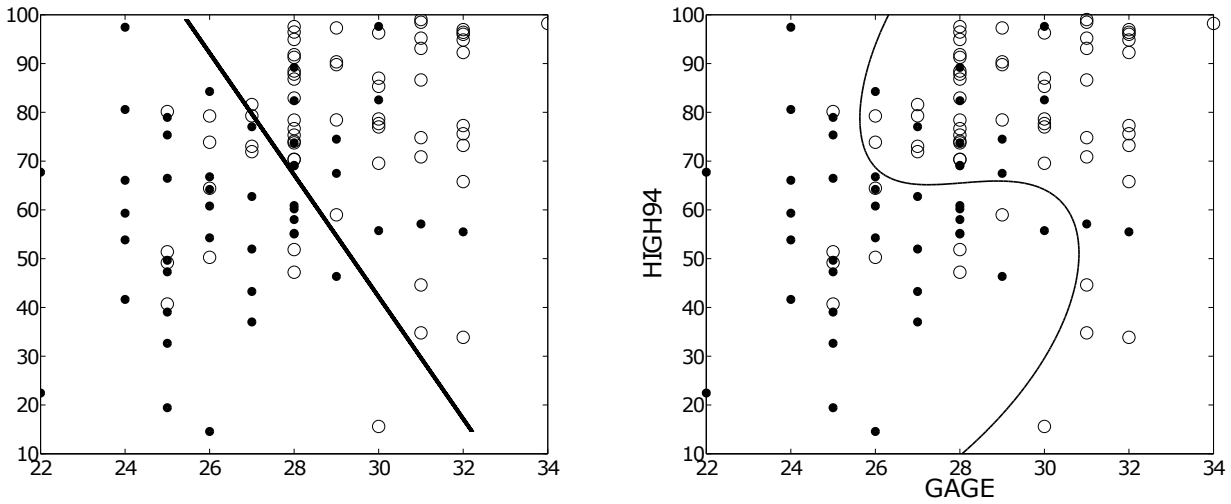
5.81 while for LR it is rarely as high as 1.74. In other words, SVM fits to data very well and even minor random changes of data causes instability of results accuracy. For this reason, when we execute our test procedure for certain model few times, we get results which differ even 2%. According to our observations such an effect occurs mostly for a very high or very low parameter models. For logit regression effect has not been noticed, which may be encouraging to select that algorithm.

We also observed that for almost all cases the more data we exclude from the test results the worse accuracy and deviation we achieve (results with Jacknife exclusion of 10 samples are generally better), which is promising - it shows that overfitting does not occur and classifier is well generalizing to cases not known during learning. The exception from that rule are very high-deviation and some simple two-parameter models mentioned before. In such cases, there is a concern that because of its oversimplification or excessive complication hard learning takes place. For this reason, it seems more secure to use logit regression or four- to six-parameter model.

Analyzing sensitivity and specificity of the best results from Table II we have the following observations:

- for SVM differences between TPR or SPC and ACC are from 1.32% to 6.47%, while for LR 1.06% to 7.8% which is quite similar,
- the exceptions are 2-parameter SVM models, for which this differences were up to 19.33%,
- as mentioned before only five of all the examined SVM models achieved both parameters higher than 80%, which is not a problem for many LR models - all five are presented in Table I.

We confirmed that one of the most important risk factors mentioned in literature [5], [6], [8], [9], [28], [29], [31] is the $GAGE$ which exists in almost all (all presented in Table I and II) of the models with acceptable accuracy. Most classifiers presented in the literature consist $BWEIGHT$ and $RESPIMV$ parameters. Unexpectedly, $RESPIMV$ parameter is indeed present in SVM models in Table I with very good (over 82% of ACC) results, but it is not in any LR model worth to present. However, $BWEIGHT$-containing one was indeed on the third place with 82.67% ACC, but it was the only one in the first twenty, which is a group with ACC higher than 82%. Among the most frequently mentioned parameters there is FiO_2 which depends on the AA (Eq. 1) feature used in our work that is indeed present in the best model, but also only in five others of the best twenty. On the other hand,



(a) Logistic regression

(b) Support Vector Machine

Fig. 4. Prediction result sample for two-parameter ($GAGE$, $HIGH94$).

PDA feature appeared in 15 out of the best 20 models which is consistent with published results [8], [9], [25], [28]. The same situation occurs with $BPMMEAN$ parameter, which is the second most important factor after GAGE especially in LR models - it is in all of the first 20 models. The only one parameter related to SpO_2 which is in more than half of the best 20 classifiers is LOW85. The average importance features are $SURFACT$ and $HIGH94$ while $SPO2DEV$, $SPO2DEV\_TR$, $SPO2MEAN\_TR$ parameters seem to have even less effect on the occurrence of diseases.

As a final conclusion we confirmed [32] that prediction of BPD after 7th day of life is possible with the accuracy higher than 82%, not only with LR but also using Support Vector Machine algorithm. Results are slightly worse than in the Logit Regression method and more attention should be paid on model selection because many of them are sensitive to even small data changes. However, it can be very useful when we have limited set of parameters (above RESPIMV example), which are not so important in LR models as in SVM. Having wide scope of algorithms we can choose the one which is more suitable for our parameter set and thereby obtain better classification results. With that knowledge it is a good idea to construct expert system that would advise which algorithm and model to use having certain parameters measured.

REFERENCES

[1] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011, 2:27:1–27:27. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm doi: 10.1145/1961189.1961199

[2] Ochab Marcin and Wiesław Wajs. Bronchopulmonary Dysplasia Prediction Using Support Vector Machine and Logit Regression. Information Technologies in Biomedicine, Volume 4. Springer International Publishing, 2014. 365-374. doi: 10.1007/978-3-319-06596-0_34

[3] Horbar JD, Badger GJ, Carpenter JH, Fanaroff AA, Kilpatrick S, LaCorte M, Phibbs R, Soll RF; Members of the Vermont Oxford Network. Trends in mortality and morbidity for very low birth weight infants, 1991-1999. Pediatrics 2002;110:143—151. doi: 10.1542/peds.110.1.143

[4] Stoll BJ, Hansen NI, Bell EF, Shankaran S, Laptook AR, Walsh MC, Hale EC, Newman NS, Schibler K, Carlo WA, et al.; Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network. Neonatal outcomes of extremely preterm infants from the NICHD Neonatal Research Network. Pediatrics 2010;126:443—456. doi: 10.1542/peds.2009-2959

[5] Jobe, A. H.: The new bronchopulmonary dysplasia. Current opinion in pediatrics, 2011, 23(2), 167 doi: 10.1097/MOP.0b013e3283423e6b

[6] Groothuis, J. R., Makari, D.: Definition and outpatient management of the very low-birth-weight infant with bronchopulmonary dysplasia. Advances in therapy, 2012, 29(4), 297—311 doi: 10.1007/s12325-012-0015-y

[7] Walsh, M. et al. Summary proceedings from the bronchopulmonary dysplasia group. Pediatrics 117(3), S52—56 (2006) doi: 10.1542/peds.2005-0620I

[8] Tapia, J.L., Agost, D., Alegria, A., Standen, J., Escobar, M., Grandi, C., et al.: Bronchopulmonary dysplasia: incidence, risk factors and resource utilization in a population of South American very low birth weight infants. Journal de Pediatria (Rio J). 2006, 82(1), 15—20 doi: 10.1590/S0021-75572006000100005

[9] Farstad, T., Bratlid, D., Medbø, S., Markestad, T.: Bronchopulmonary dysplasia–prevalence, severity and predictive factors in a national cohort of extremely premature infants. Acta Paediatrica, 2011, 100(1), pp. 53—58 doi: 10.1111/j.1651-2227.2010.01959.x

[10] Ryan SW, Nycyk J, Shaw BN. Prediction of chronic neonatal lung disease on day 4 of life. Eur J Pediatr 1996;155:668—671. doi: 10.1007/BF01957150

[11] Subhedar NV, Hamdan AH, Ryan SW, Shaw NJ. Pulmonary artery pressure: early predictor of chronic lung disease in preterm infants. Arch Dis Child Fetal Neonatal Ed 1998;78:F20—F24. doi:10.1136/fn.78.1.F20

[12] Romagnoli C, Zecca E, Tortorolo L, Vento G, Tortorolo G. A scoring system to predict the evolution of respiratory distress syndrome into chronic lung disease in preterm infants. Intensive Care Med 1998;24: 476—480. doi: 10.1007/s001340050599

[13] Toce SS, Farrell PM, Leavitt LA, Samuels DP, Edwards DK. Clinical and roentgenographic scoring systems for assessing bronchopulmo- nary dysplasia. Am J Dis Child 1984;138:581—585. doi:10.1001/archpedi.1984.02140440065017

[14] Corcoran JD, Patterson CC, Thomas PS, Halliday HL. Reduction in the risk of bronchopulmonary dysplasia from 1980—1990: results of a mul-

tivariate logistic regression analysis. Eur J Pediatr 1993;152:677—681. doi: 10.1007/BF01955247

[15] Noack G, Mortensson W, Robertson B, Nilsson R. Correlations between radiological and cytological findings in early development of bron- chopulmonary dysplasia. Eur J Pediatr 1993;152:1024—1029. doi: 10.1007/BF01957230

[16] Yuksel B, Greenough A, Karani J. Prediction of chronic lung disease from the chest radiograph appearance at seven days of age. Acta Paediatr 1993;82:944—947. doi: 10.1111/j.1651-2227.1993.tb12605.x

[17] Bhutani VK, Abbasi S. Relative likelihood of bronchopulmonary dysplasia based on pulmonary mechanics measured in preterm neonates during the first week of life. J Pediatr 1992;120:605—613. doi: 10.1016/S0022-3476(05)82491-6

[18] Kim YD, Kim EA, Kim KS, Pi SY, Kang W. Scoring method for early prediction of neonatal chronic lung disease using modified respiratory parameters. J Korean Med Sci 2005;20:397—401. doi: 10.3346/jkms.2005.20.3.397

[19] Bhering CA, Mochdece CC, Moreira ME, Rocco JR, Sant'Anna GM. Bronchopulmonary dysplasia prediction model for 7-day-old infants. J Pediatr (Rio J) 2007;83:163—170. doi: 10.1590/S0021-75572007000200011

[20] Rojas MA, Gonzalez A, Bancalari E, Claure N, Poole C, Silva-Neto G. Changing trends in the epidemiology and pathogenesis of neonatal chronic lung disease. J Pediatr 1995;126:605—610. doi: 10.1016/S0022-3476(95)70362-4

[21] Marshall DD, Kotelchuck M, Young TE, Bose CL, Kruyer L, O'Shea TM. Risk factors for chronic lung disease in the surfactant era: a North Carolina population-based study of very low birth weight infants. North Carolina Neonatologists Association. Pediatrics 1999; 104:1345—1350. doi: 10.1542/peds.104.6.1345

[22] Oh W, Poindexter BB, Perritt R, Lemons JA, Bauer CR, Ehrenkranz RA, Stoll BJ, Poole K, Wright LL; Neonatal Research Network. Association between fluid intake and weight loss during the first ten days of life and risk of bronchopulmonary dysplasia in extremely low birth weight infants. J Pediatr 2005;147:786—790. doi: 10.1016/j.jpeds.2005.06.039

[23] Ambalavanan N, Van Meurs KP, Perritt R, Carlo WA, Ehrenkranz RA, Stevenson DK, Lemons JA, Poole WK, Higgins RD. NICHD Neo- natal Research Network, Bethesda, MD. Predictors of death or bronchopulmonary dysplasia in preterm infants with respiratory failure. J Perinatol 2008;28:420—426 doi: 10.1038/jp.2008.18

[24] Gilbert R., Keighley J., The arterial/alveolar oxygen tension ratio. An index of gas exchange applicable to varying inspired oxygen concentrations., Am Rev Respir Dis., 1974, 109, 142-145.

[25] Stoch, P.: Prediction of BronchoPulmonary Dysplasia in preterm neonates using statistical and artificial neural network tools. (Thesis or Dissertation style) Ph.D. dissertation, AGH University of Science and Technology, Kraków, 2007 (in Polish), pp. 60—72

[26] Kuenzel L. Predicting and undestanding bronchopulmonary dysplasia in permature infants. Stanford Undergraduate Research Journal

[27] Burges, C. J.C.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2, Kluwer Academic Publishers, Boston 1998, pp. 121—167 doi: 10.1023/A:1009715923555

[28] Sosenko, I. R., Bancalari, E.: New Developments in the Pathogenesis and Prevention of Bronchopulmonary Dysplasia. The Newborn Lung: Neonatology Questions and Controversies: Expert Consult-Online and Print, 2012, 217

[29] Cunha, G.S., Mezzacappa-Filho, F., Ribeiro, J. D.: Risk Factors for Bronchopulmonary Dysplasia in very Low Birth Weight Newborns Treated with Mechanical Ventilation in the First Week of Life. Journal of Tropical Pediatrics, 2005, 51(6), 334—340 doi: 10.1093/tropej/fmi051

[30] Jones, H. L.: Jacknife estimation of functions of stratum means. Biometrika, 1974, 61(2), 343—348 doi: 10.1093/biomet/61.2.343

[31] Ali, Z., Schmidt, P., Dodd, J., Jeppesen, D. L.: Bronchopulmonary dysplasia: a review. Archives of gynecology and obstetrics, 2013, 1—9 doi: 10.1007/s00404-013-2753-8

[32] Laughon, Matthew M., et al. "Prediction of bronchopulmonary dysplasia by postnatal age in extremely premature infants." American journal of respiratory and critical care medicine 183.12 (2011): 1715. doi: 10.1164/rccm.201101-0055OC

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 209–217
DOI: 10.15439/2014F249
ACSIS, Vol. 2

Improving the performance of machine learning classifiers for Breast Cancer diagnosis based on feature selection

Noel Pérez*, Miguel A. Guevara†, Augusto Silva†, Isabel Ramos‡ and Joana Loureiro‡
*Institute of Mechanical Engineering and Industrial Management (INEGI)
Campus da FEUP, 4200-465 Porto, Portugal
Email: nperez@inegi.up.pt
†Institute of Electronics and Telematics Engineering of Aveiro (IEETA)
Campus Universitário de Santiago, 3810-193 Aveiro, Portugal
Email: {mguevaral, augusto.silva}@ua.pt
‡Faculty of Medicine - Centro Hospitalar São João (FMUP-HSJ)
Al. Prof. Hernâni Monteiro, 4200-319 Porto, Portugal
Email: radiologia.hs@mail.telepac.pt; joanaploureiro@gmail.com

***Abstract*—This paper proposed a comprehensive algorithm for building machine learning classifiers for Breast Cancer diagnosis based on the suitable combination of feature selection methods that provide high performance over the Area Under receiver operating characteristic Curve (AUC). The new developed method allows both for exploring and ranking search spaces of image-based features, and selecting subsets of optimal features for feeding Machine Learning Classifiers (MLCs). The method was evaluated using six mammography-based datasets (containing calcifications and masses lesions) with different configurations extracted from two public Breast Cancer databases. According to the Wilcoxon Statistical Test, the proposed method demonstrated to provide competitive Breast Cancer classification schemes reducing the number of employed features for each experimental dataset.**

I. Introduction

BREAST CANCER is a major concern and the second-most common and leading cause of cancer deaths among women [1]. According to published statistics, Breast Cancer has become a major health problem in both developed and developing countries over the past 50 years. Its incidence has increased recently with an estimated of 1,152,161 new cases in which 411,093 women die each year [2]. At present, there are no effective ways to prevent it, because its cause remains unknown. However, an efficient diagnosis in its early stages can give a woman a better chance of full recovery [2]. Therefore, its early detection can play an important role in reducing the associated morbidity and mortality rates.

Breast Cancer screening has proved to be the best way to detect cancer early. A useful and suggested approach is the double reading of mammograms (two radiologists read the same mammograms) [3], which has been advocated to reduce the proportion of missed cancers, but the workload and cost associated are high. With the support of Breast Cancer Computer-Aided Diagnosis (CADx) systems only one radiologist is needed to read each mammogram rather than two.

There is good evidence in the literature that Breast Cancer CADx systems can improve the AUC performance of radiologists [4] [5] [6] [7] [8], e.g. in [9], it was presented an evaluation of the variation of performance in terms of sensitivity and specificity of two radiologists with different experience in mammography (6 and 2 years respectively), with and without the assistance of two different CADx systems (SecondLook and CALMA). The evaluation was made on a dataset composed by 70 images of patients with cancer (biopsy proven) and 120 images of healthy breasts (with a three years follow up). The results showed that the use of a CADx allows for a substantial increment in sensitivity (up to 15.6%) and a less pronounced decrement in specificity, which was more significant for the least experienced of the radiologists.

However, the performance of current and future commercial CADx systems still needs to be improved so that they can meet the requirements of clinics and screening centers [10] [11].

In this work, we proposed a new method supported on the combination of five Feature Selection Methods (FSMs) and MLCs respectively, for building Breast Cancer classification schemes (i.e. calcifications and mases) that provide the high performance over the AUC curve. The selected FSMs are filter-based methods, which use heuristics (statistics) based on general characteristics of the data rather than a MLC (as wrapper or embedded paradigm) to evaluate the merit of features [12] [13] [14] [15]. As an optimal subset of features is always relative to a certain evaluation function [16], it was used different FSMs with different evaluation function: the traditional CHI-Square Discretization (CHI2) [17] based on the chi-square statistic function, Information Gain (IG) [18] based on the information measure, One-Rule (1Rule) [19] based on rules as a principal evaluation function and Re-

This work was supported by the Breast Cancer Digital Repository Consortium (BCDR - http://bcdr.inegi.up.pt))

lief [20] based on the distance measure. Also, it was used an algorithm developed in previous work [21] named RMean based on a voting function for indexing relevant features (see Algorithm 2, which is revisited here). The proposed method dynamically form subsets of features extracted from resultant rankings (one by each FSM applied) for feeding five machine learning models: Feed Forward Back-Propagation (FFBP) neural network [22], Support Vectors Machine (SVM) [23], Naive Bayes (NB) [24], Linear Discriminant Analysis (LDA) [25] and k-Nearest Neighbors (kNN) [26] respectively. Finally, the selection of the best classification scheme is based on the Wilcoxon Statistical Test [27] [28] following two criteria: (1) the higher obtained AUC value and (2) if there is classification performances tied, the one using less number of employed features is preferred. The method was evaluated on six datasets containing calcifications and masses lesions (with differents configurations) extracted from two public Breast Cancer databases and and it is included a statistical comparison of achieved results.

The remainder of the work is ordered as follows: the Materials and Methods section, overviews the employed databases, FSMs and MLCs. Also, describes in detail the proposed method and the experimental setup design for its evaluation. The Results and Discussion section presents an exploratory comparison based on the obtained AUC scores using the Wilcoxon statistical test [27] [28] to assess the meaningfulness of differences between classification schemes. Finally, Conclusions and Future work are drawn in the last section.

II. Materials and Methods

A. Databases

This work is supported on two public databases: the Breast Cancer Digital Repository (BCDR), which is the first wide-ranging annotated Portuguese Breast Cancer database, with anonymous cases from medical historical archives supplied by Faculty of Medicine - Centro Hospitalar de São João at University of Porto, Portugal [29] and the Digital Database for Screening Mammography (DDSM). For convenience, the DDSM images used in this study were obtained from the Image Retrieval in Medical Applications (IRMA) project (courtesy of TM Deserno, Dept. of Medical Informatics, RWTH Aachen, Germany) where the original LJPEG images of DDSM were converted to 16 bits Portable Network Graphics (PNG) format [30] [31].

The BCDR is composed of 1734 patient cases with mammography and ultrasound images, clinical history, lesion segmentation and selected pre-computed image-based descriptors; each case may have one or more Region of Interest (ROI) with associated Pathological Lesion (PL) segmentations (for different PLs), typically in Mediolateral Oblique (MLO) and Craniocaudal (CC) images of the same breast.

On the other hand, the DDSM database is composed by 2620 patient cases divided into three categories: normal cases (12 volumes), cancer cases (15 volumes) and benign cases (14 volumes); like in the BCDR, each case may have one or more associated PL segmentations, usually in MLO and CC image views of the same breast.

B. Feature Selection Methods

Several types of extracted features (e.g. intensity statistics, shape and texture) from mammograms have been combined to form subsets of features, which extensively provided significant information for lesions classification [32] [33] [34] [35]. However, selecting the most appropriate subset of features is still a very difficult task; usually a satisfactory instead of the optimal feature subset is searched.

The selected methods were all derived from the filter paradigm, because its execution is a one step process without any data exploration (search) involved and are also independent of classifiers [36].

1) CHI2 Discretization: This method consists on a justified heuristic for supervised discretization [17]. Numerical features are initially sorted by placing each observed value into its own interval. Then the chi-square statistic (x^2) is used to determine whether the relative frequencies of the classes in adjacent intervals are similar enough to justify merging. The extent of the merging process is controlled by an automatically set x^2 threshold. The threshold is determined through attempting to maintain the fidelity of the original data.

2) IG method: The IG measurement normalized with the symmetrical uncertainty coefficient [18] is a symmetrical measure in which the amount of information gained about Y after observing X is equal to the amount of information gained about X after observing Y (a measure of feature-feature intercorrelation). This model is used to estimate the value of an attribute Y for a novel sample (drawn from the same distribution as the training data) and compensates for information gain bias toward attributes with more values.

3) 1Rule: This method estimates the predictive accuracy of individual features building rules based on a single feature (can be thought of as single level decision trees) [19]. As it is used training and test datasets, it is possible to calculate a classification accuracy for each rule and hence each feature. Then, from classification scores, a ranked list of features is obtained. Experiments with choosing a selected number of the highest ranked features and using them with common machine learning algorithms showed that, on average, the top three or more features are as accurate as using the original set. This approach is unusual due to the fact that no search is conducted.

4) Relief: This method uses instance based learning to assign a relevance weight to each feature [20]. Each feature weight reflects its ability to distinguish among the class values. The feature weight is updated according to how well its values distinguish the sampled instance from its nearest hit (instance of the same class) and nearest miss (instance of opposite class). The feature will receive a high weight if it differentiates between instances from different classes and has the same value for instances of the same class. For nominal features it is defined as either 1 (the values are different) or 0 (the values are the same), while for numeric features the difference is the actual difference normalized to the interval $[0..1]$.

C. Machine Learning Models

The discrimination between samples of two classes may be formulated as a supervised learning problem, which is defined as the prediction of the value of a function for any valid input after training a learner using examples of input and target output pairs [37]. For the problem at hand, the function has only two discrete values: benign or malignant. Hence the problem can be modeled as a two-class classification problem. A variety of MLCs have been applied in CADx approaches for Breast Cancer detection/classification. The Artificial Neural Networks (ANN) [14] [23] [38] [39], SVM [14] [22] [23] [37] [39] [40] and LDA [41] [42] seem to be the most commonly used type of classifiers. Other less popular, but perform very well are NB [43] [44] [45] and kNN [25] [43] classifiers respectively.

A brief description of these MLCs is given here:

1) FFBP Neural Network Classifier: The FFBP neural network is a particular model of ANN, which provides a nonlinear mapping between its input and output according to the back-propagation error learning algorithm. This model has demonstrated to be capable of approximating an arbitrarily complex mapping within a finite support using only a sufficient number of neurons in few hidden layers (all layers using a sigmoid function as kernel type) [22].

2) SVM Classifier: SVMs are based on the definition of an optimal hyperplane, which linearly separates the training data. In comparison with other classification methods, a SVM aims to minimize the empirical risk and maximize the distances (geometric margin) of the data points from the corresponding linear decision boundary [23].

3) NB Classifier: The NB classifier is based on probabilistic models with strong (Naive) independence assumptions [24]. It assumes that c is a class variable depending on n input features: $x_1, x_2, \cdots, x_n$. The prediction of c can be described by the following conditional model: $p(c|x_1, x_2, \cdots, x_n)$ and according to the Bayes' theorem:

$$p(c|x_1, x_2, \cdots, x_n) = \frac{p(c)p(x_1, x_2, \cdots, x_n|c)}{p(x_1, x_2, \cdots, x_n)}$$

where $p(c)$ is the prior probability of c, $p(x_1, x_2, \cdots, x_n|c)$ is the conditional probability depending on c, and $p(x_1, x_2, \cdots, x_n)$ is the probability of input features. If each feature x_i is conditionally dependent, as the denominator $p(x_1, x_2, \cdots, x_n)$ does not depend on c, which is actually a constant when features are given; the conditional probability over the class variable c can be expressed as:

$$p(c|x_1, x_2, \cdots, x_n) = \frac{1}{z}p(c)\prod_{i=1}^{n} p(x_i|c)$$

where z is a normalization constant. The above NB classifier can be trained based on the relative frequencies shown in the training set to get an estimation of the class priors and feature probability distributions. For a test sample, the decision rule will be picking the most probable hypothesis (value of c) which is known as the maximum a posteriori decision rule using the above model.

4) LDA Classifier: LDA is a traditional method for classification [25]. The basic idea is to try to find an optimal projection (decision boundaries optimized by the error criterion), which can maximize the distances between samples from different classes and minimize the distances between samples from the same class. For the binary classification, observations are classified by the following linear function:

$$g_i(x) = W_i^T x - c_i \quad 1 \leq i \leq 2$$

where W_i^T is the transpose of a coefficient vector, x is a feature vector and c_i is a constant as the threshold. The values of W_i^T and c_i are determined through the analysis of a training set. Once these values are determined, they can be used to classify the new observations (smallest $g_i(x)$ is preferred).

5) kNN Classifier: The kNN classifier is a nonparametric technique called a "lazy learning" because little effort goes into building the classifier and most of the work is performed at the time of classification. The kNN assigns a test sample to the class of the majority of its k-neighbors; that is, assuming that the number of voting neighbors is $k = k_1 + k_2 + k_3$ (where k_i is the number of samples from class i in the k-sample neighborhood of the test sample, usually computed using the Euclidean distance), the test sample is assigned to class m if $k_m = max(k_i), i = 1, 2, 3$ [26].

D. Proposed Method

The proposed method is supported on the combination of five FSMs and MLCs respectively, for building Breast Cancer classification schemes that provide the high performance over the AUC curve.

The employed FSMs are filter methods, which use heuristics (statistics) based on general characteristics of the data rather than a MLC (as wrapper or embedded paradigm) to evaluate the merit of features [12] [13]. As an optimal subset of features is always relative to a certain evaluation function [16], the selected FSMs were: CHI2 discretization [17] based on the chi-square statistic function, IG [18] based on the information measure, 1Rule [19] based on rules as a principal evaluation function, Relief [20] based on the distance measure and the recently developed RMean method [21] based on a voting function (averaging each feature position) for indexing relevant features (see Algorithm 2, which revisited here).

As it is shown in the Algorithm 1, the dataset D and the total of features in the initial subset nS constituted the starting point of the proposed method. Once, this method is a multistep modelling procedure, the application of the k-fold Cross Validation (CV) method [46] to the entire sequence of modelling steps guarantee reliable results [47]. Thus, it was applied 10 times 10-CV before features ranking to avoid giving an unfair advantage to predictors, and before classification step to prevent overfitting of classifiers to the training set [46] (see Algorithm 1 step 3 and 13). The application of FSMs on the processed dataset S_{cv} produced five different ranking of features (see Algorithm 1 step 4 to 8). Then, from each ranking of features, it were dynamically built ranked subset of features with different size S_{ini}.

Algorithm 1 Proposed method

Require: $D[f_1, f_2, f_3, ..., f_n] : n \geq 2$;
$nS \leftarrow$ maximum number of features in the initial subset;
Ensure: $C_{(best)}$; Best classification scheme
1: $C_{(best)} = []; C_{(aux)} = []; S_{ini} = []; S_{cv} = []; D_{cv} = []; R_{CHI2} = []; R_{IG} = []; R_{1R} = []; R_{Rel} = []; L = []$;
2: $nF \leftarrow nS$; Initializing nF
3: $D_{cv} \leftarrow$10-$CV(D)$; Applying 10 times 10-CV
4: $R_{CHI2} = eval(CHI2, D_{cv})$; Ranking by CHI2
5: $R_{IG} = eval(IG, D_{cv})$; Ranking by IG
6: $R_{1R} = eval(1R, D_{cv})$; Ranking by 1R
7: $R_{Rel} = eval(Relief, D_{cv})$; Ranking by Relief
8: $R_{RMean} = eval(RMean, D_{cv})$; Ranking by RMean
9: $L = [R_{CHI2}, R_{IG}, R_{1R}, R_{Rel}, R_{RMean}]$; List of ranking
10: **for** $i = 1 : length(L)$ **do**
11: **for** $j = 1 : trunc(L_i/nS)$ **do**
12: $S_{ini} \leftarrow extract(nF, L_i)$; Extract the first nF features from L_i
13: $S_{cv} \leftarrow$10-$CV(S_{ini})$; Applying 10 times 10-CV
14: $C_{(i,j,FFBP)} \leftarrow eval(FFBP, S_{cv})$ Applying the FFBP
15: $C_{(i,j,SVM)} \leftarrow eval(SVM, S_{cv})$ Applying the SVM
16: $C_{(i,j,NB)} \leftarrow eval(NB, S_{cv})$ Applying the NB
17: $C_{(i,j,LDA)} \leftarrow eval(LDA, S_{cv})$ Applying the LDA
18: $C_{(i,j,kNN)} \leftarrow eval(kNN, S_{cv})$ Applying the kNN
19: $nF \leftarrow nF + nS$; Updating the number of features nF
20: **end for**
21: $C_{(aux)} \leftarrow C_{(aux)} + max(C)$; Higher statistically
22: **end for**
23: $C_{(best)} \leftarrow max(C_{(aux)})$; Higher statistically

These ranked subsets of features were processed by the 10-CV method before feeding the FFBP neural network [22], SVM [23], NB [24], LDA [25] and kNN [26] classifiers respectively (see Algorithm 1 step 13 to 18). In the last step, two important criteria are evaluated in order to select the best classification scheme: (1) the higher obtained AUC value and (2) if there is classification performances tied, the one using less number of employed features is preferred. Both criteria were conducted using the Wilcoxon Statistical Test, i.e. a non-parametric alternative test to the paired t-test, which ranks the differences in performances of two classifiers [27] [28]. This test provided a fairly comparison among all obtained AUC performances, and therefore a reasonable selection of $C_{(best)}$.
The implementation of the proposed method was in JAVA language and the source code of all employed FSMs and MLCs are available in the WEKA data mining software version 3.6 [48].

E. Experimental Setup

This section outlines the experimental evaluation design of the proposed method using two public Breast Cancer databases. That involves the datasets creation and machine learning models configurations are important aspects to be described here.

Algorithm 2 RMean

Require: $D[f_1, f_2, f_3, ..., f_n] : n \geq 2$;
Ensure: R_{Mean};
1: $R_{Mean} = []; R_{CHI2} = []; R_{IG} = []; R_{1R} = []; R_{Rel} = []; D_{cv} = []$;
2: $D_{cv} \leftarrow$10-$CV(D)$; Applying 10 times 10-CV
3: $R_{CHI2} \leftarrow eval(CHI2, D_{cv})$; Ranking by CHI2
4: $R_{IG} \leftarrow eval(IG, D_{cv})$; Ranking by IG
5: $R_{1R} \leftarrow eval(1R, D_{cv})$; Ranking by 1R
6: $R_{Rel} \leftarrow eval(Relief, D_{cv})$; Ranking by Relief
7: $R_{Mean} \leftarrow (R_{CHI2} + R_{IG} + R_{1R} + R_{Rel})/4$; Averaging the features position throughout resultant rankings from steps 3,4,5 and 6.
8: $R_{Mean} \leftarrow sort(R_{Mean}, 'ascendant')$; Sorting in ascendant way the resultant ranking from the step 6.

1) Datasets Creation: A set of 23 image-based descriptors (features) were extracted from the BCDR and DDSM databases to be used in this work. Selected descriptors included intensity statistics, shape and texture features, computed from segmented calcifications and masses in both MLO and CC mammography views. The intensity statistics and shape descriptors were selected according to the radiologists experience (similar to the clinician procedure) and the American College of Radiology (BIRADS-Mammography atlas) [49], which described in detail how to detect/classify pathological lesions. Additionally, texture descriptors were the Halarick's descriptors extracted from the grey-level co-occurrence matrices [50]. An overview of the mathematical formulation for computing features is presented below:

- *Skewness*:
$$f_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^3}$$
with x_i being the i^{th}-value and $\bar{x}$ the sample mean.
- *Kurtosis*:
$$f_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^2} - 3$$
with x_i being the i^{th}-value and $\bar{x}$ the sample mean.
- *Area Fraction* (f_3): is the percentage of non-zero pixels in the image or selection.
- *Circularity*:
$$f_4 = 4\pi\frac{area}{perimeter^2}$$
- *Perimeter*: $f_5 = length(E)$ with $E \subset O$ being the edge pixels.
- *Elongation*: $f_6 = \left(\frac{m}{M}\right)$ with m being the minor axis and M the major axis of the ellipse that has the same normalized second central moments as the region surrounded by the contour.

- *Standard Deviation*:

$$f_7 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

with x_i being the grey level intensity of the i^{th}-pixel and $\bar{x}$ the mean of intensity.

- *Roughness*:

$$f_8 = \frac{perimeter^2}{4\pi * area}$$

- *Minimum* (f_9) and *Maximum* (f_{10}): the minimum and maximum intensity value in the region surrounded by the contour.
- *Shape*:

$$f_{11} = \frac{perimeter * elongation}{8 * area}$$

- *X Centroid*:

$$f_{12} = \frac{min(x) + max(x)}{2}$$

with x being the set of X coordinates of the object's contour.

- *Entropy*:

$$f_{13} = \sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j)log(p(i,j))$$

with *L* being the number of grey-levels, and $p(i,j)$ being the probability of pixels with grey-level i occur together to pixels with grey-level j.

- *X Center Mass* (f_{14}): normalized X coordinates of the center of mass of O.
- *Angular Second Moment*:

$$f_{15} = \sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j)^2$$

with L being the number of grey-levels, and $p(i,j)$ being the probability of pixels with grey-level i occur together to pixels with grey-level j.

- *Median*:

$$f_{16} = \begin{cases} \frac{n+1}{2} & \text{if } length(X) \text{ is odd} \\ \frac{X(\frac{n}{2}) + X(\frac{n}{2}+1)}{2} & \text{if } length(X) \text{ is even} \end{cases}$$

with X being the set of intensities.

- *Contrast*:

$$f_{17} = \sum_{i}\sum_{j} p(i,j)(i-j)^2$$

with $p(i,j)$ being the probability of pixels with grey-level i occur together to pixels with grey-level j.

- *Correlation*:

$$f_{18} = \frac{\sum_i \sum_j [ijp(i,j)] - \mu_x\mu_y}{\sigma_x\sigma_y}$$

with μ_x, μ_y, σ_x and σ_y being the means and standard deviations of the marginal distribution associated with $p(i,j)$.

- *Mean*:

$$f_{19} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

with n being the number of pixels inside the region delimited by the contour and x_i being the grey level intensity of the i^{th} pixel inside the contour.

- *Inverse Difference Moment*:

$$f_{20} = \sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j)$$

with $p(i,j)$ being the probability of pixels with grey-level i occur together to pixels with grey-level j.

- *Y Center Mass* (f_{21}): normalized Y coordinates of the center of mass of O.
- *Area*: $f_{22} = |O|$ with O being the set of pixels that belong to the segmented lesion.
- *Y Centroid*:

$$f_{23} = \frac{min(y) + max(y)}{2}$$

with y being the set of Y coordinates of the object's contour.

Conformable to the number of patient cases of used databases, it were created six datasets containing calcifications and masses lesions with different configurations: (1) two balanced datasets (same quantity of benign and malignant instances), (2) two unbalanced datasets containing more benign than malignant instances and (3) two unbalanced datasets holding more malignant than benign instances, representatives of BCDR and DDSM respectively. The BCDR supplies several datasets for scientific purposes (Availaible on http://bcdr.inegi.up.pt), we used the BCDR-F01 distribution to form the BCDR1 dataset holding 374 features vectors; BCDR2 and BCDR3 datasets with a total of 287 features vectors respectively.

Due to the wide range of information in the DDSM database, it were considered only two volumes of cancer and benign cases (random selection) to form the DDSM1 dataset holding 582 features vectors; DDSM2 and DDSM3 datasets with a total of 491 features vectors respectively. Figure 1 shows a detailed description of the datasets creation workflow.

2) MLCs Configuration: For all MLCs with the exception of the NB (which is parameterless), 10-CV method [46] was performed on the training set for optimizing the classifiers parameters.

The FFBP neural network was used with a total of hidden layers determined according to the equation (*attributes* + *number of classes*)/2; one output layer associated with the binary classification (benign or malignant); transfer function for all layers based on the sigmoid function and the number of iterations (epochs) were optimized in the range of 100 to 1000 epochs (with an interval increment of 100 units).

The SVM classifier was used with the regularization parameter *C* (cost) optimized in the range of 10^{-3} to 10^3 and the kernel

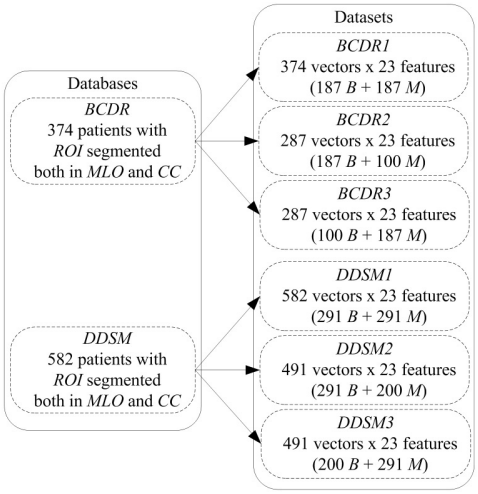


Fig. 1. Datasets creation flowchart; *B* and *M* represent Benign and Malignant class instances.

type based on a linear function, which provided better results respect to others kernel such as: radial basis, polynomial and sigmoid function (from our experimental experience).
The kNN classifier included the estimation of an optimal value k for the size of the neighborhood varying from 1 to 20, and the contribution of each neighbor was always weighted by the distance to the instance being classified.

III. Results and Discussion

According to the experimental setup section, a total of 750 ranked subsets of features containing image-based features extracted from segmented calcifications and masses lesions were analyzed using the proposed method and the straightforward statistical comparison based on the mean of AUC performances over 100 runs highlighted interesting results for balanced and unbalanced datasets respectively.

3) Performance on Balanced Datasets: The higher AUC value obtained in the BCDR1 dataset was formed by the SVM classifier and the RMean method using a total of 15 features (AUC value of 0.8365). This result was statistically superior to the majority of the remaining classification schemes. However, there were others schemes with similar performances statistically at p=0.05 (see Table I). From these results, it is possible to select the FFBP neural network in combination with the CHI2 discretization and RMean method with 5 features as the most appropriate classification schemes in this dataset. They reached an AUC value of 0.8264 and 0.8272 using the minimum number of features respectively.
For DDSM1 dataset, the higher AUC value was obtained by the combination of the LDA classifier and the RMean method using 20 features (AUC value of 0.807). However, this result did not provide statistical evidence to be better than others combinations (see Table I). Similar to the DDSM1 dataset, the combinations formed by the FFBP neural network in conjunction with the Relief and RMean methods using 10 features provided similar performances

TABLE I
Classification schemes with nonsignificant difference in AUC performances for BCDR1 and DDSM1 balanced datasets

Dataset	Best Scheme	Other Scheme	AUC	p=0.05
BCDR1	SVM+RMean+15F (0.8365)	FFBP+CHI2+5F	0.8264	p=0.574
		FFBP+RMean+5F	0.8272	p=0.494
		FFBP+CHI2+10F	0.8219	p=0.359
		SVM+Relief+15F	0.8831	p=0.603
		LDA+RMean+15F	0.8284	p=0.295
		LDA+RMean+20F	0.8279	p=0.286
DDSM1	LDA+RMean+20F (0.807)	FFBP+Relief+10F	0.8061	p=0.592
		FFBP+RMean+10F	0.8056	p=0.475
		SVM+RMean+20F	0.7939	p=0.139

TABLE II
Classification schemes with nonsignificant difference in AUC performances for BCDR2 and DDSM2 unbalanced datasets

Dataset	Best Scheme	Other Scheme	AUC	p=0.05
BCDR2	SVM+RMean+10F (0.8389)	FFBP+RMean+10F	0.8352	p=0.573
		LDA+CHI2+10F	0.8278	p=0.365
		LDA+RMean+10F	0.8284	p=0.403
DDSM2	FFBP+RMean+5F (0.8406)	FFBP+IG+10F	0.8405	p=0.682

statistically (AUC value of 0.8061 and 0.8056 respectively). These results were obtained using a less number of employed features. Thus, both classification schemes were selected as the most appropriated classification schemes in this dataset.

4) Performance on Unbalanced Datasets: The higher AUC performance for BCDR2 dataset was formed by the SVM classifier and the RMean method using 10 features, reaching an AUC value of 0.8389. This result was not statistically superior to obtained results by others classification schemes

TABLE III
Classification schemes with nonsignificant difference in AUC performances for BCDR3 and DDSM3 unbalanced datasets

Dataset	Best Scheme	Other Scheme	AUC	p=0.05
BCDR3	LDA+RMean+5F (0.8611)	FFBP+RMean+5F	0.8562	p=0.592
DDSM3	LDA+RMean+20F (0.807)	FFBP+Relief+10F	0.8061	p=0.592
		FFBP+RMean+5F	0.78	p=0.094
		SVM+Relief+10F	0.7879	p=0.139
		SVM+Relief+15F	0.7853	p=0.126
		SVM+RMean+10F	0.786	p=0.129
		SVM+RMean+15F	0.783	p=0.128
		NB+Relief+10F	0.785	p=0.125
		NB+RMean+10F	0.783	p=0.118
		LDA+Relief+10F	0.7883	p=0.145
		LDA+Relief+15F	0.7877	p=0.139
		LDA+1R+20F	0.7845	p=0.12
		LDA+RMean+10F	0.789	p=0.153
		LDA+RMean+15F	0.7861	p=0.135

using the same number of employed features(see Table II). Therefore, the four combinations presented in the Table II could be considered as the most appropriated schemes for lesions classification in the BCDR2 dataset.

Besides, for DDSM2 dataset the best classification performance was obtained by the combination of the FFBP neural network classifier and the RMean method using 5 features; reaching AUC value of 0.8406. However, this result did not statistically outperform the obtained result by the combination of the FFBP neural network classifier and the IG method with 10 features, attainment an AUC value of 0.8405 (see Table II). Despite the small and insignificant difference in term of AUC performances, the first combination was selected as the most appropriated classification scheme because it reached this results using a less number of features.

The best classification performance for BCDR3 dataset was provided by the combination of the LDA classifier and the RMean method with 5 features, accomplishment an AUC score of 0.8611 (see Table III). This result was not statistically superior to the obtained result by the combination of the FFBP neural network classifier and the RMean method with 5 features, which achieved an AUC score of 0.8562. As both classification schemes reached these results using the minimum number of employed features could be considered as the most appropriated classification schemes for BCDR3 dataset.

In the DDSM3 dataset, the higher AUC value was obtained by the combination of LDA classifier and the RMean method with a total of 10 features (AUC value of 0.7889). This classification result showed nonsignificant difference respect to others combinations, which reached similar AUC performances (see Table III). Despite the several combinations which can be used as a good classification scheme for this dataset. Only the scheme formed by the FFBP neural network and the RMean method stretched the result using the minimum number of features (5). Thus, it was considered as the most appropriated classification scheme in the DDSM3 dataset (see Table III).

Regarding of the classifiers performance, results show that the selection of the most appropriated classifier is dependent on the dataset and the FSM. From Table I, II and III, it possible to read that the best MLC was the FFBP neural network classifier, appearing consistently on every appropriated classification scheme for all datasets. These results were expected since this classifier has demonstrated to be capable of generalizing decision boundary in a more complex features space [25]. Meanwhile the best FSM was the RMean method (see Algorithm 2), which appeared consistently on every successful classification scheme, providing in most cases the minimal subset of features.

IV. Conclusions and Future Work

In this work, it is made a statistical exploration of different classification schemes within the context of Breast Cancer classification. The main contribution it was developed a new and robust method for building machine learning classifiers that combines suitably several feature selection methods with different evaluation function. This method was effective in providing competitive classification schemes for balanced and unbalanced datasets: the FFBP neural network and the RMean method using 5 features was the best scheme for BCDR1, BCDR3, DDSM2 and DDSM3 datasets, attainment an AUC value of 0.8264, 0.8562, 0.8406 and 0.78 respectively. Also, the FFBP neural network and the RMean method with 10 features in the DDSM1 dataset, reaching an AUC value of 0.8056, and the SVM classifier with the RMean method using 10 features for the BCDR2, stretching an AUC value of 0.8399. Regarding to MLCs and FSMs, the FFBP neural network classifier and the RMean method were the best, appearing consistently in the majority of successful schemes. In future work, we plan to assess the performance using others benchmarking datasets with different experimental setup: including clinical and more image-based features to evaluate the sensibility and generalization of the proposed method. Also, it's further integration in a real Breast Cancer CADx system.

Acknowledgment

MSc. Pérez acknowledges "Fundação para a Ciência e a Tecnologia (grant SFRH/BD/48896/2008)" for financial support. Prof. Guevara acknowledges the Cloud Thinking project (CENTRO-07-ST24-FEDER-002031), co-funded by QREN, "Mais Centro" program. The institutions participating in the Breast Cancer Digital Repository Consortium express their gratitude for the support of the European Regional Development Fund.

References

[1] M. D. Althuis, J. M. Dozier, W. F. Anderson, S. S. Devesa, and L. A. Brinton, "Global trends in breast cancer incidence and mortality 1973-1997", Int. J. Epidemiol., vol. 34, pp. 405-412, April 1, 2005, http://dx.doi.org/10.1093/ije/dyh414.

[2] F. Kamangar, G. M. Dores, and W. F. Anderson, "Patterns of cancer incidence, mortality, and prevalence across five continents: defining priorities to reduce cancer disparities in different geographic regions of the world", Journal of clinical oncology, vol. 24, pp. 2137-2150, 2006, http://dx.doi.org/10.1200/JCO.2005.05.2308.

[3] J. Brown, S. Bryan, and R. Warren, "Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms", BMJ (Clinical research ed.), vol. 312, pp. 809-812, 1996, http://dx.doi.org/10.1136/bmj.312.7034.809.

[4] Z. Huo, M. L. Giger, C. J. Vyborny, and C. E. Metz, "Breast cancer: effectiveness of computer-aided diagnosis observer study with independent database of mammograms", Radiology, vol. 224, pp. 560-8, Aug 2002, http://dx.doi.org/10.1148/radiol.2242010703.

[5] L. Hadjiiski, H. P. Chan, B. Sahiner, M. A. Helvie, M. A. Roubidoux, C. Blane, et al., "Improvement in radiologists' characterization of malignant and benign breast masses on serial mammograms with computer-aided diagnosis: an ROC study", Radiology, vol. 233, pp. 255-65, Oct 2004, http://dx.doi.org/10.1148/radiol.2331030432.

[6] L. Hadjiiski, B. Sahiner, M. A. Helvie, H. P. Chan, M. A. Roubidoux, C. Paramagul, et al., "Breast masses: computer-aided diagnosis with serial mammograms", Radiology, vol. 240, pp. 343-56, Aug 2006, http://dx.doi.org/10.1148/radiol.2401042099.

[7] K. Horsch, M. L. Giger, C. J. Vyborny, L. Lan, E. B. Mendelson, and R. E. Hendrick, "Classification of breast lesions with multimodality computer-aided diagnosis: observer study results on an independent clinical data set", Radiology, vol. 240, pp. 357-68, Aug 2006, http://dx.doi.org/10.1148/radiol.2401050208.

[8] L. A. Meinel, A. H. Stolpen, K. S. Berbaum, L. L. Fajardo, and J. M. Reinhardt, "Breast MRI lesion classification: Improved performance of human readers with a backpropagation neural network computer-aided diagnosis (CAD) system", Journal of Magnetic Resonance Imaging, vol. 25, pp. 89-95, 2007, http://dx.doi.org/10.1002/jmri.20794.

[9] A. Lauria, M. E. Fantacci, U. Bottigli, P. Delogu, F. Fauci, B. Golosio, et al., "Diagnostic performance of radiologists with and without different CAD systems for mammography", in Medical Imaging 2003, 2003, pp. 51-56, http://dx.doi.org/10.1117/12.480079.

[10] E. D. Pisano, C. Gatsonis, E. Hendrick, M. Yaffe, J. K. Baum, S. Acharyya, et al., "Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening", N Engl J Med, vol. 353, pp. 1773-1783, October 27, 2005, http://dx.doi.org/10.1056/NEJMoa052911.

[11] S. Ciatto, N. Houssami, D. Gur, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, et al., "Computer-Aided Screening Mammography", N Engl J Med, vol. 357, pp. 83-85, July 5, 2007, http://dx.doi.org/10.1056/NEJMc071248.

[12] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection", J. Mach. Learn. Res., vol. 3, pp. 1157-1182, 2003.

[13] I. Guyon and A. Elisseeff, "An Introduction to Feature Extraction", in Feature Extraction. vol. 207, I. Guyon, M. Nikravesh, S. Gunn, and L. Zadeh, Eds., ed: Springer Berlin Heidelberg, 2006, pp. 1-25, http://dx.doi.org/10.1007/978-3-540-35488-8_1.

[14] N. Pérez, M. A. Guevara, and A. Silva, "Improving breast cancer classification with mammography, supported on an appropriate variable selection analysis", in SPIE Medical Imaging, 2013, pp. 867022-867022-14, http://dx.doi.org/10.1117/12.2007912.

[15] N. Pérez, M. A. Guevara, and A. Silva, "EVALUATION OF FEATURES SELECTION METHODS FOR BREAST CANCER CLASSIFICATION", Icem15: 15th International Conference on Experimental Mechanics, p. 10, 2012.

[16] M. Dash and H. Liu, "Feature Selection for Classification", Intelligent Data Analysis, vol. 1, pp. 131-156, Jan 1, 1997.

[17] H. Liu and R. Setiono, "Chi2: Feature Selection and Discretization of Numeric Attributes", 1995, pp. 388-388, http://dx.doi.org/10.1109/TAI.1995.479783.

[18] B. P. Flannery, W. H. Press, S. A. Teukolsky, and W. Vetterling, "Numerical recipes in C", Press Syndicate of the University of Cambridge, New York, 1992.

[19] R. Holte, "Very Simple Classification Rules Perform Well on Most Commonly Used Datasets", Machine Learning, vol. 11, pp. 63-90, 1993/04/01 1993, http://dx.doi.org/10.1023/A:1022631118932.

[20] K. Kira and L. A. Rendell, "A practical approach to feature selection", presented at the Proceedings of the ninth international workshop on Machine learning, Aberdeen, Scotland, United Kingdom, 1992.

[21] N. Pérez, M. A. Guevara, A. Silva, and I. Ramos, "Ensemble features selection method as tool for Breast Cancer classification", Computing and Informatics, 2013, unpublished (under review).

[22] Y. H. Hu and J.-N. Hwang, "Introduction to Neural Networks for Signal Processing", in Handbook of neural network signal processing, ed: CRC press, 2001.

[23] A. Papadopoulos, D. I. Fotiadis, and A. Likas, "Characterization of clustered microcalcifications in digitized mammograms using neural networks and support vector machines", Artificial Intelligence in Medicine, vol. 34, pp. 141-150, Jun 2005, http://dx.doi.org/10.1016/j.artmed.2004.10.001.

[24] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification (2nd Edition): Wiley-Interscience, 2000.

[25] M. A. Alolfe, A. M. Youssef, Y. M. Kadah, and A. S. Mohamed, "Computer-Aided Diagnostic System based on Wavelet Analysis for Microcalcification Detection in Digital Mammograms", in Biomedical Engineering Conference, 2008. CIBEC 2008. Cairo International, 2008, pp. 1-5, http://dx.doi.org/10.1109/CIBEC.2008.4786080.

[26] S. Wang and R. M. Summers, "Machine learning and radiology", Medical Image Analysis, vol. 16, pp. 933-951, 2012, http://dx.doi.org/10.1016/j.media.2012.02.005.

[27] J. Demšar, "Statistical comparisons of classifiers over multiple data sets", The Journal of Machine Learning Research, vol. 7, pp. 1-30, 2006.

[28] M. Hollander and D. A. Wolfe, Nonparametric statistical methods, 2nd Edition ed.: Wiley-Interscience, 1999.

[29] R. Ramos-Pollan, M. A. Guevara-Lopez, C. Suarez-Ortega, G. Diaz-Herrero, J. M. Franco-Valiente, M. Rubio-Del-Solar, et al., "Discovering mammography-based machine learning classifiers for breast cancer diagnosis", J Med Syst, vol. 36, pp. 2259-69, Aug 2012, http://dx.doi.org/10.1007/s10916-011-9693-2.

[30] J. E. de Oliveira, A. M. Machado, G. C. Chavez, A. P. Lopes, T. M. Deserno, and A. Araujo Ade, "MammoSys: A content-based image retrieval system using breast density patterns", Comput Methods Programs Biomed, vol. 99, pp. 289-97, Sep 2010, http://dx.doi.org/10.1016/j.cmpb.2010.01.005.

[31] Júlia E. E. Oliveira, Mark O. Gueld, Arnaldo de A. Araújo, Bastian Ott, and T. M. Deserno., "Towards a Standard Reference Database for Computer-aided Mammography", in SPIE - Medical Imaging 2008: Computer-Aided Diagnosis, 69151Y, 2008, http://dx.doi.org/10.1117/12.770325.

[32] H. Soltanian-Zadeh, F. Rafiee-Rad, and S. Pourabdollah-Nejad D, "Comparison of multiwavelet, wavelet, Haralick, and shape features for microcalcification classification in mammograms", Pattern Recognition, vol. 37, pp. 1973-1986, Oct 2004, http://dx.doi.org/10.1016/j.patcog.2003.03.001.

[33] S.-K. Lee, P.-c. Chung, C.-I. Chang, C.-S. Lo, T. Lee, G.-C. Hsu, et al., "Classification of clustered microcalcifications using a Shape Cognitron neural network", Neural Networks, vol. 16, pp. 121-132, Jan 2003, http://dx.doi.org/10.1016/S0893-6080(02)00164-8.

[34] Y. López, Novoa, Andra., Guevara, Miguel., Silva, Augusto, "Breast Cancer Diagnosis Based on a Suitable Combination of Deformable Models and Artificial Neural Networks Techniques", in Progress in Pattern Recognition, Image Analysis and Applications. vol. Volume 4756/2008, ed: Springer Berlin / Heidelberg, 2008, pp. 803-811, http://dx.doi.org/10.1007/978-3-540-76725-1_83.

[35] Y. López, Novoa, Andra., Guevara, Miguel., Quintana, Nicolás., Silva, Augusto, "Computer Aided Diagnosis System to Detect Breast Cancer Pathological Lesions", in Progress in Pattern Recognition, Image Analysis and Applications. vol. Volume 5197/2008, ed: Springer Berlin / Heidelberg, 2008, pp. 453-460, http://dx.doi.org/10.1007/978-3-540-85920-8_56.

[36] L. Talavera, "An Evaluation of Filter and Wrapper Methods for Feature Selection in Categorical Clustering", in Advances in Intelligent Data Analysis VI. vol. 3646, A. F. Famili, J. Kok, J. Peña, A. Siebes, and A. Feelders, Eds., ed: Springer Berlin Heidelberg, 2005, pp. 440-451, http://dx.doi.org/10.1007/11552253_40.

[37] M. Elter and A. Horsch, "CADx of mammographic masses and clustered microcalcifications: a review", Medical physics, vol. 36, pp. 2052-2068, 2009, http://dx.doi.org/10.1118/1.3121511.

[38] Z. Ping, B. Verma, and K. Kuldeep, "A neural-genetic algorithm for feature selection and breast abnormality classification in digital mammography", in Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on, 2004, pp. 2303-2308 vol.3, http://dx.doi.org/10.1109/IJCNN.2004.1380985.

[39] M. E. Mavroforakis, H. V. Georgiou, N. Dimitropoulos, D. Cavouras, and S. Theodoridis, "Mammographic masses characterization based on localized texture and dataset fractal analysis using linear, neural and support vector machine classifiers", Artif Intell Med, vol. 37, pp. 145-62, Jun 2006, http://dx.doi.org/10.1016/j.artmed.2006.03.002.

[40] J. C. Fu, S. K. Lee, S. T. C. Wong, J. Y. Yeh, A. H. Wang, and H. K. Wu, "Image segmentation feature selection and pattern classification for mammographic microcalcifications", Computerized Medical Imaging and Graphics, vol. 29, pp. 419-429, Sep 2005, http://dx.doi.org/10.1016/j.compmedimag.2005.03.002.

[41] J. Shi, B. Sahiner, H. P. Chan, J. Ge, L. Hadjiiski, M. A. Helvie, et al., "Characterization of mammographic masses based on level set segmentation with new image features and patient information", Med Phys, vol. 35, pp. 280-90, Jan 2008.

[42] J. L. Jesneck, J. Y. Lo, and J. A. Baker, "Breast mass lesions: computer-aided diagnosis models with mammographic and sonographic descriptors", Radiology, vol. 244, pp. 390-8, Aug 2007, http://dx.doi.org/10.1148/radiol.2442060712.

[43] D. Moura and M. Guevara López, "An evaluation of image descriptors combined with clinical data for breast cancer diagnosis", International Journal of Computer Assisted Radiology and Surgery, vol. 8, pp. 561-574, Jul 2013, http://dx.doi.org/10.1007/s11548-013-0838-2.

[44] G. I. Salama, M. Abdelhalim, and M. A.-e. Zeid, "Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers", Breast Cancer (WDBC), vol. 32, p. 2, 2012, .

[45] A. Christobel, "An Empirical Comparison of Data mining Classification Methods", International Journal of Computer Information Systems, vol. 3, 2011.

[46] F. García López, M. García Torres, B. Melián Batista, J. A. Moreno Pérez, and J. M. Moreno-Vega, "Solving feature subset selection problem by a parallel scatter search", European Journal of Operational Research, vol. 169, pp. 477-489, 2006, http://dx.doi.org/10.1016/j.ejor.2004.08.010.

[47] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The elements of statistical learning: data mining, inference and prediction", The Mathematical Intelligencer, vol. 27, pp. 83-85, 2005.

[48] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update", ACM SIGKDD explorations newsletter, vol. 11, pp. 10-18, 2009, http://dx.doi.org/10.1145/1656274.1656278.

[49] "American College of Radiology (ACR) ACR BIRADS - Mammography", in ACR Breast Imaging Reporting and Data System, Breast Imaging Atlas, Reston, VA, 2003.

[50] R. M. Haralick, Shanmuga.K, and I. Dinstein, "Textural Features for Image Classification", IEEE Transactions on Systems Man and Cybernetics, vol. Smc3, pp. 610-621, 1973, http://dx.doi.org/10.1109/Tsmc.1973.4309314.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 219–225
DOI: 10.15439/2014F204
ACSIS, Vol. 2

Topological MRI Prostate Segmentation Method

Done Stojanov
Faculty of Computer Science,
University Goce Delcev-Stip
ul. Krste Misirkov 10/A, 2000 Stip, Macedonia
Email: done.stojanov@ugd.edu.mk

Saso Koceski
Faculty of Computer Science,
University Goce Delcev-Stip
ul. Krste Misirkov 10/A, 2000 Stip, Macedonia
Email: saso.koceski@ugd.edu.mk

Abstract—The main aim of this paper is to advance the state of the art in automated prostate segmentation using T2 weighted MR images, by introducing a hybrid topological MRI prostate segmentation method which is based on a set of pre-labeled MR atlas images. The proposed method has been experimentally tested on a set of 30 MRI T2 weighted images. For evaluation the automated segmentations of the proposed scheme have been compared with the manual segmentations, using an average Dice Similarity Coefficient (DSC). Obtained quantitative results have shown a good approximation of the segmented prostate.

Keywords: Prostate segmentation, MRI T2, hybrid topological method

I. Introduction

PROSTATE cancer is one of the major healthcare problems affecting men's population and is the second most common cancer in men worldwide. An estimated 1.1 million men worldwide were diagnosed with prostate cancer in 2012, accounting for 15% of the cancers diagnosed in men. Considering this worrying data, it is predicted that the number of cases will almost double by 2030 [1]. Consequently, there is an increased demand and interest in advancements and enhancements of current methodologies for prostate cancer diagnosis and treatment planning.

Determination of proper information about the prostate location, its volume and shape of prostate gland are basic task and play essential role in numerous clinical applications. This information is crucial for cancer detection, localization and staging, guided biopsy, radiation treatment planning, but also for surgical planning and image-guided robotic-aided laparoscopic prostatectomy (RALP) with augmented reality (AR). In order to provide accurate information various imaging techniques are used in the clinical practice. Nowadays, trans rectal ultrasound (TRUS) is probably the most common and widespread medical imaging technique employed for cancer detection [2], [3], [4] as well as for guided needle biopsy [5]. This is mainly due to its low cost, portability and real-time acquisition. However, this technique has its own drawbacks. Namely, due to the low sensitivity prostate cancer visualization is poor, its false negative rate is high [6] and often resulting in high rates of rebiopsies.

Therefore, the Computer Tomography (CT) has been proposed as an alternative, and it is mainly used in prostate brachytherapy to determine the placement of the radioactive seeds and also to confirm the seed location post-procedure [7]. On the other hand, CT requires ionising radiation and nephrotoxic contrast media and could not provide differentiation between external and internal prostate anatomy because of the poor soft-tissue resolution.

Therefore, in the last decade high-resolution MRI have been promoted as a valuable alternative to before mentioned imaging techniques, which offers physicians better evaluation of the prostate diseases. In the clinical practice nowadays three different modalities of MR images are normally produced: T2-weighted, diffusion-weighted and dynamic contrast enhanced images. Recently, many scientific works have proved that MRI has very high accuracy in the detection of prostate diseases [8], [9] significantly improving the diagnostic rates. It enables easier image segmentation and determination of prostate shape and boundaries which is the basic step in clinical applications.

Usual MRI prostate examination results with a series of multiple images which are presenting plenty of anatomical and functional data regarding the prostate tissues. Analysis and segmentation of these images in major percentage of the cases in the clinical practice, currently is performed by experienced radiologists who based on their knowledge of the anatomy.

However, manual segmentation of prostate boundaries on multiple images in the MRI series could be extremely difficult and time consuming task, especially for series containing large number of images. Manual segmentation is subjective and could be performed differently by different experts and thus could produce different outcomes.

Because of this, currently there is a huge demand for fast and accurate automatic or semi-automatic segmentation methods for clinical applications.

Development of automatic segmentation algorithms and methodologies faces huge challenges, mainly owing to variability of prostate size and shape from patient to patient,

variable intensity ranges inside the prostate region and tissues of surrounding organs, as well as the absence of clear prostate boundaries.

The main aim of this paper is to advance the state of the art in automated prostate segmentation using T2 weighted MR images, by introducing a topological MRI prostate segmentation method using a set of pre-labeled MR atlas images.

The rest of the paper is organized as follows: in part II we present the current state of the art in automatic medical image segmentation methods, in part III we present the proposed topological method for MR image segmentation, in part IV the evaluation of the proposed method its results and findings about its efficiency are presented. Part V presents the work conclusions and the references are in part VI.

II. Related Work

Prostate segmentation methods based on images acquired using ultrasound, magnetic resonance and computed tomography could be generally divided into four major categories: contour and shape based methods, region based methods, supervised and un-supervised classification methods, hybrid methods [2].

Contour and shape based methods are using the boundary features to segment the prostate. This is very difficult problem since MRI exhibits high soft tissue contrast. To cope with this Zwiggelaar et al. [10] used first and second order Lindeberg directional derivatives, in a polar coordinate system to identify the edges. On the other hand, Samiee et al. [11] used prior information of the prostate shape to refine the prostate boundary. Without prior shape information segmentation was error prone and often significantly different from the anatomical structure. Therefore, Cootes et al. [12] proposed to segment prostate in MR slices using the active shape model (ASM). Slightly different approach which combines two and three dimensional ASMs to segment the prostate using MR images was proposed by Zhu et al. [13]. A three dimensional ASM was built that represented the shape variance of the prostate.

One of the commonly used methods for region based segmentation is the one which lies upon the set of manual segmentations of anatomical structures registered to a common coordinate frame called atlas, which is afterwards used as a reference. These methods are trying to map the pre-segmented images to the querying image by finding a one to one transformation. However, due to variations in image intensities and differences in shapes this matching remains to be a challenging research topic.

For this purposes, various multi-atlas segmentation methods have been analyzed in order to improve the selection of the atlas images which are most similar to the querying one [14]. It should be stressed that the weighting coefficients should favor the atlas images which are most similar to the querying one and thus should contribute more in the segmentation.

Having in mind this, Klein et al. [15] has proposed a multi-atlas approach to segment the prostate using localized mutual information. The registration of the training volumes to the querying one was performed using affine and non-rigid registration.

Álvarez et al. [16] improved this method by taking the advantage of both the inter-individual shape variation and intra-individual salient point representation.

Langerak et al. [17] focused their work pre-selection of atlases before registration by assigning them to clusters and registering only some of these clusters. They are analyzing and registering instances from each cluster and then combining them to an estimate of the target segmentation. By doing so, they claim to achieve the same accuracy with atlas reduction of even 60%.

Sjöberg and Ahnesjö [18] proposed a new multi-atlas based segmentation using probabilistic label fusion with adaptive weighting of image similarity measures. Namely, their method is based on probabilistic weighting of distance maps. Relationships between image similarities and segmentation similarities are estimated in a learning phase and used to derive fusion weights that are proportional to the probability for each atlas to improve the segmentation result.

Xie and Ruan [19] recently proposed a method where they first perform an affine registration to minimize the global mean squared error to coarsely align each atlas image to the target. Afterwards, they use a target-specific regional mean squared error, in order to select a relevant subset from the training atlas. Then non-rigid registration between the training images and the querying one are performed inside previously identified subset only. At the end, using the estimated deformation fields, structure labels are transferred from training to querying images and they are fused based on a weighted combination of regional and local mean squared error, with proper total-variation-based spatial regularization.

Makni et al. [20] proposed a modified alternative of the evidential C-means algorithm to cluster voxels in multispectral MRI, including T2 weighted, diffusion weighted and contrast enhanced images.

In contract to the previously mentioned methods, hybrid ones are combining a priory boundary and feature information. These methods are proven to give superior results in contrast to others in presence of shape and texture variations.

Vikal et al. [21] proposed a method for building an average shape model using the prior shape and size information from manually marked contours. In order to reduce the noise and enhance the contrast they used a stick filter. On the enhanced images they detected the edges by applying the Canny filter. The constructed average shape model was used to discriminate pixels which are out of the model orientation. By applying polynomial interpolation the contour was further refined. The segmented contours obtained in the middle slices were used to initialize other slices towards the peripheral zones in both directions.

III. METHODOLOGY

In order to enable accurate multi-atlas based prostate segmentation, the proposed methodology relies on most similar atlases which can provide robust and precise transformation to the target image. The proposed methodology consists of several steps as presented on the diagram in Fig. 1.

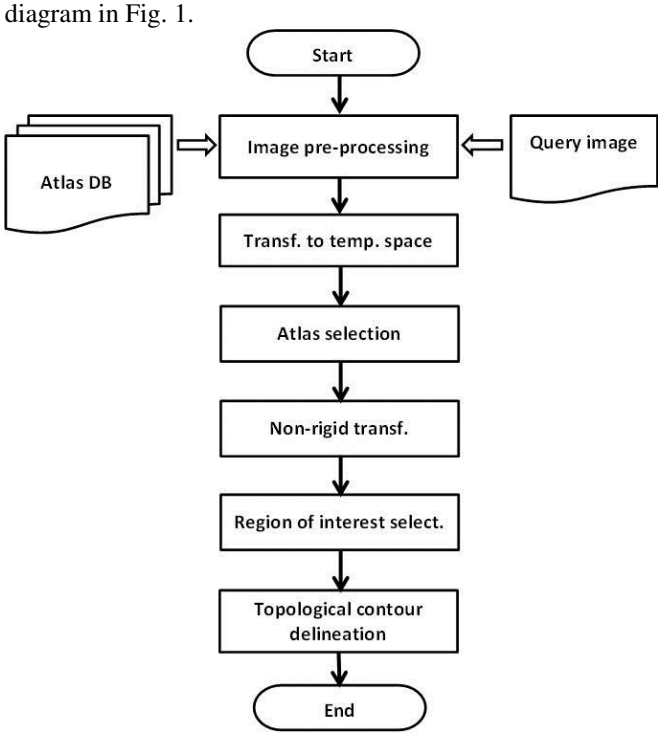


Fig. 1 Activity diagram which describes the whole methodology

Initial step of the proposed method are similar to those proposed in [15], namely appearance-specific atlas selections and a patch-based local weighting strategy for atlas fusion. After some image preprocessing which aims at inhomogeneity correction the top 5 similar atlases are selected for atlas registration based on the intensity differences in the surrounding region of the prostate. Once selected, the similar atlases are non-rigidly registered to a target image. Using the calculated transformation anatomical structure labels of the atlas are propagated to the space of the querying image. The aim of this step in our methodology is to order to derive a region of interest formed by the interception and union of the a priori shapes in the selected atlases, where the prostate contour of a non-segmented sample is supposed to be positioned. The final step is to delineate the prostate contour in the determined region by pixel classification. Namely, pixels within the region of interest are classified as prostate-likely-belonging pixels or prostate-unlikely-belonging pixels, taking into consideration the number of same-position pixels, being part of segmented samples and the intensity difference between a pixel of a non-segmented sample and the same-position pixels of segmented samples. The prostate contour is found as a set of pixels, separating column pixels (row pixels) within the region of interest in two disjunctive sets, having maximized the number of prostate-unlike-belonging pixels in the first set and the number of prostate-likely-belonging pixels in the other set exclusively. Prostate contour of a non-segmented sample is determined in three steps, described as follows:

Step 1: Determine prostate interception and union shape model, according to Definition 1 and Definition 2, over a set of segmented samples, acquiring prostate shapes' knowledge of n segmented same size and type prostate MR images.

Definition 1: A pixel $p[i, j]$ is an interception model pixel if:

$ss_1[i, j] \in$ prostate segment ss_1,
$ss_2[i, j] \in$ prostate segment ss_2,
…
$ss_{n-1}[i, j] \in$ prostate segment ss_{n-1},
$ss_n[i, j] \in$ prostate segment ss_n,
where $ss_i, 1 \le i \le n$ is a segmented sample.

Definition 2: A pixel $p[i, j]$ is a union model pixel if:

$ss_1[i, j] \in$ prostate segment ss_1 or
$ss_2[i, j] \in$ prostate segment ss_2 or
…
$ss_{n-1}[i, j] \in$ prostate segment ss_{n-1} or
$ss_n[i, j] \in$ prostate segment ss_n,
where $ss_i, 1 \le i \le n$ is a segmented sample.

According to previous definitions, if a pixel at position i,j is found as a prostate pixel in all segmented samples, then the pixels is considered as a part of the interception model. If at least one pixel at position i,j is found as a prostate pixel, then the pixel is a union model pixel. Interception pixels are considered as a part of the prostate of a non-segmented MR image.

Step 2: Classify each pixel within the union, but out of the interception as a prostate-likely-belonging pixel or prostate-unlikely-belonging pixel, exclusively according Eq. 1 and Eq.2.

If Eq.1 is satisfied,

$$(n - n_{plb})\text{diff}_{plb} \le (n - n_{pulb})\text{diff}_{pulb} \quad (1)$$

classify the pixel $p[i, j]$ as a prostate-likely-belonging pixel.

If Eq.2 is satisfied,

$$(n - n_{plb})\text{diff}_{plb} > (n - n_{pulb})\text{diff}_{pulb} \quad (2)$$

classify the pixel $p[i, j]$ as a prostate-unlikely-belonging pixel.

Equation 1, 2 parameters are the following ones:

n : Number of segmented samples.

n_{plb}: Number of segmented samples, where pixel at position i,j is part of the prostate segmented region.

n_{pulb}: Number of segmented samples, where pixel at position i,j is not a prostate pixel.

The intensity difference between a pixel $p[i, j]$ of a non-segmented sample and the mean intensity of pixels at position i,j, part of a prostate in segmented samples, is calculated according Eq. 3.

$$diff_{plb} = p[i, j] - (\sum_{k=1}^{n} ss_k[i, j]) / n_{plb} \quad (3)$$

where $ss_k[i, j] \in$ prostate segment of ss_k

The intensity difference between a pixel $p[i, j]$ of a non-segmented sample and the mean intensity of pixels at position i,j, out of the prostate in segmented samples, is calculated according Eq. 4.

$$diff_{pulb} = p[i, j] - (\sum_{k=1}^{n} ss_k[i, j]) / n_{pulb} \quad (4)$$

where $ss_k[i, j] \notin$ prostate segment of ss_k

Value-opposite differences: $(n - n_{plb})$ and $(n - n_{pulb})$ serve as a weight factor for pixel intensity differences: $diff_{plb}$ and $diff_{pulb}$. The smaller $(n - n_{plb})$ is, greater difference $(n - n_{pulb})$ is obtained. Relatively small pixel intensity difference $diff_{plb}$ increases pixel prostate belonging expectation. On the contrary, small pixel intensity difference $diff_{pulb}$ decreases pixel prostate belonging expectation. Combining previous parameters in a single equation (Equations 1, 2), a prostate pixel classifier is derived.

Step 3: Determine prostate contour shape as a set of pixel, separating the union, out of the interception in two disjunctive sets, such as the number of prostate-unlike-belonging pixels in the first set and the number of prostate-likely-belonging pixels in the other set is exclusively maximized.

Applying Equations 1, 2 for pixels of a non-segmented sample, out of the interception, but within the union, each pixel in the region is classified as a prostate-likely-belonging or prostate-unlikely-belonging pixel, exclusively.

Representing with 1 prostate-likely-belonging classified pixels, while with 0 prostate-unlikely-belonging classified pixels, the problem of identification of a prostate contour pixel is simplified to identification of prostate contour pixels, separating same row pixels (same column pixels), out of the interception, but within the union, in two disjunctive sets, such as the number of prostate-unlikely-belonging pixels and prostate-likely-belonging pixels in the sets is exclusively maximized.

For example, if $CP = \begin{bmatrix} p[i, j] \\ p[i+1, j] \\ p[i+2, j] \\ p[i+3, j] \\ p[i+4, j] \\ p[i+5, j] \end{bmatrix}$ is a six pixel same column set, out of the interception, but within the union, being accordingly classified as: $CCP = \{0,1,0,0,1,1\}$, there are 4 prostate contour candidate pixels, without taking into consideration the first and the last pixel. Pixel $p[i+1, j]$ separates classified set CPP in two disjunctive sets: $CPP_1 = \{0,1\}$, $CPP_2 = \{0,0,1,1\}$. The number of prostate-unlikely-belonging pixels in the first set is 1, while the number of prostate-likely-belonging pixels in the second set is 2. The sum equals 3. Similarly, pixel $p[i+2, j]$ separates set CPP in two disjunctive sets: $CPP_1 = \{0,1,0\}$, $CPP_2 = \{0,1,1\}$. Now the number of prostate-unlikely-belonging pixels in the first set is 2, while the number of prostate-likely-belonging pixels in the second set is 2. The sum equals 4. Choosing pixel $p[i+3, j]$ as a prostate contour pixel, the following disjunctive sets are obtained: $CPP_1 = \{0,1,0,0\}$, $CPP_2 = \{1,1\}$. The number of prostate-unlikely-belonging pixels in the first set is 3, while the number of prostate-likely-belonging pixels in the second set is 2. Their sum equals 5. Pixel $p[i+4, j]$ assumed as a prostate contour pixel, decreases the number of prostate-likely-belonging pixels in CPP_2, while the number of prostate-unlikely belonging pixels in CPP_1 remains unchanged.

Therefore, pixel $p[i+3, j]$ is chosen as a prostate contour pixel, since the sum of prostate-unlikely-belonging pixel and prostate-likely-belonging pixels in the disjunctive sets is maximized in that case (Fig.2).

-1	-1	-1
-1	8	-1
-1	-1	-1

Fig. 2 Point detection mask

If the condition given by Eq.5 is satisfied

$$R = \frac{1}{8}(8\,p[i, j] - p[i-1, j-1] - p[i-1, j] - p[i-1, j+1] - p[i, j-1] - p[i, j+1] - p[i+1, j-1] - p[i+1, j] - p[i+1, j+1]) > T \quad (5)$$

then the pixel p[i, j] is a prostate contour outlying pixel.

Discontinuous prostate contour curves are linked together applying standard image morphological operations, such as multiple Dilatation at first, then Erosion, in order to derive one pixel-thin prostate contour. Figure 3 represents the structuring element used in the morphological operations.

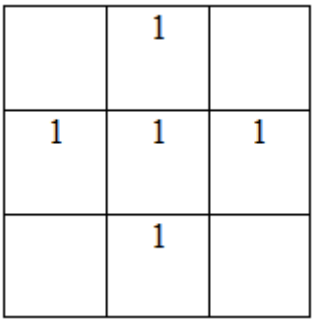


Fig. 3 Morphological operations' structuring element

IV. Experimental Evaluation and Results

For evaluation purposes all the steps described of the proposed methodology and presented in Fig. 1 are implemented in C# programming language. The program was executed on laptop with 4GB RAM memory and equipped with Intel Core i3 CPU with 2.4GHz and 64 bit Windows 7 OS. It has also ATI Mobility Radeon HD 4650 with 1GB dedicated memory. The proposed method was evaluated on 30 training MRI prostate images. The image series used for this evaluation were T2 FSE AXIALS 256x256 pixel. They were obtained from the online Prostate MR Image Database [22]. For each training image, manual segmentation is provided.

A leave-one-out study has been implemented based on each of the training scans using the remaining 29 images as the atlas database. In the sub-database, the top 5 most similar atlases are chosen. Based on these atlases the union and the interception shape model are constructed.

For better visual representation of the obtained results the following coloring convention was used: the red colored region represents the interception model, while the white colored region represents the union, out of the interception, Figure 4.

Pixels within the interception are considered as a part of the prostate. Each pixel within the white colored region is exclusively classified as a prostate-likely-belonging pixel or prostate-unlikely-belonging pixel.

Taking image 000046.00001.001.0013 from the Prostate MR Image Database (http://prostatemrimagedatabase.com/Database/000046/00001/001/0013.html) as a querying image, the result of the classification is shown on Figure 5. Red pixels, out of the interception, but within the union are classified as prostate-likely-belonging pixels, satisfying Equation 1, while the white pixels in the same region are classified as prostate-unlikely-belonging pixels, satisfying Equation 2.

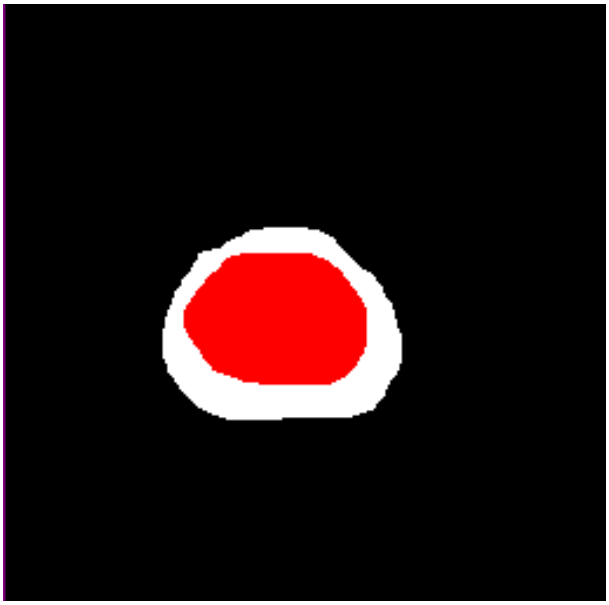


Fig. 4 Interception and union derived model

In Figure 5 blue colored pixels are the prostate contour pixels, determined in the third processing step of the proposed method.

Filtering prostate contour outlying pixels and applying standard region closing morphological operation, using Figure 3 structuring element, the prostate contour of a non-segmented T2 FSE AXIALS database image 000046.00001.001.0013 is obtained, Figure 6.

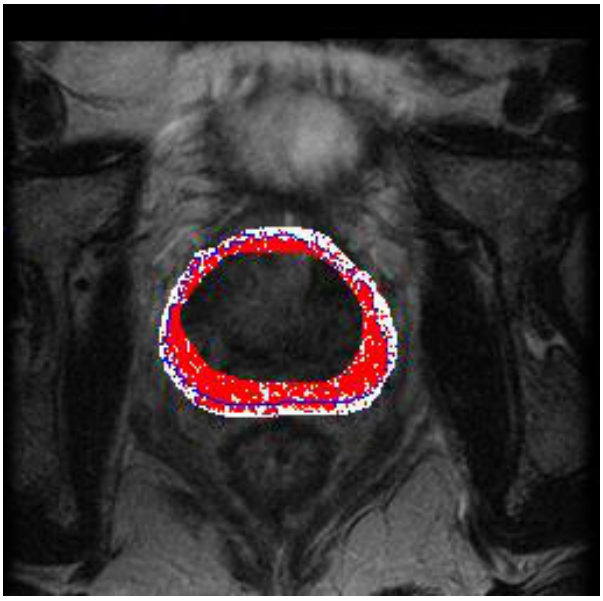


Fig. 5 Method application results

Fig. 6 Identified prostate contour (green), gold standard contour delineated manually by the expert (yellow).

For evaluation of the proposed method, we have used Dice similarity coefficient (DSC) defined as in Eq. 6 compared with the manual expert segmentation.

$$DSC = 2 \cdot \frac{A \cap B}{A + B} \qquad (6)$$

It has been applied on the apex, central and the base region of the prostate. The average of this metric with its standard deviations calculated from the selected image series compared with the corresponding manual segmentation is given in Table1.

TABLE I.
PROSTATE SEGMENTATION QUANTITATIVE RESULTS FOR TRAINING DATASETS

Region	**DSC**
Apex	0.81±0.13
Central	0.82±0.10
Base	0.79±0.17

V. CONCLUSION

Comparing the obtained results a conclusion for prostate shape accordance can be derived, based on the prostate edges' direction compatibility, prostate contour position and prostate surface.

In general, prostate segmentation result in this case depends of two factors. The number of segmented samples used and segmented samples' prostate shape variability, based on what the interception and the union shapes are determined. More segmented samples are considered, with wider prostate shape variability, more accurate prostate contour is obtained. A drawback of the proposed method is the incapacity of detecting prostate segments, out of the derived union region, being part of prostate of a non-segmented sample. On the opposite, prostate segmentation running time is significantly improved, since relatively small segment of a non-segmented prostate MR image is processed.

ACKNOWLEDGMENT

The research presented in this paper has been partially supported and conducted in the framework of the project "Development of novel algorithms and software library for biomedical engineering applications" funded by the University Goce Delcev – Stip.

REFERENCES

[1] GLOBOCAN 2012 (online at http://globocan.iarc.fr, last visited 19.04.2014).

[2] Ghose, S., Oliver, A., Martí, R., Lladó, X., Vilanova, J. C., Freixenet, J., Mitra, J., Sidibé, D., Meriaudeau, F. A survey of prostate segmentation methodologies in ultrasound, magnetic resonance and computed tomography images. Computer methods and programs in biomedicine, 108(1), 262-287, 2012, http://dx.doi.org/10.1016/j.cmpb.2012.04.006

[3] Kim, S. G., Seo, Y. G., A TRUS Prostate Segmentation using Gabor Texture Features and Snake-like Contour. Journal of Information Processing Systems, 9(1), 2013, http://dx.doi.org/10.3745/JIPS.2013.9.1.103

[4] Natarajan, S., Marks, L. S., Margolis, D. J., Huang, J., Macairan, M. L., Lieu, P., & Fenster, A. Clinical application of a 3D ultrasound-guided prostate biopsy system. In Urologic Oncology: Seminars and Original Investigations, Vol. 29, No. 3, pp. 334-342, June 2011, Elsevier, http://dx.doi.org/10.1016/j.urolonc.2011.02.014

[5] Woodruff, A.J., Morgan, T.M., Wright, J.L., Porter, C.R.., Prostate volume as an independent predictor of prostate cancer and high-grade disease on prostate needle biopsy. J Clin Oncol 26: 5165, 2008

[6] Catalona, W. J., Smith, D. S., Ratliff T. L., Dodds, K. M., Coplen, D. E., Yuan, J. J., Petros, J.A., Andriole, G. L., "Measurement of prostate-specific antigen in serum as a screening test for prostate cancer", New England Journal of Medicine, vol. 324, no. 17, pp. 1156–1161, 1991, http://dx.doi.org/10.1056/NEJM199104253241702

[7] Halpern, E. J., Cochlin, D. L., Goldberg, B. B., Imaging of the prostate, Informa Healthcare, United Kingdom, first edition, 2002

[8] Ozer, S., Langer, D. L., Liu, X., Haider, M. A., Van der Kwast, T. H., Evans, A. J., Yang, Y., Wernick, M. N., Yetik,I. S., Supervised and Unsupervised Methods for Prostate Cancer Segmentation with Multispectral MRI, Medical Physics 37 (2010) 1873–83, http://dx.doi.org/10.1118/1.3359459

[9] Puech, P., Betrouni, N., Makni, N., Dewalle, A.S., Villers, A., Lemaitre, L., Computer-Assisted Diagnosis of Prostate Cancer Using DCE-MRI Data: Design, Implementation and Preliminary Results, International Journal of Computer Assisted Radiology and Surgery 4 (2009) 1–10, http://dx.doi.org/10.1007/s11548-008-0261-2

[10] Zwiggelaar, R., Zhu, Y., Williams, S., Semi-Automatic Segmentation of the Prostate, in: F. J. Perales, A. J. Campilho, N. P. de la Blanca, A. Sanfeliu (Eds.), Pattern Recognition and Image Analysis, Proceedings of First Iberian Conference, IbPRIA, Springer, Berlin and Heidelberg and New York and Hong Kong and London and Milan and Paris and Tokyo, 2003, pp. 1108–16, http://dx.doi.org/10.1007/978-3-540-44871-6_128

[11] Samiee, M., Thomas, G., Fazel-Rezai, R., Semi-Automatic Prostate Segmentation of MR Images Based on Flow Orientation, in: IEEE International Symposium on Signal Processing and Information Technology, IEEE Computer Society Press, USA, 2006, pp. 203–7, http://dx.doi.org/10.1109/ISSPIT.2006.270797

[12] Cootes, T. F., Hill, A., Taylor, C. J., Haslam, J., The Use of Active Shape Model for Locating Structures in Medical Images, Image and

Vision Computing 12 (1994)355–66, http://dx.doi.org/10.1007/BFb0013779

[13] Zhu, Y., Williams, S., Zwiggelaar, R., A Hybrid ASM Approach for Sparse Volumetric Data Segmentation, Pattern Recognition and Image Analysis 17 (2007) 252–8, http://dx.doi.org/10.1134/S1054661807020125

[14] Ghose, S., Oliver, A., Marti, R., Llado, X., Freixenet, J., Vilanova, J. C., Meriaudeau, F. A probabilistic framework for automatic prostate segmentation with a statistical model of shape and appearance. In Image Processing (ICIP), 2011 18th IEEE International Conference on (pp. 713-716). 2011, September IEEE, http://dx.doi.org/10.1109/ICIP.2011.6116653

[15] Klein, S., Van der Heide, U. A., Lipps, I. M., Vulpen, M. V., Staring, M., Pluim, J. P. W.,"Automatic Segmentation of the Prostate in 3D MR Images by Atlas Matching Using Localized Mutual Information", Medical Physics 35 (2008) 1407–17, http://dx.doi.org/10.1118/1.2842076

[16] Álvarez, C., Martínez, F., Romero, E., "A novel atlas-based approach for MRI prostate segmentation using multiscale points of interest ", Proc. SPIE 8922, IX International Seminar on Medical Information Processing and Analysis, 89220O (November 19, 2013), http://dx.doi.org/10.1117/12.2035462

[17] Langerak, T. R., Berendsen, F. F., Van der Heide, U. A., Kotte, A. N., Pluim, J. P. Multiatlas-based segmentation with preregistration atlas selection. Medical physics, 40(9), 2013, 091701, http://dx.doi.org/10.1118/1.4816654

[18] Sjöberg, C., & Ahnesjö, A. Multi-atlas based segmentation using probabilistic label fusion with adaptive weighting of image similarity measures. Computer methods and programs in biomedicine, 110(3), 2013, pp.308-319, http://dx.doi.org/10.1016/j.cmpb.2012.12.006

[19] Xie, Q., Ruan, D. Low-complexity atlas-based prostate segmentation by combining global, regional, and local metrics. Medical physics, 41(4), 2014, 041909, http://dx.doi.org/10.1118/1.4867855

[20] Makni, N., Iancu, A., Colot, O., Puech, P., Mordon, S., Betrouni, N., et al.: Zonal segmentation of prostate using multispectral magnetic resonance images. Medical Physics 38(11), 6093 (2011), http://dx.doi.org/10.1118/1.3651610

[21] Vikal, S., Haker, S., Tempany, C., Fichtinger, G., Prostate Contouring in MRI Guided Biopsy, in: J. P. W. Pluim, B. M. Dawant (Eds.), Proceedings of SPIE Medical Imaging: Image Processing, SPIE, USA, 2009, pp. 7259–72594A, http://dx.doi.org/ 10.1117/12.812433.

[22] Prostate MR Image Database, The Brigham and Women's Hospital, 2008 (Online at: http://prostatemrimagedatabase.com; last accessed 19.04.2014).

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 227–234
DOI: 10.15439/2014F338
ACSIS, Vol. 2

EMG Speller with Adaptive Stimulus Rate and Dictionary Support

M. Vasiljevas, R. Turčinas, R. Damaševičius
Software Engineering Department, Kaunas University of Technology, Kaunas, Lithuania
Email: {mindaugas.vasiljevas, rutenis.turcinas, robertas.damasevicius}@ktu.lt

Abstract—Ambient Assisted Living (AAL) aims to improve the quality of daily life for all humans in different periods of life. Neural-Computer Interface (NCI) can be used within AAL environments to provide alternative communication means for impaired persons bypassing the need for speech and other motor activities. By monitoring, analyzing and responding to muscular activity (EMG signals) of users, NCI systems are able to monitor, diagnose and respond to the cognitive, emotional and physical states of users in real time. In this paper we analyze and develop a speller application based on the EMG interface. We analyze requirements for developing interfaces for disabled users and interfaces of known speller applications, and describe the development of the EMG-based speller as a benchmark application. The developed speller has adaptive stimulus rate and allows word selection from dictionary. We evaluate performance and usability of the developed speller using a set of empirical (accuracy, information transfer speed, input speed), ergonomic (NASA-TLX scale) and conceptual (humanistic intelligence) attributes.

I. Introduction

Ambient Assisted Living (AAL) environments comprise assisted technology devices, communication protocols and interfaces used to improve the quality of daily life for humans in different periods of their life [1]. Considering predictions of the demographic changes in society, AAL particularly focuses on elderly people though people with minor disabilities such as motor impairments can benefit, too. The AAL systems are user-centered and specifically are based on the concept of User Interfaces for All [2]. The concept aims at efficiently and effectively addressing the accessibility problems in human interaction with software applications and services while meeting the individual requirements of the users in general, including disabled and elderly people. Following the vision of e-Inclusion, the aim to "leave no-one behind" when enjoying the benefits of information and communication technology [3].

In the AAL environments, Neural-Computer Interfaces (NCI) can be used to provide alternative communication means for persons with disabilities bypassing the need for speech and other motor activities. NCI is similar to Brain-Computer Interface (BCI) in methods used as well as in applications, however it uses the Electromyography (EMG) data rather than the Electroencephalography (EEG) data to establish an interface between human peripheral neural system and computers by recording electrical signals governing muscular movements of a subject. The concepts are particularly suited to the needs of the handicapped as the cores of the smart environments and virtual reality applications. The state of a user is captured using sensors attached to the body. Then a physiological computing system creates a bio-cybernetic neurofeedback loop involving both human users and computers [4], which allows to produce a representation of the user's operational context. The loop may be designed to offer assistance if the user is frustrated or unable to perform the task due to excessive mental workload, adapt the level of challenge to sustain or increase task engagement if the user is bored or demotivated by the task, incorporating an emotional display element into the user interface, or alert for help if the user is not responsive [5].

When developing NCI systems for older adults (over 60) one has to consider that older people often have multiple, minor motor and cognitive function impairments or have slow control over their motor functions [6]. Given the often reported lower skin hydration in the elderly, the skin conductance is lower which leads to lower amplitudes and signal-to-noise ratio of measured EMG signals [7]. They also may also have slower control over muscle activities of the hands, fingers, etc. and decision-making may also be slower. For this group, the motivation to use NCI is completely different from the first group. Therefore, the design of a NCI for older adults should reflect their non-typical EMG profiles or slower response times.

Speller is a typical example of NCI/BCI application, which establishes a communication channel for people unable to use traditional keyboard and still remains a benchmark for BCI and NCI methods [8]. The speller is aimed to help those disabled persons unable to activate muscles traditionally used in communication (hands, tongue) to spell words by utilizing their neural activity. Typically, spellers use signal amplitude information, however integrating it with signal processing methods such as noise and dimensionality reduction methods and user intent prediction techniques can improve the results [9].

In this paper, we analyze the requirements for developing interfaces for impaired users and visual interfaces of known speller applications, describe the development of a speller as a typical benchmark application, and evaluate its performance and usability of the developed speller using a set of empirical (accuracy, information transfer speed, input speed), ergonomic (NASA-TLX scale) and conceptual (humanistic intelligence) attributes.

This work was not supported by any organization

The structure of the remaining parts of the paper is as follows. Section II analyzes the requirements for NCI systems and, specifically, NCI spellers. Section III discusses the interfaces of speller applications. Section IV describes the development of speller application. Section V presents the experimental results. Section VI evaluates the results. Finally, Section VII presents conclusions.

II. Analysis of Requirements for Speller Application

The requirements for speller application can be categorized at different levels depending upon the physical abilities of its users [10]: 1) Users with no physical disability, who may use NCI for entertainment or other conditions where physical movement is restricted. 2) Users with minor impairments (such as older persons). 3) Users with severe physical disabilities, who may wish to use NCI as a secondary input. 4) Users who are almost locked-in (having limited muscle control), who may need to use NCI as a method for communication.

First, the speller must follow general requirements for smart systems to be integrated into the AAL environments. Next, the specific requirements for impaired users (and, specifically, for older persons) must be followed. Impaired users need assistance such as automatic learning of user's behavior to estimate his/her current needs.

Since humans often make mistakes or errors in interacting with machines, for any human-operated system, user interfaces should be designed such that prevent errors whenever possible; deactivate invalid commands; make errors easy to detect and show users what they have done; and allow undoes, reverse, correct errors easily [11].

For smart systems, the following principles (also called "operational modes") of Humanistic Intelligence Framework [12] must be satisfied:

1) **Constancy:** the interface should operate continuously to read signals from human to computer and to provide a constant user-interface.

2) **Augmentation:** the primary task is increasing the intelligence of the system rather than computing tasks.

3) **Mediation:** the interface mediates between human senses, emotions and perceptions and acts as information filter by blocking or attenuating undesired input to decrease negative effects of interaction (such as fatigue, information overload, etc.) as well as to increase positive effects (such as user satisfaction) by amplifying or enhancing desired inputs.

According to Lopes [13], the user interface for persons with disabilities must: support user variability allowing to provide the means to adapt to user-specific requirements; support of a wide range of input devices and output modes; provide minimal user interface design; promote interaction and retain user attention on the tasks; and provide strong feedback mechanisms that may provide rewarding schemes for correct behavior (results).

The requirements for interfaces for impaired users can be formulated as follows [1]: 1) **Limited access to details:** complex and vital details of the system have to be hidden to avoid user overwhelming and trapping. 2) **Self-learning:** detected common patterns in the behavior of the user should be used to automatically create rules or shortcuts that speed and ease up the use of the system. 3) **System interruption:** Impaired users have in most cases no idea how the system is working, therefore easy cancellation of system's activities must be ensured.

In the questionnaire-based study of potential BCI user requirements towards assisted technologies [14], the participants rated participants rated "functionality" (aka effectiveness) as the most important requirement, followed by "possibility of independent use" and "easiness of use".

III. Review of Speller Interfaces

Many different variants of interfaces have been proposed and designed for speller, a de-facto benchmark application of BCI/NCI. Based on the complexity and visual representation of symbols to input, they can be categorized into the following classes:

Linear (or single character) speller: all symbols are shown and each symbol is flashed individually until symbol selection is done [15].

Matrix (or Row-Column) Speller: All letters are arranged in a matrix . First, speller flashes an entire column (Fig. 1, left) or row of characters (Fig. 1, right). Then, single letters are flashed in a sequence, and can be selected. Different matrix sizes can be used, e.g., a 6x6 matrix containing all 26 letters of the alphabet and 10 digits (0-9) [16], or even a full QWERTY keyboard [17].

Fig. 1. Example of matrix speller interface.

Chekerboard Speller [18]: the 8x9 matrix is virtually superimposed on a checkerboard (Fig. 2, left), which the participants never actually see. The items in white cells of the 8 x 9 matrix are segregated into a white 6 x 6 matrix and the items in the black cells are segregated into a black 6 x 6 matrix. Before each sequence of flashes, the items in Fig. 2 (left) randomly populate the white or black matrix, respectively, as shown in Fig. 2 (middle). The checkerboard layout controls for adjacency-distraction errors, because adjacent items cannot be included in the same flash group. The users see random groups of six items flashing (as opposed to rows and columns) because the virtual rows and columns depicted in Fig. 2 (middle) flash.

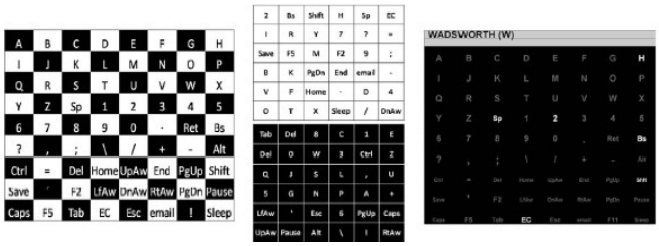


Fig. 2. Checkerboard intherface of the speller [18]

Frequency-based layout accounts for the relative frequency of character occurrence in a language as in Bremen SSVEP-based BCI-speller [19]. It has in the middle of the screen a virtual keyboard with 32 symbols (see Fig. 3) surrounded by five boxes flickering at different frequencies. These boxes correspond to commands for navigating the cursor (indicated by red color), and for selecting the intended character. The application starts with a cursor in the central position corresponding to the most frequent character in English (i.e., "E"). Letters with the higher frequency of occurrence are positioned closer to the center while the less frequent ones are further away. The user can navigate the cursor to the desired letter and confirm his/her choice with the "Select" command. The further the character is located from the center, the more command selections (cursor movements) are required.

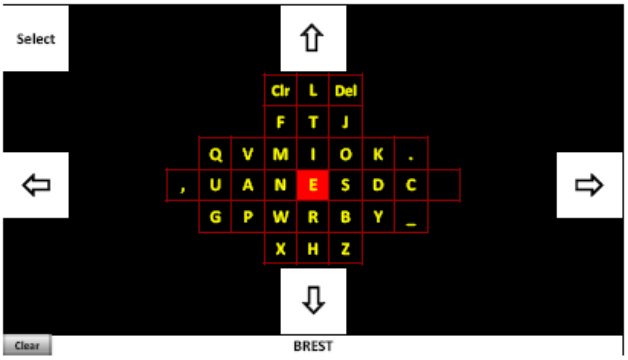


Fig. 3. Interface of the Bremen BCI speller [19].

Region-based Speller: groups of characters arranged into different regions, which contain different subsets of characters (Fig. 4). When the user confirms the selection of the group, the characters of the group are divided into new groups until the desired character is selected. Examples of such interface are 27 symbol triads [20] or 64 symbol quadrants [21].

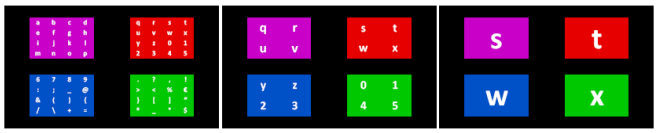


Fig. 4. Consecutive stages to select symbol in the quadrant region-based speller [21]

The Rotate-Extend (REx) paradigm [10] consists of a wheel divided into segments (see Fig. 5). An arrow in the centre of the wheel controls the selection of target segments. One mental class is used to control the rotation of the arrow, and the other class extends the arrow to select the target segment. Example of Rex interface is Hex-o-Spell speller [22], which allows 30 different characters to be typed in. The characters are shown in six adjacent hexagons distributed around a circle. Each hexagon contains five characters and a „go back" command. For the selection of the hexagons, there is an arrow in the center of the circle. After selection, the characters in all hexagons, except for the selected one disappear, while the remaining characters and the „go back" command are mapped into six hexagons around the circle. Using the same arrow-based strategy, the user selects the desired character or decides to go back to the previous level of the interface to correct a mistake. Another implementation of REx interface is Oct-o-spell, where a larger set of symbols is used [23].

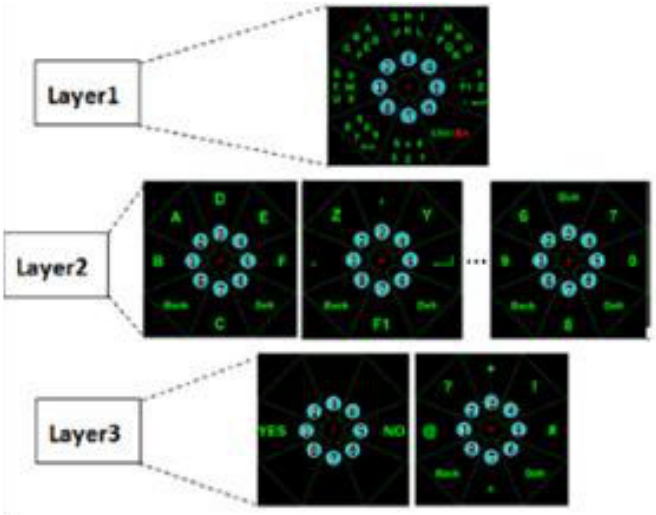


Fig. 5. Interface of Oct-o-spell speller [23]

The overview of visual interfaces of spellers can be summarized as follows. The interfaces can classified according to interface paradigm (linear/single, matrix, chekerboard, frequency layout, region-based, Rotate-Extend), stimulus type (the way each individual character changes, e.g., flashing frequency, color change, distance to target, etc.), stimulus rate (the speed at which individual characters change), stimulus pattern (grouping of symbols in interface), character set (usually includes all letters of the alphabet as well as additional symbols such as numbers, separation marks, etc., enhancing intelligence (additional techniques for improving accuracy of the system and rate of communication such as using language model, word autocomplete, spelling correction or word prediction).

IV. Development of EMG speller

A NCI system generally comprises the following components: (i) a device that records the muscular activity signals; (ii) a signal preprocessor that reduces noise and artifacts; (iii) a decoder that classifies the de-noised signal into a control commands for (iv) an external device or application (e.g., a robotic actuator, a computer program etc.), which provides feedback to the user [24].

Our speller application has three layers as follows: 1) on the lowest layer, the physiological signal is sampled into a data stream of physiological data. Downsampling can be used to decrease amount of data and increase information processing speed at higher levels. 2) On the intermediate layer, data is aggregated and events corresponding to specific patterns of data are generated. Machine learning techniques such as artificial neural networks may be used to recognize such events and generate decisions. 3) On the highest layer, decisions are processed and used to generate control commands for external applications (systems).

The architecture of the developed speller application is shown in Fig. 6. The speller has 6 components: MainReader is responsible for control of data reader which is selected to use. ReaderAPI is public external interface of third-party EMG data reader modules. MainController is responsible for selected control module (executes commands). NiaReader is a third-party module implemented for the OCZ NIA data reader device. SpellingSquare is a third-party module implemented for text input in symbol matrix using the EMG-based commands.

The dashed rectangle separates system components from external components. Components inside the rectangle are considered as system components. Components outside the rectangle are considered as external components. System was developed with respect to maintenance so that external components were easy to add or remove. The external components are sensors (EMG readers), actuators (robot, etc.) controllers or external software. The NetBeans framework was used for development. It provides the opportunity to add third-party components on-demand.

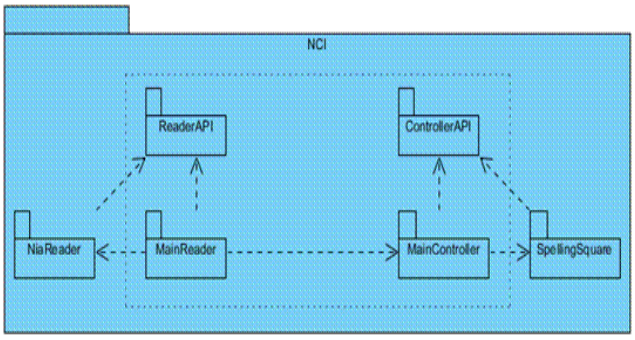


Fig. 6 Architecture of physiological computing system

The feedback to the user is an important aspect of the NCI as it provides the user with information about his/her mistakes as well as motivates the user to increase attention and engagement in the task. The main element of the developed speller application that supports feedback is visual interface (Fig. 7). It contains the representation of the symbol matrix (the size and character set of the matrix is adaptable). The red-colored column indicates the current position of the speller cursor. The cursor moves sequentially from column to column until the user activates the "Select" command. Next, the cursor moves through each symbol in the selected column. After another "Select" command the particular symbol is selected and appears in the text output area. The stimuli rate (the speed of the cursor can vary from 500 to 1500 ms) depends upon the number of input mistakes the user does (the speed increases or decreases automatically to keep the number of mistakes low). The mistake is considered as the "Cancel" command, which exits the selected column or deletes the selected symbol.

The control commands are initiated by the movements of facial muscles (left eye blink for "select" and right eye blink for "cancel"). The user can see the EMG signal feedback in the EMG signal view area of the interface (Fig. 8). The particular control command is performed when the amplitude of the EMG signal is equal or higher than the specified threshold value (marked with yellow horizontal lines). The upper threshold (high positive amplitude) indicates the "Select" command, while the lower threshold (high negative amplitude) indicates the "Cancel" command. The threshold values can be adjusted by the user via settings.

The signal view of EMG, while spelling the word "hello", is presented in Figs. 8 & 9. In Fig. 8, the word "hello" is spelled without mistakes. In Fig. 9, the same example is presented but in this case it contains a few mistakes (wrong selections). For correction of those mistakes cancellation commands must be performed. The spikes indicate the "select" command. One trial (selection of one character) contains two positive signal spikes, the first spike is for column selection, the second for letter selection in the corresponding column.

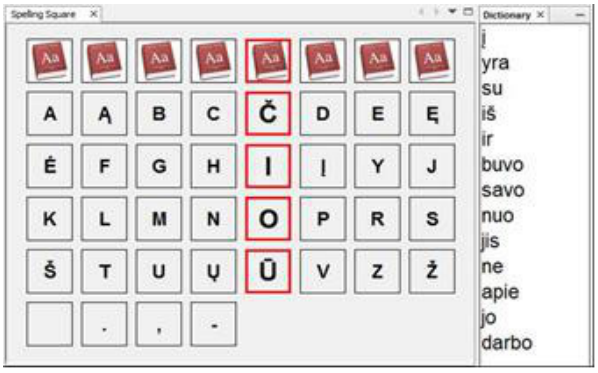


Fig. 7. Interface of developed speller application

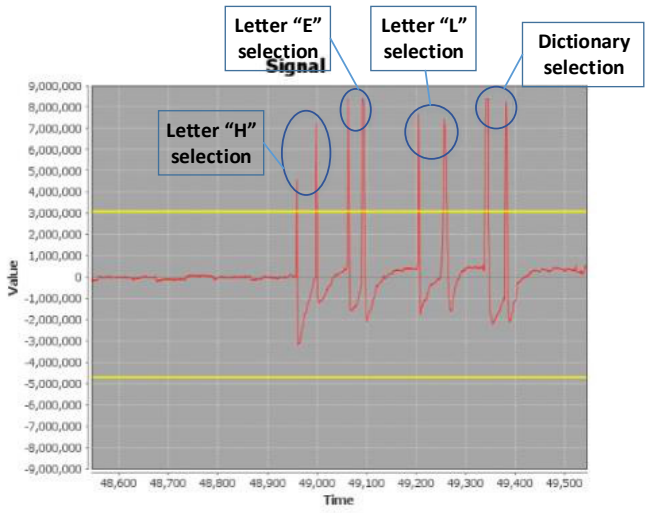


Fig. 8. Signal view of spelling word "hello". In this example no spelling mistakes were made and only three characters ("hel") were selected from the symbol matrix. Dictionary selection was made to complete the word.

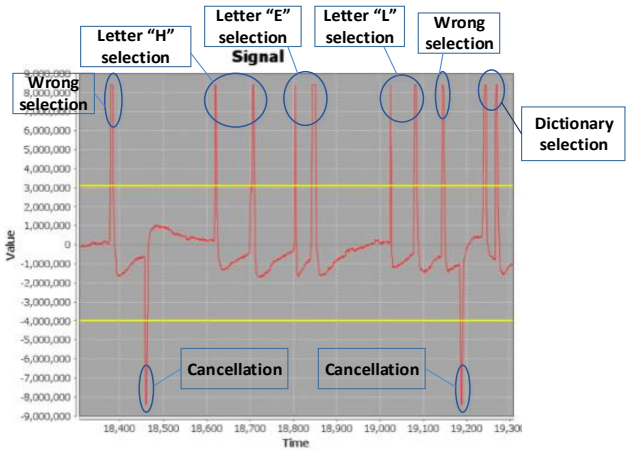


Fig. 9. Signal view of spelling word "hello". This example contains two spelling mistakes therefore after each wrong selection cancellation command was performed. Three characters ("hel") were selected from symbol matrix. Dictionary selection was made to complete the word.

The external dictionary can be used while entering symbols. The dictionary words are filtered using the already entered part of the word and shown to the user based on their frequency in the text corpora of the given language. The frequency value of the word is increased based on its usage frequency by the user. If the user enters a word absent in the dictionary, the dictionary is updated.

V. Experimental Setting

The experiments we performed with 5 subjects (3 males, 2 females), aged 24–54 (mean = 33) year. Subjects did not have any neurological abnormalities, reported normal or corrected to normal vision, and did not use medication. All subjects gave informed consent prior to the experiment.

EMG was recorded using OCZ Neuro Impulse Actuator equipment, which employs three electrodes across the forehead (see Fig. 10). It uses carbon interface fibers injected into a soft plastic as sensors to capture a combination of muscle, skin, and nerve biopotentials.

Fig. 10. The test subject with OCZ NIA device

The test subjects were seated in front of a table, 100 cm away from the liquid crystal display (LCD) showing stimuli. Visual stimuli were presented on a 13.3" size LCD screen with 1360 × 768 pixel resolution and a refresh rate of 60 Hz. Contrast and brightness are set to maximum. The size of each character was 1.5 × 1.5 cm (0.86 × 0.86° visual angle) and the entire speller matrix was 9.5 × 13 cm (5.44 × 7.42° visual angle). Stimuli consisted of intensifications of the rows and columns in sequential order. Intensification was achieved by increasing the size of all characters in the row or column with a factor 500 for 1500 ms. A trial is defined here as spelling of one character. All trials started with the speller being displayed on the screen, together with an instruction indicating which letter to select. Each stimulation sequence was followed by feedback on the screen, showing which letter or group of letters had been selected.

Three text paragraphs were given to the experiment participants. Their task was to input the proposed text paragraphs using speller. All text paragraphs were presented in Lithuanian. The first text paragraph contained 126 characters and its content covered a daily conversation. The second text paragraph contained 111 characters and its content covered a scientific speech. The third text paragraph contained 120 characters and covered a scientific speech with mathematical equations. Each experiment participant repeated the experiment 4 times with different speller settings. The first test was made with basic speller settings. The second test was made with adaptable stimulus time (the stimulus time varied from 500 ms to 1500 ms depending on the amount of mistakes). The third test was made with dictionary. The fourth test was made with both dictionary and adaptable stimulus time. The average accuracy, input speed and bit rate values were calculated. The results of experiments are presented in Section VI.

VI. Evaluation of Results

Quantitatively, the performance of speller application can be evaluated using accuracy, information transfer speed and input speed metrics. Accuracy is calculated as the percentage of correct decisions. Information transfer rate (or bit rate) indicates how much information can be communicated per time unit and is calculated using Walpaw's formula [25]. Finally, input speed is measured as the average time required to input a set of benchmark texts.

The accuracy results of BCI/NCI-based speller applications achieved by other authors are within 80-95% range (80% using EEG-based P300 speller [26], 82.77% using ECoG [9], 84.22% using invasive BCI [27], 87.58% using SSVEP based BCI [17], 87.8% for EOG-based speller [28], 89.5% [29], 91.80% [30], 94.8% for RSVP based speller [31], 90.81% for SSVEP-based speller [32], 95.18% for Oct-o-spell [23]).

The information transfer rate (aka bit rate) of the BCI/NCI-based speller applications achieved by other authors are within 7-41 bits/min (7.43 bits/min [33], 17.13 bits/min [29], 19.18 bits/min [30], 11.58-37.57 bits/min [32], 40.72 using SSVEP based BCI [17], 41.02 using ECoG [9]).

The symbol input speed of the BCI/NCI-based speller applications achieved by other authors are within 1-12 CPM (1.38 CPM for EOG-based speller [28], 1.43 CPM for RSVP based speller [31], 4.33 CPM [30], 4.91 CPM [32], 9.39 CPM using SSVEP based BCI [17], 10.16 CPM [23], 12.75 CPM [34]).

The results of the evaluation developed speller are given in Table I and summarized in Figs. 11-13. Best results in terms of both average and peak information transfer rate and input rate values are achieved when adaptable stimulus rate is used together with the dictionary. However, higher input speed inevitably lead to larger number of errors, therefore, accuracy is lower than using the speller with basic settings.

Table I.
Evaluation of Speller Application

Quantitative metric	Average Value	Peak value
BASIC SETTINGS		
Accuracy	96.29	98.25
Information transfer rate	34.78	41.83
Input speed	6.37	7.57
ADAPTABLE STIMULUS RATE		
Accuracy	88.61	93.64
Information transfer rate	42.53	49.79
Input speed	8.19	9.60
WITH DICTIONARY		
Accuracy	92.65	96.06
Information transfer rate	43.55	49.26
Input speed	8.22	9.35
WITH ADAPTABLE STIMULUS RATE AND DICTIONARY		
Accuracy	89.16	92.53
Information transfer rate	58.69	65.53
Input speed	11.35	12.42

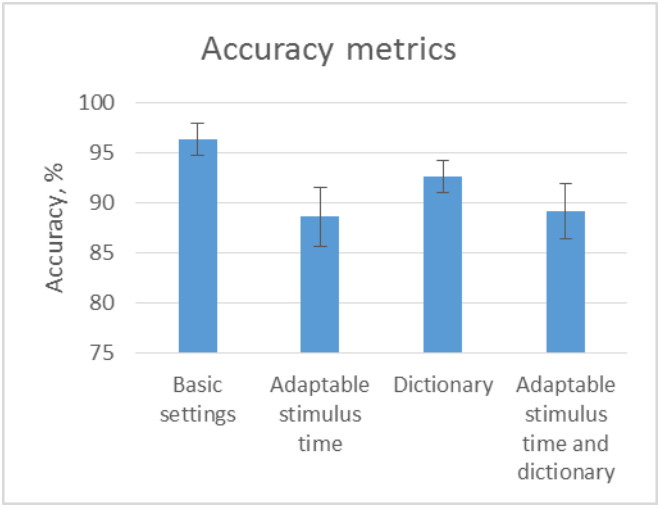


Fig. 11. Accuracy.

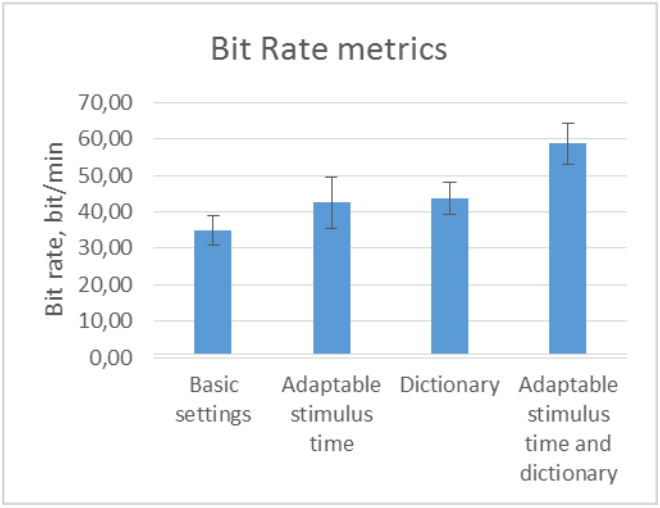


Fig. 12. Information transfer rate

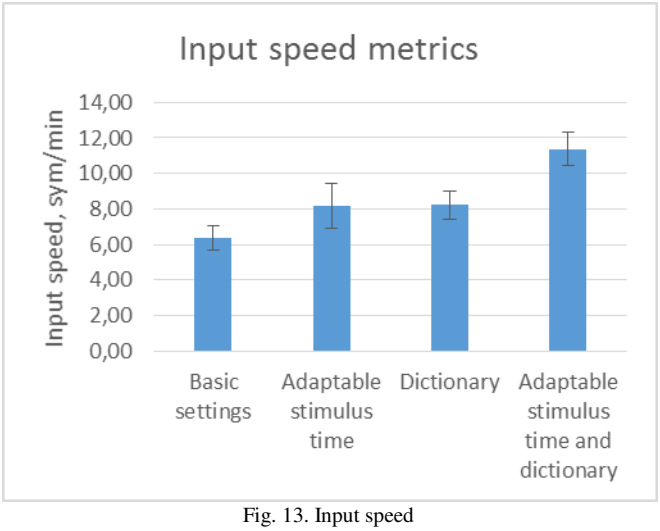


Fig. 13. Input speed

Qualitatively, the speller application can be evaluated based on user's mental workload required to work with this application. Here we use the NASA Task Load Index (TLX) questionnaires [35], a multi-dimensional rating procedure with six subscales: Mental Demands, Physical Demands, Temporal Demands, Performance, Effort, and Frustration.

TABLE II.
EVALUATION OF SPELLER APPLICATION USING NASA-TLX INDEX

NASA-TLX scale	Average value
Mental Demands	52
Physical Demands	66.6
Temporal Demands	45
Performance	48
Effort	67
Frustration	55

The users' subjective workload was assessed with the NASA Task Load Index (TLX), which identifies (1) the overall workload in the different tasks and (2) the main sources of workload. Workload in the TLX is defined as a "hypothetical construct that represents the cost incurred by a human operator to achieve a particular level of performance." The TLX is specifically adequate when interested in detecting the sources of workload. Workload is estimated with six subscales (mental, physical, and temporal demand and performance, effort, and frustration). Participants rated subjective workload for each dimension on twenty step bipolar scales with scores from 0 to 100. A weighting procedure was used to combine the six individual ratings into a global score. To do so, the six scales were combined to 14 pairs and subjects had to indicate which scale of the pair contributed more to their workload. A weighted average technique was then used to compute an overall measure of workload (between 0 and 100) and the relative contribution of each subscale to overall workload.

The NASA-TLX contains six factors (shown in Table III), each of which has 20 levels and is scored from 0 to 100. Small score represents low workload and vice versa. The speller application was evaluated by 5 healthy subjects and the results are presented in Table III. Main sources of workload were physical demands and effort. The average score for the workload factors are less than 56%. This observation indicates that the speller interface is acceptable for all subjects. High scores for "Temporal demand" given by most of the subjects indicate that the speed of the system must be improved further. Low scores for "Frustration" indicate that subjects are interested in using speller application and that the results meet their expectations.

TABLE IV.
EVALUATION OF SPELLER APPLICATION BASED ON THE MANN'S ATTRIBUTES OF HUMANISTIC-INTELLIGENCE SYSTEM

Attribute	Evaluation	Comment
Unmonopolizing	Yes	Speller does not cut the user off from the outside world
Unrestrictive	Yes	User can use other channels of communication at the same time while using the speller
Observable	Yes	Speller can get the user's attention continuously if and the output medium is constantly perceptible
Controllable	Yes	User can control the speller anytime
Attentive	Yes	Speller is context aware, multimodal, and multisensory
Communicative	Yes	Speller allows to communicate directly to other users or spellers

Conceptually, the speller application can be evaluated based on the attributes (Unmonopolising, Unrestrictive, Observable, Controllable, Attentive, Communicative), which every humanistic-intelligence system must have, as formulated by Mann [12] (see Table IV).

VII. CONCLUSION

In this paper we have described the development of the speller application for an assisted living environment using the EMG interface. This system is controlled by voluntary muscular movements, particularly the orbicular ones (i.e., eye blinking). These movements are translated into instructions which allow the text input.

The developed speller application is adaptive (text input speed can be adapted dynamically in response to the user's state) and intelligent (machine learning techniques are used to analyze input data to achieve high accuracy of selection as well as to increase text input speed by using word complete and word frequency features).

The speller has been evaluated empirically (using accuracy, information transfer speed and input speed), ergonomically (using the NASA-TLX scale of subjective workload) and conceptually (using the attributes of Mann's Humanistic Intelligence Framework [12]).

The achieved empirical results are within range of results achieved by other authors, while the ergonomic evaluation suggests that users are interested in using speller application and that the results meet their expectations, yet the speed of the system as well as the usability of its interface could be improved further.

This system can aid people with reduced mobility, extending the time that older people and disabled people can live in their home environment, increasing their autonomy and their confidence.

Future work will focus on going beyond low-level typing to graphical-symbol matrix that allows selection of concepts rather than stand-alone letters. Also the performance characteristics of the speller application will be researched further aiming to maximize usability of the product both in terms of increased speed as well as better ease of use.

ACKNOWLEDGMENT

The work described in this paper has been carried out within the framework the Operational Programme for the Development of Human Resources 2007-2013 of Lithuania „Strengthening of capacities of researchers and scientists" project VP1-3.1-ŠMM-08-K-01-018 „Research and development of Internet technologies and their infrastructure for smart environments of things and services" (2012-2015), funded by European Social Fund (ESF).

The authors would like to acknowledge the contribution of the COST Action IC1303 – Architectures, Algorithms and Platforms for Enhanced Living Environments (AAPELE).

REFERENCES

[1] A. Marinc, C. Stocklöw, A. Braun, C. Limberger, C. Hofmann, and A. Kuijper, „Interactive personalization of ambient assisted living environments," Proc. of the 2011 Int. Conf. on Human interface and the management of information - Volume Part I (HI'11), LNCS vol. 6771, Springer-Verlag, Berlin, Heidelberg, 2011, 567-576. DOI: 10.1007/978-3-642-21793-7_64

[2] C. Stephanidis, "Towards User Interfaces for All: Some Critical Issues," Advances in Human Factors/Ergonomics, 01/1995, 20:137-142. DOI:10.1016/S0921-2647(06)80024-9

[3] C. Zickler, A. Riccio, F. Leotta, S. Hillian-Tress, S. Halder, E. Holz, P. Staiger-Sälzer, E.J. Hoogerwerf, L. Desideri, D. Mattia, and A. Kübler, "A brain-computer interface as input channel for a standard assistive technology software," Clinical EEG and Neuroscience, 2011, 42(4), 236-44. DOI:10.1177/155005941104200409

[4] N.B. Serbedzija, and S.H. Fairclough, "Biocybernetic loop: From awareness to evolution," IEEE Congress on Evolutionary Computation, 2009, 2063-2069. DOI: 10.1109/CEC.2009.4983195

[5] S.H. Fairclough, "Fundamentals of physiological computing", Interacting with Computers (IWC), 2009, 21(1-2):133-145.

[6] R. Adams, R. Comley, and M. Ghoreyshi, "The Potential of the BCI for Accessible and Smart e-Learning," Proc. of the 5th Int. Conference on Universal Access in Human-Computer Interaction. Part II: Intelligent and Ubiquitous Interaction Environments (UAHCI '09), 2009, 467-476. DOI: 10.1007/978-3-642-02710-9_51

[7] J. Kemp, O. Després, T. Pebayle, and A. Dufour, "Age-related decrease in sensitivity to electrical stimulation is unrelated to skin conductance: an evoked potentials study," Clinical Neurophysiology, 2014, 125(3):602-7. DOI: 10.1016/j.clinph.2013.08.020

[8] H. Cecotti, "Spelling with non-invasive Brain–Computer Interfaces – Current and future trends", Journal of Physiology-Paris, 2011, 105(1–3), 106-114. DOI: 10.1016/j.jphysparis.2011.08.003

[9] W. Speier, I. Fried, and N. Pouratian, "Improved P300 speller performance using electrocorticography, spectral features, and natural language processing," Clinical Neurophysiology, 2013, 124(7), 1321-1328. DOI: 10.1016/j.clinph.2013.02.002

[10] M. Quek, J. Höhne, R. Murray-Smith, and M. Tangermann, "Designing Future BCIs: Beyond the Bit Rate", in Allison, B., Dunne, S., Leeb, R., Millan, J.D.R. and Nijholt, A. (eds.), Towards Practical Brain-Computer Interfaces: Bridging the Gap from Research to Real-world Applications, Springer, 2013, 173-196. DOI: 10.1007/978-3-642-29746-5_9

[11] J. Johnson, Designing with the Mind in Mind: a Simple Guide to Understanding User Interface Design Rules. Morgan Kaufmann, Burlington, 2011.

[12] S. Mann, "Wearable computing: Toward humanistic intelligence", IEEE Intelligent Systems, 2001, 16(3): 10-15. DOI: 10.1109/5254.940020

[13] J.B. Lopes, "Designing user interfaces for severely handicapped persons," Proc. of the 2001 EC/NSF workshop on Universal accessibility of ubiquitous computing: providing for the elderly (WUAUC'01), ACM, New York, NY, USA, 2001, 100-106. DOI: 10.1145/564526.564553

[14] C. Zickler, V. Kaiser, A. Al-Khodairy, S. Kleih, A. Kübler, M. Malavasi, D. Mattia, S. Mongardi, C. Neuper, M. Rohm, R. Rupp, P. Staiger-Sälzer, and E.-J. Hoogerwerf, "BCI-Applications: Requirements of Disabled End-Users and Professional Users", First TOBI Workshop, Graz, Austria, February 2010.

[15] R. Ortner, R. Prueckl, V. Putz, J. Scharinger, M. Bruckner, A. Schnuerer, and C. Guger, "Accuracy of a P300 Speller for Different Conditions: A Comparison," Proc. of the 5th Int. Brain-Computer Interface Conference, 2011, Graz, Austria, p. 196.

[16] L. A. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials," Electroencephalography and clinical neurophysiology, 70(6), 510-523, Dec 1988. DOI: 10.1016/0013-4694(88)90149-6

[17] H.-J. Hwang, J.-H. Lim, Y.-J. Jung, H. Choi, S.W. Lee, and C.-H. Im, "Development of an SSVEP-based BCI spelling system adopting a QWERTY-style LED keyboard," Journal of Neuroscience Methods, 208(1), 2012, 59-65. DOI: 10.1016/j.jneumeth.2012.04.011

[18] G. Townsend, B. K. LaPallo, C. B. Boulay, D. J. Krusienski, G. E. Frye, C. K. Hauser, N. E. Schwartz, T. M. Vaughan, J. R. Wolpaw, and E. W. Sellers, "A novel p300-based brain-computer interface stimulus presentation paradigm: moving beyond rows and columns," Clinical Neurophysiology, 121(7), 2010, 1109–1120. DOI: 10.1016/j.clinph.2010.01.030

[19] I. Volosyak, H. Cecotti, D. Valbuena, and A. Gräser, "Evaluation of the Bremen SSVEP Based BCI in Real World Conditions", in Proc. of the 11th Int. Conference on Rehabilitation Robotics, Kyoto, Japan, 2009, 322–331. DOI: 10.1155/2012/967305

[20] T. D'Albis, R. Blatt, R. Tedesco, L. Sbattella, and M. Matteucci, "A predictive speller controlled by a brain-computer interface based on motor imagery," ACM Transaction of Computer-Human Interaction, 19(3):20:1–20:25, October 2012. DOI: 10.1145/2362364.2362368

[21] H. Segers, A. Combaz, N.V. Manyakov, N. Chumerin, K. Vanderperren, S. Van Huffel, and M.M. Van Hulle, "Steady State Visual Evoked Potential (SSVEP)-Based Brain Spelling System with Synchronous and Asynchronous Typing Modes," 15th Nordic-Baltic Conference on Biomedical Engineering and Medical Physics (NBC 2011), Aalborg, Denmark, 14–17 June 2011, 164–167. DOI: 10.1371/journal.pone.0073691

[22] M. Treder, and B. Blankertz. "(C)overt attention and visual speller design in an ERP-based brain–computer interface," Behavioral and Brain Functions, 2010, 6, 1–13. DOI: 10.1186/1744-9081-6-28

[23] C. Cheng, J. Yang, Y. Huang, J. Li, and B. Xia, "A Cursor Control Based Chinese-English BCI Speller," Neural Information Processing, Lecture Notes in Computer Science, 8226, 2013, 403-410. DOI: 10.1007/978-3-642-42054-2_50

[24] A. Mora-Cortes, N.V. Manyakov, N. Chumerin and M.M. Van Hulle, "Language Model Applications to Spelling with Brain-Computer Interfaces," Sensors 2014, 14, 5967-5993. DOI 10.3390/s140405967.

[25] J. R. Wolpaw, N. Birbaumer, W. J. Heetderks, D. J. McFaralnd, P. H. Peckham, G. Schalk, E. Donchin, L. A. Quatrano, C. J. Robinson, and T. M. Vaughan, "Brain–computer interface technology: A review of the first international meeting," IEEE Trans. On Rehabilitation Engineering, 2000, 8, 164–173. DOI: 10.1109/tre.2000.847807

[26] B. Rivet, H. Cecotti, M. Perrin, E. Maby, and J. Mattout, "Adaptive training session for a P300 speller brain–computer interface", Journal of Physiology-Paris, 2011, 105(1–3), 123-129. DOI: 10.1016/j.jphysparis.2011.07.013

[27] D. Zhang, H. Song, R. Xu, W. Zhou, Z. Ling, and B. Hong, "Toward a minimally invasive brain-computer interface using a single subdural channel: A visual speller study," NeuroImage 2013, 71:30-41.

[28] Y. Liu, Z. Zhou, and D. Hu,"Comparison of stimulus types in visual P300 speller of brain-computer interfaces," IEEE ICCI 2010:273-279.

[29] Y. Shahriari, and A. Erfanian, "Improving the performance of P300-based brain–computer interface through subspace-based filtering," Neurocomputing, 2013, 121, 434-441. DOI: 10.1016/j.neucom.2010.01.018

[30] G. Pires, U. Nunes, and M. Castelo-Branco, "Comparison of a row-column speller vs. a novel lateral single-character speller: Assessment of BCI for severe motor disabled patients," Clinical Neurophysiology, 2012, 123(6), 1168-1181. DOI: 10.1016/j.clinph.2011.10.04

[31] L, Acqualagna, B, Blankertz, "Gaze-Independent BCI-Spelling Using Rapid Visual Serial Presentation (RSVP)," Clinical Neurophysiology, 124(5):901-908, 2013. DOI:10.1016/j.clinph.2012.12.050

[32] A. Vilic, T.W. Kjaer, C.E. Thomsen, S. Puthusserypady, and H.B. Sorensen, "DTU BCI Speller: An SSVEP-based Spelling System with Dictionary Support," 35th Annual Int. Conf. of the IEEE EMBS, Osaka, Japan, 3 - 7 July, 2013. DOI: 10.1109/EMBC.2013.6609975.

[33] I. Käthner, C.A. Ruf, E. Pasqualotto, C. Braun, N. Birbaumer, and S. Halder, "A portable auditory P300 brain–computer interface with directional cues," Clinical Neurophysiology, 2013, 124(2), 327-338. DOI: 10.1016/j.clinph.2012.08.006

[34] P.T. Wang, C.E. King, A.H. Do, and Z. Nenadic, "Pushing the Communication Speed Limit of a Noninvasive BCI Speller", CoRR abs/1212.0469 (2012)

[35] S.G. Hart, L.E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," in P.A. Hancock & N. Meshkati (Eds.), Human mental workload (pp. 139 – 183). Amsterdam: North-Holland, 1988.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 235–240
DOI: 10.15439/2014F296
ACSIS, Vol. 2

Feature Selection for Classification Incorporating Less Meaningful Attributes in Medical Diagnostics

Agnieszka Wosiak
Lodz University of Technology
Institute of Information Technology
ul. Wólczańska 215, 90-924 Łódź, Poland
Email: agnieszka.wosiak@p.lodz.pl

Danuta Zakrzewska
Lodz University of Technology
Institute of Information Technology
ul. Wólczańska 215, 90-924 Łódź, Poland
Email: danuta.zakrzewska@p.lodz.pl

***Abstract*—In medical diagnostics there is a constant need of searching for new methods of attribute acquiring, but it is difficult to asses if these new features can support the existing ones and can be useful in medical inference. In the paper the methodology of discovering features which are less informative while considering independently, however meaningful for diagnosis making, is investigated. The proposed methodology can contribute to better use of attributes, which have not been considered in the diagnostics process so far. The experimental study, which concerns arterial hypertension as one of the civilization diseases demanding early detection and improved treatment is presented. The experiments confirmed that additional attributes enable obtaining the diagnostic results comparable to the ones received by using the most obvious features.**

I. Introduction

IN MEDICAL research the process of diagnosis is usually provided by experts with the necessary knowledge. Towards facilitating this task some automatic actions may be performed, such as feature selection for choosing the set of attributes appropriate for particular diagnostics problem. In most instances the results of the automated selection cover or even improve the expert judgment [1], [2], [3].

The paper deals with cases where attributes regarded as the most informative should be excluded to discover new dependencies and, as a consequence, new therapies. We consider the methodology, which aims at indicating these features from among less meaningful for medical classification, that can be used in automated diagnosis of the disease. The proposed method includes result evaluation by using clustering taking into account independently features indicated by the method and the recognized ones. The obtained clusters are compared to check if the presented methodology can contribute to better use of attributes acquired from new diagnostics process. Presented approach will be investigated for arterial hypertension, which is acknowledged as one of the civilization diseases demanding early detection and improved treatment [4]. The proposed methodology is evaluated by experiments carried out on real data.

The rest of the paper is organized as follows. In Section II relevant work is presented. Next we describe the proposed methodology and discuss classification and clustering techniques, which are expected to be the most appropriate for the considered case. Section IV is dedicated to the experiments conducted on real data. Finally, in Section V, the results and some concluding remarks are discussed.

II. Related Work

The paper addresses an issue of feature selection methods for medical diagnosis supporting. The new contribution of this work is the exclusion of the most informative features to find out additional dependencies among the attributes derived from a modern process of data acquiring. Such an approach was not considered in the literature so far, however the feature selection analysis was the subject of interests of many researchers.

A survey on feature selection methods was presented in [5]. The main objective was to provide a generic introduction to variable elimination which can be applied to a wide range of machine learning problems. The authors described filter, wrapper and embedded methods. Moreover they applied some of the feature selection techniques on standard data sets to demonstrate the applicability of the proposed methods. As the conclusion they pointed out that comparison between feature selection algorithms can only be done on the same data set since each underlying algorithm behaves differently depending on data characteristic.

The problem of factors that are considered as less important for disease diagnosis but still, according to medical literature, deserve to be included in diagnostics process, was mentioned in [6]. However the main goal of that paper was to compare classifiers for the detection of heart disease. The paper presented effects of using automated feature selection and a medical knowledge based on Motivated Feature Selection (MFS) process. MFS combined with the Computerized Feature Selection (CFS) process was analyzed and good performance was observed for Naive Bayes, k-nearest-neighbors and SMO classifiers.

In [7] the research was based on data of clinical diagnosis, symptoms and medical intervention classification for the patients after surgical intervention with recurrent pelvic cyst. The decision tree was used to find the meaningful characteristic as well as classification rules. The experiment results were to help the clinical faculty doctors in effective diagnosing and providing treatment reference for future patients.

The authors of [8] presented a study of a diverse set of machine learning algorithms on a large number of biomedical

978-83-60810-58-3/$25.00

datasets. They concluded that the nature of a given dataset plays an important role on the classification accuracy of algorithms. Therefore it is necessary to choose an appropriate algorithm for a particular data set. However they identified some general rules for machine learning technique selection: using resampling based classifier enhancement techniques (bagging and boosting) over individual classifiers, using boosting on stable algorithms like SMO, JRip, and J48 and recommended using bagging MLP for classification if the nature of a biomedical data set is unknown.

In [9] the efficiency of the classification methods including SVM, RBF Neural Nets, MLP Neural Nets, Bayesian, Decision Tree and Random Forrest methods were compared. Some of the common clustering techniques including K-means, DBC, and EM algorithms were applied to the datasets and the efficiency of these methods has been analyzed. In each case these methods were applied to eight different binary (two class) microarray datasets. As a conclusion the authors stated that the classification success depends on the choice of feature selection methods, the number of attributes and the number of cases (samples). Results revealed the importance of attribute selection in accurately classifying new samples and the importance of integration of the feature selection and classification algorithms.

The advantages of using and perspectives of applications for AdaBoost algorithm were discussed in [10]. The authors stated that the main significance of AdaBoost concerns providing new ideas to theoretical study as well as practical problems. Moreover while most of the machine learning algorithms tend to seek complicated classifiers to improve the accuracy, AdaBoost takes the approach to obtain an accurate classifier by combining simple and weak classifiers whose accuracies are slightly better than random guessing. Besides AdaBoost does not need any parameters except the number of iterations and therefore the authors suggested that it can be used in many practical applications.

The overview of recent research done to analyze the different machine learning schemes on various medical domains leads to the conclusion that experiments are usually carried out by using the limited choice of algorithms from the machine learning repository. Then the technique which gives relatively the best results for the considered domain is selected. There is no guidelines which indicates the best classifier for a particular type of data.

III. Materials and Methods

The proposed methodology of indicating less meaningful features to use in diagnostics process consists of four steps:

- data preparation, which results in the initial dataset,
- classification process, which enables the selection of the set of attributes crucial for the automated diagnosis,
- clustering based on the attributes derived from the previous step,
- verification process by using expert feature selection.

The research will focus on arterial hypertension case study.

A. Data Description

Arterial hypertension is a significant problem in pediatric practice. Therefore, finding effective methods which support early diagnosis of hypertension is a crucial issue for the researchers to solve. The assessment of arterial hypertension include physical examination, manual and twenty-four hour blood pressure measurements and medical imaging derived from two- and three-dimensional echocardiography.

The tissue Doppler echocardiography (TDE) allows exact evaluation of a number of additional parameters that indicate myocardial functions. Many studies confirmed that regional analysis comes from the method of tissue Doppler imaging as a sensitive way to detect clinically silent changes when standard echocardiographic parameters are still within the normal range [11]. This method is mainly used for research carried out for adult population, however Zamojska et al. considered assigning this approach also for children [12].

The initial cardiac data can be characterized by over 50 attributes. All patients undergo physical examination, manual arterial blood pressure measurements (RRmanSBP, RRmanDBP), ambulatory blood pressure monitoring (ABPM-S, ABPM-D), echocardiographic examination to evaluate cardiac function using standard parameters (ejection fraction - EF, shortening fraction - SF and myocardial performance index - MPI) and tissue Doppler examination (systolic mitral annular velocity profile and regional function parameters: velocity, strain, strain rate).

The aim of medical analysis is to evaluate the characteristics of the variables in the data sets of healthy and diagnosed children and to discover the relationships between all the parameters. The process of diagnosis performed by medical expert is mainly based on the blood pressure measurements (either manual or ambulatory monitored). The rest of the attributes are usually supportive for medical staff as each of them separately cannot indicate the disease and multivariate analysis is difficult to perform without any computer support.

B. Classification Task

For the classification purpose we will consider two approaches: decision trees and adaptive boosting. Decision trees represent one of the main techniques for discriminant analysis in data mining and knowledge discovery [13], [14]. They predict the class membership (dependent variable) of an instance using its measurements of predictor variables. They provide higher classification accuracy and offer an easy way to understand graphic representation of gathered knowledge [15]. Moreover decision trees are easy to understand and analyze, as they reflect a hierarchical way of human decision making. Therefore they are the opposite of the 'black-box' approaches where model parameters are not understandable [16] and can be easily understood by human experts [17].

In the paper we have chosen C4.5 for a decision tree algorithm as one of the most popular. Namely J48 algorithm, which is the open source Java implementation of the C4.5 in the Waikato Environment for Knowledge Analysis (WEKA) data mining tool [18] has been chosen.

1) J48 Algorithm: The J48 algorithm is the WEKA implementation of the C4.5 top-down decision tree learner proposed by Quinlan [19]. The algorithm uses the greedy technique. It deals with numeric attributes by determining where thresholds for decision splits should be placed. J48 algorithm employs an automatic procedure capable to select relevant features from the training data. It is able to cut the poor or non-meaningful branches into an efficient pruning process as well as able to handle both continuous and discrete attributes. In handling continuous attributes, J48 creates a threshold and then splits the list into those attributes, which values are above the threshold and the ones, which are less than or at least equal to the threshold value. It enables handling training data with missing attribute values by employing gain and entropy calculations. Therefore the J48 algorithm may cut the poor and non-meaningful branches into an efficient pruning process [20].

2) AdaBoost Algorithm: The possibility of boosting the prediction quality of a weak learner was firstly introduced by Freund and Schapire [21]. The adaptive boosting algorithm (AdaBoost) solved many practical shortcomings of earlier algorithms [21]. The AdaBoost is a machine learning algorithm which feeds an input training set to a weak learner algorithm repeatedly. During these repeated calls, the algorithm maintains and updates a set of weights, which indicate how difficult it is for the weak learner to identify a particular element of the training data set. Initially, all weights are equal. However, after each call, the weights are updated, in the way, which guarantee that the weights of misclassified training set elements grow. This forces the weak learner to concentrate on the difficult elements of the training set. In the study, we use a decision stump as a weak learner algorithm for the AdaBoost classifier. This model is composed of a single-level decision tree (DT), which uses one of the input parameters [22].

C. Clustering

Cluster analysis algorithms group objects taking into account a certain similarity metric. They divide the objects into a predetermined number of groups in a manner that maximizes a similarity function. During investigations of the proposed methodology, two different approaches, commonly used in medical studies ([9]) will be considered: the Expectation Maximization (EM) probabilistic approach and deterministic k-means algorithm.

1) k-means Algorithm: The k-means algorithm divides a data set into k clusters, where k is a user-defined value. The algorithm starts with k random clusters, and then move objects between those clusters to minimize variability within clusters and maximize variability between clusters. In other words, the similarity rules apply maximally to the members of one cluster and minimally to members belonging to the rest of the clusters. Usually, the means for each cluster on each dimension are calculated for assigning objects into the closest ones [23]. In most of the cases Euclidean metric is considered as the distance function for k-means algorithm [24], [25].

2) EM Algorithm: An expectation-maximization (EM) algorithm finds maximum likelihood estimates of parameters in probabilistic models. EM performs repeatedly between an expectation (E) and maximization (M) steps. Within the E step an expectation of the likelihood of the observed variables is computed and then the M step computes the maximum expected likelihood found on the E step. EM assigns a probability distribution to each instance which indicates the probability of its belonging to each of the clusters [25]. By cross validation, EM can decide how many clusters to create. The goal of EM clustering is to estimate the means and standard deviations for each cluster so as to maximize the likelihood of the observed data. K-means assigns observations to clusters to maximize the distances between clusters. The EM algorithm computes classification probabilities, not actual assignments of observations to clusters.

D. Verification of results

In order to confirm the correctness of the obtained results, clusters based on the most meaningful attributes selected by classification algorithm are built. They are compared with groups created by clustering using attributes indicated by experts. If the groups, which are built taking into account two different sets of attributes, are of similar characteristics, then the attributes indicated by classification can be effectively used in diagnostics process.

Methodology verification consists of the following steps:

- classification using all the available features including most informative ones, which results in the feature subset selection,
- clustering based on the attributes derived from the previous step,
- comparison of clusters obtained after exclusion of most informative features with the clusters from the previous step.

IV. Experimental Analysis and Results

The main objectives of the experiments were to prove, that by performing clustering based on particular set of less meaningful features acquired in automated classification, we can obtain the output results close to data sets acquired by using most important attributes derived from the process of feature selection and pointed out by medical experts. The presented methodology was evaluated on the real data, which were gathered for early diagnosis of arterial hypertension in children. The data set was described earlier in the section III-A.

During experiments 2 initial data sets were considered: the first one (A - Study group), consisted of data of 30 children diagnosed with primary arterial hypertension, without being overweight or obese, hospitalized in the University Hospital No 4, Department of Cardiology and Rheumatology, Medical University of Lodz. The second set (B - Control group) consisted of 30 data of children with normal blood pressure. The decision process of this initial judgment (the value of

the dependent variable for our experiments) was performed by medical experts.

A. Data Preprocessing

As the first step all the cases were put together to form one data set consisted of 60 children. We decided to exclude from the process of automatic classification these attributes that are in the straight relation to the expert judgment: manual arterial blood pressure measurements (RRmanSBP, RRmanDBP) and ambulatory blood pressure monitoring (ABPM-S, ABPM-D). Moreover we removed fundus_oculi as the feature that is usually correlated to arterial hypertension but can not determine this disease. As a result we took into consideration 42 attributes listed in table I, where the first column contains names of all selected parameters, the second one describes these parameters and the third column gives the domain definitions.

B. Classification

According to the methodology described in section III we used two classification methods: decision trees and adaptive boosting.

1) J48 Results.: The J48 algorithm has chosen for classification 9 attributes listed in table II out of all the attributes (table I).

As a result we obtained 58 correctly classified instances (96.67%) and 2 incorrectly (3.33%) which made the precision and recall equal to 0.967, the same for both classes.

2) AdaBoost - Results.: The AdaBoost algorithm choose for classification 6 attributes listed in table III out of all the parameters (table I).

Despite the fact that this method has chosen the set of attributes different from the J48 algorithm, the results were satisfactory enough. We obtained 49 instances correctly classified (81.67%) and 11 incorrectly (18.33%). The weighted average of precision was equal to 0.831 (0.757 for the 1st class and 0.913 for the 2nd class) and the weighted average of recall was equal to 0.817 (0.933 for the 1st class and 0.7 for the 2nd class).

C. Clustering

We performed clustering taking into account the sets of attributes selected by classification algorithms in the previous step of analysis (section IV-B).

1) EM Algorithm with J48 Subset of Attributes: Performing EM algorithm we firstly used the same subset of attributes as it was chosen by J48 algorithm. We obtained 2 clusters automatically by using cross-validation [25]. The first cluster consisted of 21 instances: 19 instances from the set A and 2 cases from the set B. The second cluster included 39 instances: 11 from the set A and 28 from the set B.

2) K-means Algorithm with J48 Subset of Attributes: While testing k-means technique with the same subset of attributes as it was chosen by J48 algorithm the number of 2 clusters was indicated. As a result the first cluster consisted of 35 instances. It contains 8 instances from the set A and 27 ones from the

TABLE I
THE LIST OF PARAMETERS TAKEN FOR ARTERIAL HYPERTENSION CLASSIFICATION

Parameter name	**Parameter description**	**Domain**
Group type	Dependent attribute	Integer
Body mass	Body mass	Real
BMI	Body mass index	Real
BSA	Body surface area	Real
Phys act	Physical activity	Integer
Family hist	Family history risk factor	Integer
EF	Ejection fraction	Integer
SF	Shortening fraction	Integer
IVSs	Interventricular septum-systole	Real
IVSd	Interventricular septum-diastole	Real
PWDs	Posterior wall thickness in systole	Real
PWDd	Posterior wall thickness in diastole	Real
LVDs	Left ventricular systolic diameter	Real
LVDd	Left ventricular diastolic diameter	Real
S long	Longitudinal strain	Real
MPI	Myocardial performance index	Real
LVMPI	Left ventricular myocardial performance index	Real
Sm [cm/s]	Systolic mitral annular velocity at the intraventricular septum level	Real
Sml [cm/s]	Systolic mitral annular velocity profile at the lateral level	Real
LVM Sim	Left ventricular mass by de Simone	Real
LVM Dev	Left ventricular mass by Devereux	Real
V long	Systolic longitudinal regional velocity	Integer
V circ	Systolic circumferential regional velocity	Integer
V rad	Systolic radial regional velocity	Integer
S long	Longitudinal strain	Integer
Time to peek 1	Time to peek for longitudinal strain	Integer
S circ	Circumferential strain	Integer
Time to peek 2	Time to peek for circumferential strain	Integer
S rad	Radial strain	Integer
Time to peek 3	Time to peek for radial strain	Integer
SRI long	Longitudinal strain rate	Integer
SRI rad	Radial strain rate	Integer
SRI circ	Circumferential strain rate	Integer
V long basal	Longitudinal regional systolic velocity - basal segments	Integer
V long mid	Longitudinal regional systolic velocity - middle segments	Integer
V long apex	Longitudinal regional systolic velocity - apical segments	Integer
S long basal	Longitudinal strain - basal segments	Integer
S long mid	Longitudinal strain - middle segments	Integer
S long apex	Longitudinal strain - apical segments	Integer
SRI long basal	Longitudinal strain rate - basal segments	Integer
SRI long mid	Longitudinal strain rate - middle segments	Integer
SRI long apex	Longitudinal strain rate - apical segments	Integer

TABLE II
THE LIST OF PARAMETERS CHOSEN BY J48 ALGORITHM.

Parameter name	Parameter description
Body mass	Body mass
BMI	Body mass index
EF	Ejection fraction
IVSs	Interventricular septum-systole
PWDs	Posterior wall thickness in systole
PWDd	Posterior wall thickness in diastole
Sml	Systolic mitral annular velocity profile at the lateral level
Sm	Systolic mitral annular velocity at the intraventricular septum level
S long mid	Longitudinal strain - middle segments

TABLE III
THE LIST OF PARAMETERS CHOSEN BY ADABOOST ALGORITHM.

Parameter name	Parameter description
EF	Ejection fraction
SF	Shortening fraction
IVSs	Interventricular septum-systole
PWDd	Posterior wall thickness in diastole
Family hist	Family history risk factor
Time to peek 2	Time to peek for circumferential strain

set B. The second cluster included 25 instances: 22 from the set A and 3 from the set B.

3) EM Algorithm with AdaBoost Subset of Attributes: In the third test we executed EM algorithm with the same subset of attributes as it was chosen by AdaBoost. We also obtained 2 clusters automatically by using cross-validation [25]. The first cluster consisted of 21 instances. It was built up of 18 instances from the set A and 3 cases from the set B. The second cluster included 39 instances: 12 from the set A and 27 from the set B.

4) K-means Algorithm with AdaBoost Subset of Attributes: The last run was performed by using k-means technique with the same subset of attributes as it was chosen by AdaBoost algorithm and 2 clusters indicated. Consequently we obtained the first cluster consisted of 24 instances: 18 instances from the set A and 6 cases from the set B. The second cluster included 36 instances: 12 from the set A and 24 from the set B.

The results of all the combinations of methods introduced in section IV-C are presented in table IV, where the first column describes the methods and the last two columns contain the numbers of cases obtained for particular cluster with the reference to the initial data sets A (healthy children) and B (diagnosed children).

It can be easily noticed that in more than 70% of cases group contents were consistent with groups created by the initial expert diagnosis being the result of the most informative attribute analysis.

TABLE IV
THE RESULTS OF CLUSTERING PERFORMED USING PROPOSED METHODOLOGY.

Method	cluster "0"	cluster "1"	% of cases of initial groups
J48-EM	19A / 2B	11A / 28B	78%
J48-k-means	22A / 3B	8A / 27B	82%
AdaBoost-EM	18A / 3B	12A / 27B	75%
AdaBoost-k-means	18A / 6B	12A / 24B	70%

TABLE V
THE LIST OF PARAMETERS CHOSEN BY J48 ALGORITHM OUT FROM ALL THE ATTRIBUTES.

Parameter name	Parameter description
ABPM-S	ambulatory blood pressure monitoring - systolic
ABPM-D	ambulatory blood pressure monitoring - diastolic

D. Verification of results

As the first step in the process of verification we performed classification enabling all the features - also the most informative derived from the standard echocardiography examination. As a result the J48 algorithm pointed to 2 attributes (table V) for the classification task and the AdaBoost algorithm chose 6 attributes (table VI).

After the clustering process we obtained 2 clusters for each combination of methods: EM after J48 classification, k-means after J48, EM after AdaBoost and k-means after AdaBoost. The detailed results are presented in table VII. The first column of the table describes the combination of methods for the classification and clustering. The second and the third columns contain the number of cases in reference to the initial data set of healthy children (A) and data set of diagnosed children (B).

The comparison of results gathered in tables IV and VII allows to conclude that the proposed methodology incorporating less meaningful features produces the cluster structure similar to the clustering based on most informative attributes derived from standard echocardiography. For all the applied algorithms, we obtained more than 60% of cases assigned to the corresponding clusters. The best results were obtained using the combination of J48 classification and EM clustering (80%), and the worst for AdaBoost classification with k-means clustering (65%).

V. CONCLUSIONS AND FUTURE WORK

In medicine, as well as in other fields of science, which include diagnostics techniques, there is a constant need of searching for new methods of attribute acquiring. However it may be difficult to asses if these new features can replace the existing ones and can be useful in medical inference.

In this paper we proposed the methodology of searching for features which are less informative while considering independently, but still meaningful in the process of diagnosis. This approach is mainly useful when new attributes derived from new diagnostics techniques are introduced. These features may

TABLE VI
THE LIST OF PARAMETERS CHOSEN BY ADABOOST ALGORITHM OUT FROM ALL THE ATTRIBUTES.

Parameter name	Parameter description
ABPM-D	ambulatory blood pressure monitoring - diastolic
SF	Shortening fraction
IVSs	Interventricular septum-systole
PWDd	Posterior wall thickness in diastole
SrRRmanSBP	manual arterial blood pressure measurements - systolic
SrRRmanDBP	manual arterial blood pressure measurements - diastolic

TABLE VII
THE RESULTS OF CLUSTERING PERFORMED USING ALL THE FEATURES.

Method	cluster "0"	cluster "1"	% of cases of initial groups
J48-EM	26A / 0B	4A / 30B	93%
J48-k-means	30A / 8B	0A / 22B	87%
AdaBoost-EM	29A / 2B	1A / 28B	95%
AdaBoost-k-means	29A / 7B	1A / 23B	87%

seem to be less meaningful at first and hard to be assessed by medical staff due to multivariate analysis, but the experimental studies confirmed that they enable obtaining the diagnostic results comparable to the ones received by using features recognized as the most informative.

In the first step feature set classification is applied, then taking into account the selected set of attributes clustering is performed. Two different algorithms of classification with two methods of clustering were combined: J48 + k-means, J48 + EM, AdaBoost + k-means, and AdaBoost + EM. During experiments, conducted on real data, we obtain satisfactory results in comparison to the corresponding ones received by cluster analysis carried out by using all the features. Moreover the results did not differ significantly while comparing with the initial groups created by features indicated by experts.

Despite the fact, that the mining methods chosen for the research were widely recommended in the literature as appropriate for medical data, in the future we intend to verify other approaches and build different hybrid solutions to find out methods, which enable discovering new features assuring more precise disease diagnosing.

REFERENCES

[1] I. Guyon and A. Elisseeff, *An Introduction to Variable and Feature Selection*, Mach Learn Res, vol. 3, 2003, pp. 1157-1182

[2] Z. Xu, I. King and M. R.-T. Lyu, *Discriminative Semi-Supervised Feature Selection Via Manifold Regularization*, IEEE Transactions on Neural Networks, Vol. 21, No. 7, 2010, pp. 1033-1047, DOI: 10.1109/TNN.2010.2047114

[3] A. Hamdy and A. E. Hassanien, *The importance of handling multivariate attributes in the identification of heart valve diseases using heart signals*, In: M. Ganzha, L. Maciaszek, M. Paprzycki (eds.) Proceedings of the 2012 Federated Conference on Computer Science and Information Systems, IEEE, 2012, pp. 75-79

[4] L. Ostrowska -Nawarycz and T. Nawarycz, *Prevalence of excessive body weight and high blood pressure in children and adolescents in the city of Łódź*, Kardiol Pol. Vol. 65, 2007, pp. 1079-1087

[5] G. Chandrashekar and F. Sahin F, *A survey on feature selection methods*, Computers and Electrical Engineering, Vol. 40, 2014, pp. 16-28, DOI: dx.doi.org/10.1016/j.compeleceng.2013.11.024

[6] J. Nahar, T. Imama, K.S. Tickle and Y.-P.P. Chen, *Computational intelligence for heart disease diagnosis: A medical knowledge driven approach*, Expert Systems with Applications Vol. 40, 2013, pp. 96-104, DOI: 10.1016/j.eswa.2012.07.032

[7] Y.F. Wang, M.Y. Chang, R.D. Chiang, L.J. Hwang, C.M. Lee and Y.H. Wang, *Mining Medical Data: A Case Study of Endometriosis*, J Med Syst 37:9899, 2013, DOI: 10.1007/s10916-012-9899-y, DOI: 10.1007/s10916-012-9899-y

[8] A.K. Tanwani, M.J. Afridi, M.Z. Shafiq and M. Farooq, *Guidelines to Select Machine Learning Scheme for Classification of Biomedical Datasets*, In: C. Pizzuti, M.D., Ritchie, M., Giacobini (eds.), EvoBIO, Springer, 2009, pp. 128-139, DOI: 10.1007/978-3-642-01184-9_12

[9] M. Pirooznia, J. Yang, M.Q. Yang and Y. Deng, *A comparative study of different machine learning methods on microarray gene expression data*, BMC Genomics, Vol. 9, 2008, DOI:10.1186/1471-2164-9-s1-s13

[10] Y. Cao, Q.-G. Miao, J.-Ch. Liu and L. Gao, *Advance and Prospects of AdaBoost Algorithm*, Acta Automatica Sinica, Vol. 39, 2013, pp. 745-758, DOI: 10.1016/S1874-1029(13)60052-X

[11] S. Yuda, L. Short, R. Leano and T.H. Marwick, *Myocardial abnormalities in hypertensive patients with normal and abnormal left ventricular filling: a study of ultrasound tissue characterization and strain*, Clin Sci, Vol. 103(3), 2002, pp. 283-293

[12] J. Zamojska, K. Niewiadomska-Jarosik, A. Wosiak and J. Stańczyk, *Evaluation of left ventricular systolic function with the use of tissue Doppler echocardiography in children with primary arterial hypertension (Ocena funkcji skurczowej lewej komory z wykorzystaniem metody doplera tkankowego u dzieci z nadciśnieniem tętniczym pierwotnym)*, Pol J Cardiol Vol. 4(2), 2012, pp. 95-100

[13] S.K. Murthy, *Automatic construction of decision trees from data: a multi-disciplinary survey*, Data Mining and Knowledge Discovery, vol.2, 1998, pp. 345-389, DOI: 10.1023/A:1009744630224

[14] L. Rokach and O. Maimon, *Data mining with decision trees: theory and applications*, Machine perception and artificial intelligence, vol. 69, 2008, WorldScientific Publishing (Singapore)

[15] J. Cerquides, M. López-Sánchez, S., Ontañón, E. Puertas, A. Puig, O. Pujol and D. Tost, *Classification Algorithms for Biomedical Volume Datasets*, In: R. Marín, E. Onaindía, A. Bugarín and J. Santos, (eds.) Current Topics in Artificial Intelligence. LNCS, Springer Berlin Heidelberg, vol. 4177, 2006, pp. 143-152, DOI: 10.1007/11881216_16

[16] M. Czajkowski, M. Grześ and M. Kretowski, *Multi-test decision tree and its application to microarray data classification*, Artif Intell Med, 2014, DOI: http://dx.doi.org/10.1016/j.artmed.2014.01.005

[17] C.E. Brodley and P.E. Utgoff, *Multivariate decision trees. Machine Learning*, 1995, pp. 45-77

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, R. Reutemann, I. H. Witten, *The WEKA data mining software: an update*, SIGKDD Explor. Newsl. vol. 11, 2009, pp. 10-18, DOI: 10.1145/1656274.1656278

[19] J.R. Quinlan, *Bagging, Boosting, and C4.5*, In: Thirteenth National Conference on Artificial Intelligence, AAAI Press, 2006, pp. 725-730

[20] S.R. Konda, *A Comparative Evaluation Of Symbolic Learning Methods and Neural Learning Methods*, https://www.cs.umd.edu/grad/scholarlypapers/papers/ShravyaKonda.pdf

[21] Y. Freund and R. E. Schapire, *A decision-theoretic generalization of on-line learning and an application to boosting*, In: Proceedings of the Second European Conference on Computational Learning Theory, Paul M. B. Vitányi (Ed.), Springer-Verlag, London, UK, 1995, pp. 23-37, DOI: 10.1006/jcss.1997.1504

[22] W. Iba and P. Langley, *Induction of one-level decision trees*, In: Ninth International Workshop on Machine Learning, Morgan Kaufmann PublishersInc., SanFrancisco, USA, 1992, pp.233-240, URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.23.2878

[23] J. B. MacQueen, *Some Methods for Classification and Analysis of Multi-Variate Observations*, In: Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1967, pp. 281-297,

[24] V. Ankita, R. V. Satyanarayana and K. Kamalakar, *An Experiment with Distance Measures for Clustering*, In: International Conference on Management of Data, Technical Report, 2008,

[25] I. H. Witten, E. Frank and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques (3rd ed.)*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 241–245

DOI: 10.15439/2014F95
ACSIS, Vol. 2

An Infrastructure for Efficient Reporting Workflow in Grid Based Teleradiology Architectures Using Relation Based Semantic Matching and Integer Linear Programming

Ayhan Ozan Yılmaz and Nazife Baykal
Department of Medical Informatics, Informatics Institute,
Middle East Technical University,
Ankara, Turkey
Email: e124572@metu.edu.tr, baykal@metu.edu.tr

***Abstract*—This paper proposes an infrastructure with a global workflow management algorithm in order to interconnect facilities and reporting units on a single access interface, decrease the access time of medical images and increase the efficiency of the reporting process. The inspection and radiologist attributes extracted by *Grid Agent* are modelled using a hierarchical ontology structure based on Digital Imaging and Communications in Medicine (DICOM) Conformance and DICOM Content Mapping Resource and World Health Organization (WHO) definitions. Attribute preferences rated by radiologists and technical experts or inferred by references are formed into reciprocal matrixes. Weights for entities are calculated utilizing Analytic Hierarchy Process (AHP). The assignment alternatives are processed by relation-based semantic matching (RBSM) and Integer Linear Programming (ILP). The results are evaluated based on turnaround time, workload and report quality and compared with the outcomes obtained by applying Round Robin, Shortest Queue and Random distribution policies.**

I. Introduction

DUE to lack of radiologists within the facilities and consultation needs, the business model for radiology practice around the world is formed to include the facility's employing and outsourcing radiology services to non-local radiology groups [1]. Picture Archiving and Communication System (PACS) and Radiology Information System (RIS) are typically designed to handle local radiology communication and workflow management. However, remote accesses for non-local radiologists that serve several sites need to access medical images with a single interface and return medical reports efficiently, which requires speed, quality and workload optimization. Studies on PACS based on data grids [2], [3] propose co-allocation parallel transfer strategies to improve the non-local access interface and reduce the transfer time for medical images. Integration with heterogeneous resources and systems such as RIS and Hospital Information System (HIS) is also crucial for the quality of the service. This can be achieved by employing agents that support DICOM, Health Level 7 (HL7), Hypertext Transfer Protocol (HTTP), Cross Enterprise Document Sharing (XDS) and non-standardized data at regarding sites [4]. Turnaround time of a requested report for an inspection is affected by the radiologist's availability, reporting speed and workload as well as the image transfer time. Therefore, workflow optimization should also be considered in the network and software architecture design. In previous research, multiple types of workflow optimization and semantic matching strategies are evaluated such as reinforcement learning [5], [6], machine learning (SVM, Bayes) [7] and relation based negotiation [8]. In this study, an infrastructure for medical image distribution is proposed and a RBSM algorithm enhanced by ILP is utilized to design medical image distribution strategy based on reporting workflow and efficiency. Subspecialty and quality of report are also critical parameters for teleradiology service. An inspection requiring subspecialty should be assigned to a radiologist with corresponding experience and high quality reports should be promoted in assignment process. In the proposed algorithm, experiences and subspecialties of radiologists are evaluated based on radiologist characteristics [9], [10] and report quality feedback [11] is included in the ontology map for the recalculation of weights by AHP.

II. Methods

Workflow centric network architecture with an enhanced caching, querying and retrieving mechanism is implemented by seamlessly integrating *Grid Agent* and *Grid Manager* to conventional digital radiology systems. *Grid Agent* is deployed on each site which is responsible for rendering and transferring radiology data with PACS, RIS, and Workstations using DICOM protocol and with *Grid Manager*, clients using DICOM, HL7, HTTP, and Real Time Messaging Protocol (RTMP). *Grid Manager* is responsible for the flow management of images between sites and reporting units or distribution of reports based on the report distribution workflow algorithm. *Grid Manager* also enhances the image access time by providing non-local clients to query and retrieve medical images in parallel from multiple *Grid Agents* where medical content is cached during report distribution process. For web clients,

978-83-60810-58-3/$25.00

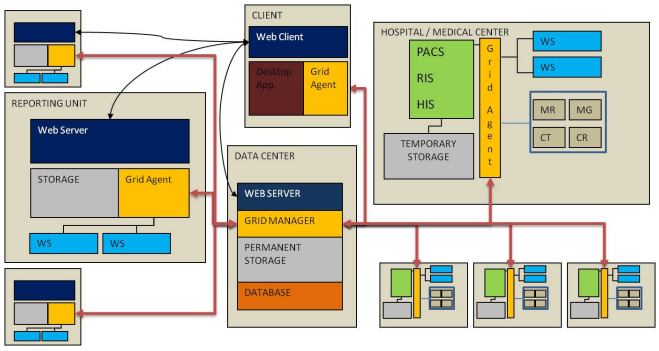


Fig. 1. Workflow centric network architecture with an enhanced caching, querying and retrieving mechanism

Java based open-source DICOM viewer software, *ImageJ*, is customized to stream image instances in parallel chunks.

A. The scenario

A typical data integration, communication and medical image delivery scenario starts with a non-local physician's request for a radiology inspection using the web interface. When the request is received, Grid Manager delivers the imaging request as an Extensible Markup Language (XML) message to the Grid Agent at the regarding medical center. Grid Agent informs the HIS and delivers the Modality Work List (MWL) request to the RIS. When the incoming patient is registered in HIS, Grid Agent is informed which afterwards gets the index of Grid Agents that have patient's data regarding previous examinations from the Grid Manager. Grid Agent pre-fetches the patient's previous medical information and synchronizes PACS and HIS in case a local radiologist examines the inspection. In parallel, Grid Manager automatically assigns the inspection to a remote radiologist after evaluating the experience, report quality, response time and technical adequecy parameters of registered radiologists and corresponding reporting units. Grid Agent at the reporting unit of the assignee receives the updated request list from the Grid Manager and fetches the patient's data including previous examination with the Grid Agent Index and synchronizes the data to PACS in the unit. The non-local radiologist can access the history of the patient independent from the vendor's software and the hospital where the data is acquired. The radiologist at the reporting unit retrieves medical images to be reported from several medical centers on a single interface and generates corresponding reports in RIS or using the web interface. The report is first delivered to the Grid Manager, then to the Grid Agent at the regarding medical center and finally to HIS.

B. The architecture

Teleradiology data serving and sharing architecture is composed of three main components: *Grid Manager* and *Data Center* forming the central node and *Grid Agent* forming the distributed network at the site nodes.

1) The Grid Agent: Grid Agent software is developed to run on an open source media server Red5 which also includes an embedded Tomcat servlet container for JEE Web Applications and supports streaming and shared object communication over RTMP. DICOM and HL7 messages are handled by asynchronous Java threads using dcm4che and HAPI Java libraries. The communication between Grid Agents and Grid Manager is accomplished using encrypted XML messages using HTTP and RTMP protocols.

2) The Grid Manager: Grid Manager is developed to run on Red5 and is specialized to send and receive encrypted XML and SOAP messages or DICOM files utilizing DICOM, HTTP or RTMP protocols. It communicates with Grid Agents and performs database, indexing and file operations at the center. Grid Manager has the Grid Index which includes the patient examination map cached by Grid Agents. The index is in shared object form so that a change in the index is pushed to all agents with the help of RTMP protocol. The caching mechanism at the agents provides the redundancy of the medical data so that the data achieve is distributed and web server maintenance costs are prevented.

3) Data Center Architecture: The central server is composed of the Application, Database and File Operation Layers. Grid Manager forms a bridge between these layers. The Database layer is implemented with open source Postgresql software. The database instances are implemented in shards to deliver large scale loads. Application layer is implemented with Red5 media server and File Operation layer is implemented with Tomcat Servlet Container.

4) The clients: Clients can query and retrieve medical images in parallel from multiple Grid Agents where medical content is cached during the pre-fetching and synchronization processes. For web clients, Java based open-source DICOM viewer software, ImageJ, is customized to stream image instances in parallel chunks. A query that is directed to the grid agent in a hospital by workstations is also directed to other grid proxies. Consequently, the query is performed at every hospital and central web server. Grid Manager provides the Grid Agent and consequently workstations with the result list and the images or data can be retrieved by the help of Grid Manager. Parallel downloading and efficient query algorithms in the Grid Manager enhances the bandwidth usage and time delay.

C. Medical Image Delivery Optimization

1) Problem statement: In order to claim that a radiologist is the optimum choice as a reporter for an inspection, parameters such as experience of the radiologist, response time, workload quota of the radiologist, technical adequacy of the reporting unit that the radiologist is located have to be evaluated.

Experience of the radiologist: Based on the expertise area of the radiologists or experiences on practice, radiologists may be better equipped in certain modalities, diseases or body systems. With reference to the studies that have been carried out on the association between radiologist characteristics and

interpretive performance of diagnostic radiology, a hierarchical structure is defined.

Response time: Response time is another important parameter that should be taken into account while estimating the most suitable reporter for the inspection. A radiology inspection for diagnostic purposed should be reported typically in 48 hours while an urgent inspection should be reported in at most 4 hours. The factors effecting the response time are inspection file delivery time depending on the inspection file size and reporting unit bandwidth, radiologist availability time based on the schedule and radiologist reporting time based on the modality, protocol and statictical data. The statistical data for response time is populated as the radiologists save their reports corresponding to inspections related with a certain set of modality, disease, body part and anatomy.

Workload quota of the radiologist: In order to achieve an efficient reporting process and to balance incomes, each radiologist should be assigned with inspections within certain workload limits. However, every reporting process is not equal in effort. The work load and payments of reporting processes are determined according to the "Performance Point Documentation (SUT)" announced by the Turkish Ministry of Health [13]. "Performance Point Extension Proposal" proposed by the Turkish Society of Radiology is used to strengthen the estimations on the average reporting time. In urgent cases, response time is much more important than the workload quota and expertise area; therefore, expertise area and workload quota are evaluated as secondary importance in emergency situations.

Technical adequacy of the reporting unit: Based on the inspection distribution scenario within this study, it is assumed that the radiologists are located in reporting units, where the assigned inspections are synchronized for access. Therefore, the technical infrastructure of the reporting unit effects the response time and the capacity of reporting service. Bandwidth of the unit affects the response time, while storage capacity and performance of the workstations determine the technical adequacy of the reporting unit. The medical monitor resolution is also taken as a requirement parameter as inspections of certain modalites need high resolutions for investigation.

2) Rendering entities into ontology maps: Ontology maps include the main nodes of Experience, Response Time, Workload and Technical as illustrated in Fig. 3. Experience is evaluated by the assessment of subquantities for each subnode Modality, Body Part, Anatomy and Disease. Similarly each node is connected hierarchically to subnodes having a weighted relation based on AHP. The input for the assessment process is provided by the inspection DICOM file. Each DICOM file provides entites that determine experience, response time, workload and technical requirements as illustrated in Fig. 2. *dcm4che* open-source DICOM Java library is used to render inspections in DICOM format. The modality of inspection, body part and anatomy examined, protocol requested, file size, series and slice numbers, resolution data are rendered into XML for RBSM and ILP processes. It is assumed that the pre-diagnosis is either embedded into the inspection or

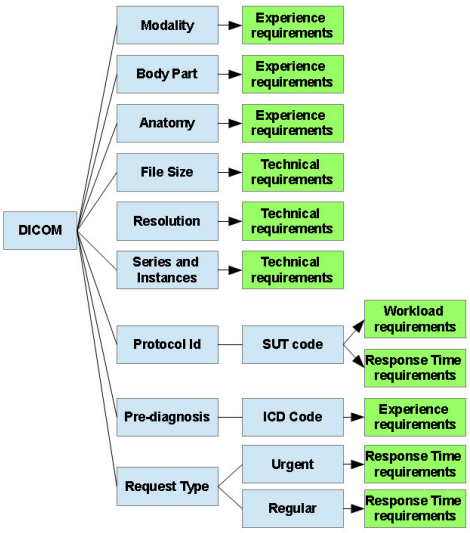


Fig. 2. DICOM file structure. The file structure is rendered to obtain components and these components are used to form the ontology map of the inspection DICOM file.

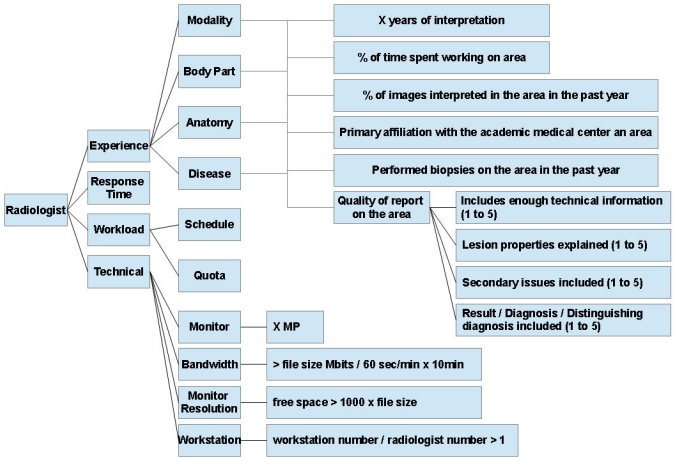


Fig. 3. Ontology map to calculate weights and ratings in inspection assignment to radiologists

entered manually by the report requester as 10^{th} revision of International Classification of Diseases (ICD-10) code. The attributes rendered in the inspection files of DICOM format are used to derive the demand criterias for the radiologist.

3) AHP Process: The pairwise comparison for each entity of the problem is represented by a pairwise comparison matrix.

The weight w_l for each entity l at hierarchical level k of the ontology map including n entities is calculated using pair-wise

comparison matrix element a_{lm} by the following equation:

$$w_l = \frac{\sqrt[n]{\prod_{m=1}^{n} a_{lm}^{(k)}}}{\sum_{l=1}^{n} \sqrt[n]{\prod_{m=1}^{n} a_{lm}^{(k)}}}, \quad l, m = 1, 2, \ldots, n \qquad (1)$$

4) RBSM Process: Ratings are evaluated with a recursive relation based semantic matching process [12] by normalizing the sum of multiplicative weights with the following equation:

$$r_l^{(k-1)} = \frac{\sum_{m=1}^{M} w_m^{(k)} r_m^{(k)} q_m c_{lm}}{\sum_{l=1}^{L} \sum_{m=1}^{M} w_m^{(k)} r_m^{(k)} q_m c_{lm}}, \qquad (2)$$

l=1, 2,...,L, m=1, 2,...,M where L and M are the number of entities at levels $k-1$ and k of the ontology map respectively. q_m is equal to 1 if the qualitative or quantitative condition is satisfied by the potential assignee on entity m and 0 otherwise. c_{lm} is the binary value representing the presence of the connection between entities l and m in the ontology map.

5) ILP Process: The ILP process is formulated by defining the function to be maximized

$$J = \sum_{i=1}^{S} \sum_{j=1}^{R} r_{ij} x_{ij}, \quad i = 1, 2, \ldots, S, \quad j = 1, 2, \ldots R \quad (3)$$

for S inspections and R radiologists where x_{ij} is 1 if inspection i is assigned to radiologist j or 0 otherwise; and r_{ij} represents the resultant rating for each potential assignment calculated by the *RBSM Process*. The constraints are defined as the following:

Assignment constraint: Each inspection is assigned to one and only one radiologist.

$$\sum_{j=1}^{R} x_{ij} = 1 \qquad (4)$$

Workload constraint: The sum of assigned inspection estimated effort e_i should not exceed radiologist's workload l_j.

$$\sum_{i=1}^{S} e_i x_{ij} \leq l_j \qquad (5)$$

Storage constraint: The storage free space s_j in the server at the reporting unit should be at least C_1 times that of the incoming inspection with file size f_i in bytes.

$$\sum_{j=1}^{R} \frac{s_j}{f_i} x_{ij} \leq C_1 \qquad (6)$$

*Bandwidth constraint:*Transfer time of an inspection with file size f_i in bytes to a reporting unit with bandwidth b_j in bits per second should not take more than C_2 seconds.

$$\sum_{j=1}^{R} \frac{8 \times f_i}{b_j} x_{ij} \leq C_2 \qquad (7)$$

Response time constraint: The response time which is equal to the sum of transfer time of the inspection file, radiologist availability time $t_{a,j}$ and radiologist's inspection specific average reporting time $t_{r,ij}$should be less than the requested time $t_{req,i}$. This is typically 4 hours for urgent cases.

$$\sum_{j=1}^{R} \left(\frac{8 \times f_i}{b_j} + t_{a,j} + t_{r,ij} \right) x_{ij} \leq t_{req,i} \qquad (8)$$

ILP process is implemented using the *lp_solve* Java library *OptimJ*.

TABLE I
SIMULATED PRE-DIAGNOSIS ENTRIES AND MODALITY, BODY PART, ANATOMY INFORMATION RENDERED FROM THE DICOM FILES FOR 100 SAMPLE RADIOLOGY INSPECTIONS.

Onthology	**Glossory**
Modality	MR, CT, DX, MG
Body Part	Head, Leg, Throax, Knee, Shoulder
Anatomy	Skull, Leg, Upper Abdomen, Lower Abdomen, Shoulder, Spine
Disease	C39 : Malign neoplasm of inspiration system and inner thorax organs, C50 : Malign neoplasm of breast, C71 : Malign neoplasm of brain, C76 : Malign neoplasm of thorax, C78 : Malign neoplasm of rectum, D43: Neoplasm of brain and central neural system, S02 : Skull and face bone fracture, S42 : Shoulder and fore arm fracture, S82 : Calf and knee fracture, S83 : Dislocation, sprain or strain of knee and ligaments

D. Experimental Design

The performance of the proposed algorithm is tested using 100 sample radiology inspections. A simulation test bed is adopted with 4 imaging facilities and 3 reporting units, 1 data center and 2 non-local clients as virtual machines on different subnets. 6 radiologists working in 3 reporting units are registered and their experience, reporting unit technical capabilities are defined using the web interface. Round robin, random, shortest queue distribution policies are compared to RBSM and ILP distribution algorithms.

III. RESULTS

The results are evaluated based on experience rating, response time success rate and workload average deviation values.

Experience rating is a normalized value between 0 and 1 where higher experience rating is required for better reporting quality.

Response time success rate for policy p is defined as

$$sr_p = \frac{1}{S} \sum_{j=1}^{R} \sum_{i=1}^{S} s_{ij,p} x_{ij,p} \qquad (9)$$

TABLE II
EXPERIENCE RATING, RESPONSE TIME SUCCESS RATE AND WORKLOAD AVERAGE DEVIATION VALUES FOR THE APPLIED DISTRIBUTION POLICIES: ROUND ROBIN, RANDOM, SHORTEST QUEUE, RBSM AND RBSM+ILP.

	Round Robin	**Random**	**Shortest Queue**	**RBSM**	**RBSM +ILP**
Experience Rating	0.34	0.37	0.40	0.75	0.71
Response Time Success Rate	0.74	0.72	1.00	0.93	0.99
Workload Avg. Dev.	2.73	3.60	1.94	0.49	0.28

where $s_{ij,p} = \begin{cases} 1 & t_{ij,rep} \leq t_{i,req} \\ 0 & otherwise \end{cases}$
for reporting time of inspection i by radiologist j, $t_{ij,rep}$ and requested reporting time for inspection i, $t_{i,req}$. The maximum possible value for sr_p is 1 and higher is the response time success rate, better is the distribution policy.

Workload average deviation for policy p, ld_p, is a measure of how efficient the radiologist resources are utilized indicating the distance from the load limit.

$$ld_p = \frac{1}{R} \sum_{j=1}^{R} \frac{\left| \sum_{i=1}^{S} e_i x_{ij,p} - l_j \right|}{l_j} \tag{10}$$

where e_i is the estimated time to report the assigned inspection i and l_j is the workload of radiologist j. The minimum possible value for ld_p is 0 which means the assignment workloads are equal to the defined workload limits for each radiologist. Therefore, distribution policy can be evaluated as more successful in terms of workload efficiency when ld_p approaches 0.

IV. DISCUSSION

The proposed architecture increases the efficiency of reporting process for teleradiology applications and provides a process centric network structure with an enhanced caching, querying and retrieving mechanism.

Shortest Queue policy has the highest response time performance; however it is inefficient in experience rating and workload distribution. Applying only RBSM gives the highest experience ratings, but integrating ILP with RBSM ratings provides a better response time success rate and the best performance for workload distribution with a small optimization trade off in experience rating. RBSM and ILP based image delivery also prevents bandwidth, storage or hardware related locks and latencies.

V. CONCLUSION

The proposed infrastructure decreases the storage costs, reporting costs, turnaround times and increases report quality and effectiveness of resultant treatments. The adaptation of medical sites and reporting groups to the architecture only requires the integration of Grid Agent into the present systems deployed on these sites which decreases integration costs and provides high interoperability.

The response time and report quality statistics for each radiologist are updated in real time. Therefore, it is considered that the proposed solution can be even more efficient and accurate in real case scenarios. Also the recalculation of weights based on the satisfaction level feedback for response time, report quality and workload distribution enhences the algorithm to make more accurate decisions.

REFERENCES

[1] B.F. Branstetter, "Basics of Imaging Informatics: Part 2," *Radiology,* vol. 244, 2007, pp. 78-84, http://dx.doi.org/10.1148/radiol.2441060995.

[2] C.T. Yang, C.H. Chen, M.F. Yang, "Implementation of a medical image file accessing system in co-allocation data grids," *Future Generation Computer Systems,* vol. 26, 2010, pp. 1127-1140, http://dx.doi.org/10.1016/j.future.2010.05.013.

[3] C.T. Yang, M.F. Yang, W.C. Chiang, "Enhancement of anticipative recursively adjusting mechanism for redundant file transfer in data grids," *Journal of Network and Computer Applications,* vol. 32, 2009, pp. 834-845, http://dx.doi.org/10.1109/ICPADS.2008.48.

[4] M. Benjamin, Y. Aradi, R. Shreiber, "From shared data to sharing workflow: Merging PACS and teleradiology," *Europian Journal of Radiology,* vol. 73, 2010, pp. 3-9, http://dx.doi.org/10.1016/j.ejrad.2009.10.014.

[5] Z. Huang, X. Lu., H. Duan, "Mining association rules to support resource allocation in business process management," *Expert Systems with Applications,* vol. 38, 2011, pp. 9483-9490, http://dx.doi.org/10.1016/j.eswa.2011.01.146.

[6] Z. Huang, W.M.P. Aalst, X. Lu, H. Duan, "Reinforcement learning based resource allocation in business process management," *Data and Knowledge Engineering,* vol. 70, 2011, pp. 127-145, http://dx.doi.org/10.1016/j.datak.2010.09.002.

[7] Y. Liu, J. Wang, Y. Yang, J. Sun, "A semi-automatic approach for workflow staff assignment," *Computers in Industry,* vol. 59, 2008, pp. 463-476, http://dx.doi.org/10.1016/j.compind.2007.12.002.

[8] W.K. Cheng, B.Y. Ooi, H.Y. Chan, "Resource federation in grid using automated intelligent agent negotiation," *Future Generation Computer Systems,* vol. 26, 2010, pp. 1116-1126, http://dx.doi.org/10.1016/j.future.2010.05.012.

[9] D.L.Miglioretti,et al, "Radiologist Characteristics Associated With Interpretive Performance of Diagnostic Mammography," *Journal of National Cancer Inst,* vol. (99)24, 2007, pp. 1854-1863, http://dx.doi.org/10.1093/jnci/djm238.

[10] M.Eduard, et al, "Association between Radiologists' Experience and Accuracy in Interpreting Screening Mammograms," *BMC Health Services Res.,* vol. (8)91, 2008, pp. 1-10, http://dx.doi.org/10.1186/1472-6963-8-91.

[11] J. Hohmann, et al, "Quality assessment of out sourced after-hours computed tomography reports in Central London University Hospital," *Europian Journal of Radiology,* vol. 81, 2012, pp. e875-e879, http://dx.doi.org/10.1016/j.ejrad.2012.04.013.

[12] S. Colucci, et al, "A Formal Approach to Ontology-Based Semantic Match of Skill Descriptions," *Journal of Universal Computer Science,* vol. (9)12, 2003, pp. 1437-1454, http://dx.doi.org/10.3217/jucs-009-12-1437.

[13] Performance Point Documentation (SUT), Turkish Ministry of Health, http://www.istanbulsaglik.gov.tr/w/sb/imis/pdf/ek_8.pdf.

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 247–251

DOI: 10.15439/2014F199
ACSIS, Vol. 2

A new approach to automatic continuous artery diameter measurement

Bartosz Zieliński and Adam Roman
Institute of Computer Science
and Computer Mathematics,
Faculty of Mathematics and Computer Science,
Jagiellonian University,
ul. Łojasiewicza 6, 30-348 Kraków, Poland
Email: {bartosz.zielinski, adam.roman}@uj.edu.pl

Agata Dróżdż, Agata Kowalewska,
and Marzena Frołow
Jagiellonian Centre for Experimental Therapeutics,
Jagiellonian University,
ul. Bobrzyńskiego 14, 30-348 Kraków, Poland
Email: {agata.drozdz, agata.kowalewska,
marzena.frolow}@jcet.eu

***Abstract*—In this paper, we present an application which aid an evaluation of the arterial diameter changes, based on ultrasound videos. The designed, implemented and verified algorithm uses the techniques of image processing, image analysis and pattern recognition, such as filtering, profile plot analysis and active contour method. Except determining the artery diameter over time it is also able to retrieve ECG from ultrasound video. The results obtained for both signals are synchronized, therefore it is possible to obtain the artery diameters in R wave points, which is a novel approach. Experiments were performed to assess the software validation by comparing the outcomes obtained with the evaluated algorithm with those manually-acquired – the correlation is high. This is the first stage of the research in which we will build the cardiovascular predictive model to search for the new cardiovascular factors.**

I. Introduction

INCREASED risk of cardiovascular diseases is related to endothelial dysfunction, which can be assessed based on the ultrasonic monitoring of brachial artery diameter changes in response to the temporal closure of the vessel. During the examination, videos containing the brachial artery representation in B-mode and M-mode are simultaneously captured (left and right side in Fig. 1). B-mode is a 2D image of the brachial artery in the longitudinal plane. A single scan line placed along B-mode is used by ultrasound device to generate M-mode representation, which describes how the structures intersect with that line in time.

Manual measurement performed on such videos is operator-dependent, prone to mistakes and time-consuming. Moreover, it is impossible to perform continuous measurement manually, due to frames quantity (30 frames per second).

In this paper, we present an application which can aid an evaluation of the arterial diameter changes, based on ultrasound videos. The designed, implemented and verified algorithm uses the techniques of image processing, image analysis and pattern recognition, such as profile plot analysis and active contour method [9]. Except determining the artery diameter over time it is also able to retrieve ECG from an M-mode ultrasound. The results obtained for both signals are synchronized, therefore it is possible to obtain the artery diameters in R wave points, which is a novel approach introduced in our algorithm.

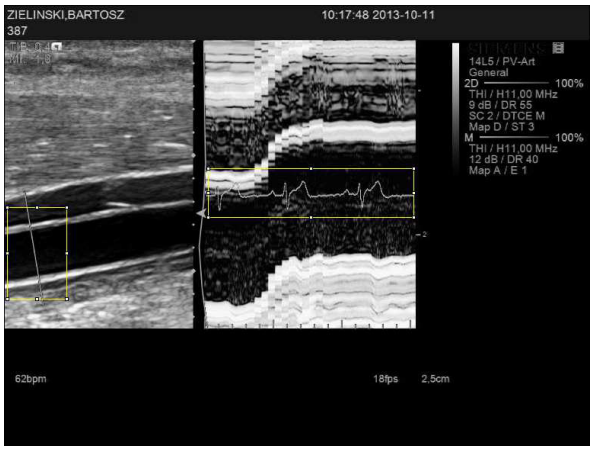


Fig. 1. The example of the first frame from the ultrasound video with B-mode and M-mode on the left and right side, respectively.

The test set includes baseline, occlusion and post-deflation videos and images obtained from 20 patients, together with the automatically and manually-measured values. Such test set was used to conduct analyses, results of which were investigated.

II. The medical aspects of the paper

The endothelium, forming the innermost, unicellular layer of arteries, plays a key role in the homeostasis of blood vessels. Properly functioning endothelial cells, in response to shear stress (SS) variations produce certain particles that may cause dilation or constriction of the vessel. Increased SS results in releasing vasodilators such as Endothelium-Derived Relaxing Faxtor (Nitric Oxide), Endothelium-Derived Hyperpolarizating Factor and prostacyclin, causing a local increase in the vessel diameter. Disorders in the excretion of these substances can be associated with the early endothelial dysfunction, which among others is related to increased likelihood of cardiovascular diseases [2], [7].

One of the most commonly used method for non-invasive assessment of endothelial function is the flow mediated-

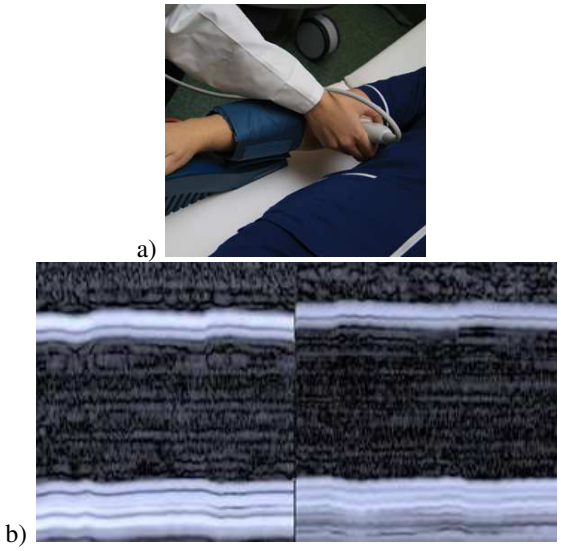


Fig. 2. Cuff and ultrasonic probe placement – a. M-modes of the arterial baseline (left side) and its dilation after releasing the cuff (right side) – b.

dilation (FMD) study. It is based on the ultrasonic monitoring of the brachial artery diameter changes in response to the temporal closure of the vessel (ischemia). Artery closure is achieved through the pressure cuff inflation on the limb (see Fig. 2a). After releasing the cuff the shear stress acting on the endothelial cells in the vessel wall is augmented as a result of increased flow, causing a reactive hyperemia [3]. The arterial diameter varies between 2 and 5 mm and its change ranges from 0 to 25% (see Fig. 2b). The accurate calculation of such a small variation would be difficult due to the limitations of the standard USG, therefore artery imaging is performed using a high frequency ultrasound probe (7-14 MHz).

III. The existing methods

Most of the FMD measurements in the literature are done by tracing the vessel boundary manually. Such a process is time consuming and error-prone. Therefore, several authors proposed some automated methods for this task. In [5] the authors proposed a robust automated measurement of the vasodilator response by automatically locating the artery using a variable window method and global constraint deformable model for vessel wall boundary detection. Geminiani et al. [6] used a robust edge detection algorithm, called "mass center of the gray level variability". Faita et al. [4] introduced the localization algorithm of the artery tunics based on a new mathematical operator called the first order absolute moment and on a pattern recognition approach. Bartoli et al. [1] utilized a spline model (deformable template) that detects the artery boundaries and track them all along the video sequence. The a priori knowledge about the image features and its content is exploited in this approach. Kaneko et al. [8] developed a method of measuring the change in the thickness and the elasticity of the brachial artery during a cardiac cycle using the so-called phased tracking method for the evaluation of the mechanical property of only the intima media region.

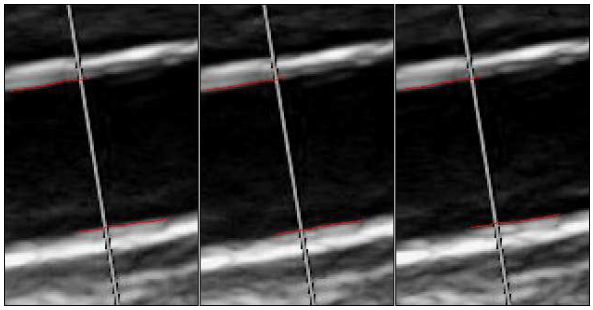


Fig. 3. The successive steps of the artery diameter analysis.

Our approach is based on profile plot analysis and active contour method. As opposed to the existing models, our algorithm, except determining the artery diameter over time, also retrieves ECG from M-mode and is resistant to the additional line in B-mode. The existing commercial medical applications for FMD measurement, such as Quipu FMD Studio[1], or MIA Vascular Tools 5[2] do not return the partial results like continues artery diameter. On the other hand, such partial product is necessary to conduct medical research. Other algorithms are sold only with the new machinery, like Ultrasonix products[3]. Our approach does not depend on any machinery.

IV. The algorithm

The presented algorithm can be employed for artery diameter measurement and FMD study. It is a semi-automatic technique, which operates on the basis of two regions of interest (ROI) selected by the operator in the first frame of the ultrasound video (see Fig. 1). First ROI contains a fragment of B-mode with scan line (left yellow rectangle in Fig. 1). The ECG representation is selected as the second ROI (right yellow rectangle in Fig. 1).

The first ROI is given together with the initial outlines of the arterial walls, which are automatically tracked at the following frames of the video (see Fig. 3). The second region of interest is analysed fully automatically.

A. Determining the artery diameter over time

The algorithm for determining the artery diameter over time works for both outlines independently and is based on the active contour method [9]. In both cases, it starts with the outline containing 2 points (beginning and end of the section given by the system operator), which is decomposed to obtain series of 10 points with the same distance between successive elements.

Representing the position of a snake parametrically by v(s), we can write its energy functional as [9]:

$$E^*_{snake}(v) = \int E_{int}(v(s)) + E_{image}(v(s)) + E_{con}(v(s))dx. \quad (1)$$

[1]http://www.quipu.eu/fmd.php
[2]http://www.mia-llc.com/products/vascular_fda.htm
[3]http://www.ultrasonix.com/research/clinical

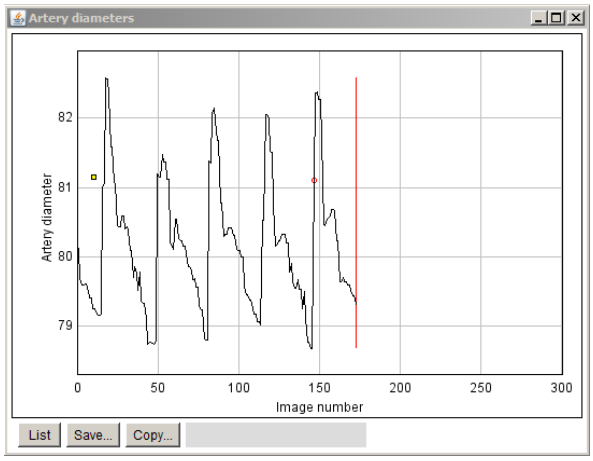


Fig. 4. The partial result of the artery diameter analysis.

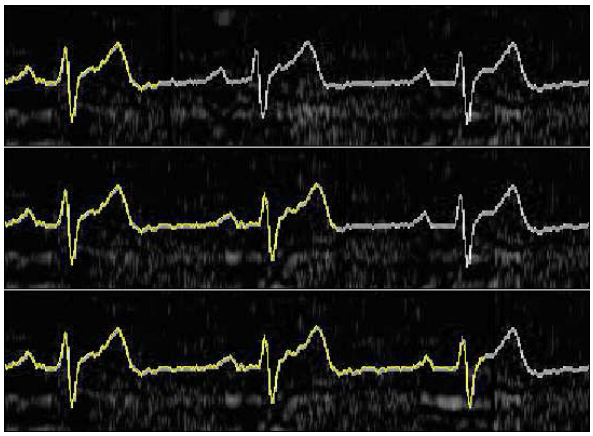


Fig. 5. The successive steps of the ECG retrieval analysis.

Internal spline energy contains a first-order term controlled by α, and a second-order term controlled by β:

$$E_{int}(v(s)) = \frac{\alpha|v_s(s)|^2 + \beta|v_{ss}(s)|^2}{2}, \quad (2)$$

The first-order term is responsible for stretching and the second-order is responsible for bending. The values of both factors were initially set to 0.5, to prevent extensive stretching and bending. Those preliminary values proved to be accurate (see Section VI) and therefore were not explored more specifically, but we are plan to do it in the future.

The total image energy can be exposed as a weighted combination of the three energy functionals:

$$E_{image}(v(s)) = w_{line}E_{line} + w_{edge}E_{edge} + w_{term}E_{term}, \quad (3)$$

$E_{line} = I$ is the image intensity itself, $E_{edge} = -|\nabla I|^2$, and E_{term} is terminal functional utilized in order to find terminations of line segments and corners. The value of w_{edge} was set to 1 to promote the edges, while both other factors were set to 0, not to promote darker or lighter regions. Moreover, the Gaussian blur is applied to an image before computing ∇I and then it is scaled to $[0, 1]$. With reference to such approach, the factor γ responsible for step size was set to 0.

Functional E_{conv} responsible for external constrain forces was currently omitted, but we plan to use it in the feature to promote parallelism between both outlines.

The endpoints (P_1 and P_{10}) are treated differently than the other points. Their x coordinate left unchanged, while their y coordinates are obtained using linear approximation computed for the outline. It is necessary, because otherwise the outlines would shrink into two points. In the last step of the iteration, each outline is modified so that the distance between the successive points is the same and successive iteration begins.

Active contour method works as long as distance between the outlines from the successive iterations is bigger than $0.000001mm$. As the result of such analysis for each frame, the sequence of the artery diameters is obtained (see Fig. 4).

B. ECG retrieval from M-mode

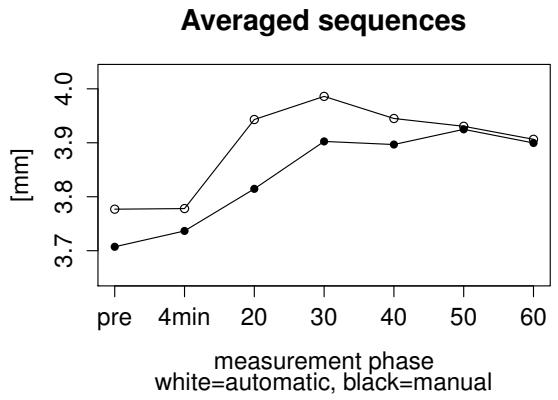
The algorithm for ECG retrieval is based on M-mode which is cyclic (see Fig. 1). Therefore, in order to retrieve ECG it finds the most differing columns of two successive video frames. For each of them the profile plot is generated and then analysed to determine the pixels with the maximal values. Thanks to such approach, the analysis is invariant to noise and equipment type (see Fig. 5).

V. Experiments

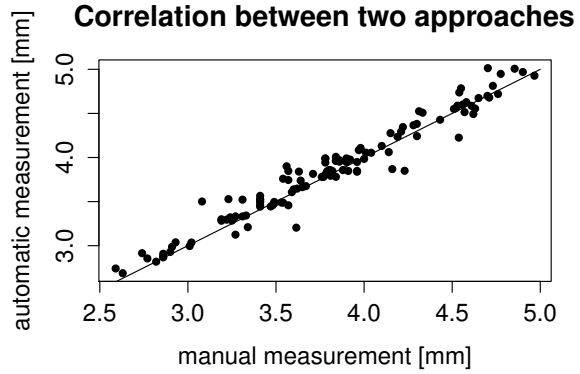
The software validation was performed by comparing the outcomes obtained with the evaluated algorithm with those manually-acquired. The test procedure is as follows. Before the test a pressure cuff is placed on the patientâĂńs forearm and the 14-MHz ultrasonic transducer is placed on the upper arm (proximal to the cuff), fixed on a tripod. During the whole test videos containing the brachial artery representation in B-mode and M-mode are simultaneously captured (Fig. 1). A baseline video is recorded for 10 seconds. Thereafter arterial occlusion is created by inflating the cuff for 5 minutes to about 50 mmHg above patientâĂńs systolic pressure. The last, 60-second video is recorded continuously from 5th second after the cuff deflation.

Hence, there are 7 points in time where the measurements are taken: baseline (called "pre" in the following graphs), arterial occlusion ("4min") and 5 points in time after cuff deflation (in 20th, 30th, 40th, 50th and 60th second). The software tracks the wall borders using B-mode video representation and it displays arterial diameter changing in time as a result. After a calibration it is possible to determine vesselâĂńs peak value. The manual diameter measurements were performed using M-mode images captured every 10 seconds after the cuff deflation, starting from the fifth second. The diameter was measured in each image at R wave points (an average from three measurements).

The data come from the tests (both manual and automatic) done for 20 patients. Fig. 6a shows the averaged sequence of measurements for both approaches. The results for the baseline, arterial occlusion and 30s are almost identical. Both



a)



b)

Fig. 6. Averaged manually and automatically obtained sequences – a. Correlation between manual and automatic measurements – b.

approaches give similar results also for 40, 50 and 60s. For 20s point we can see the larger difference between manual and automatic approach. This is due to the problems with manual measurements mentioned earlier – they are operator-dependent and more prone to mistakes. There is also a threat of missing the right value due to 10 second intervals between measurements. In the initial phase the ultrasonic transducer can be still tuned to capture the good quality image of the artery. As one can see from Fig. 6a, automatic method works better, because the pressure in the 20th second after cuff release should be much higher than it follows from the manual measurement. The automatic measurements follow the right scheme here.

Correlation between manual and automatic measurements is shown in Fig. 6b. Each point in this figure represents a measurement from one phase (baseline, cuff occlusion, 20s, 30s, 40s, 50s or 60s) for one patient. The correlation is 0.98. This means that in general the automatic method is well aligned with the manual one, so we can claim that the method is accurate. The points lying farther from the $y = x$ line refer mainly to the situation from 20th second presented in Fig. 6a, where the operator can make an inaccurate measurement (this will be one of the topic of the future research). The mean difference between manual and automatic measurement was $-0.055mm$. The mean differences split by phase is shown in Table I.

TABLE I
DIFFERENCE BETWEEN MANUAL AND AUTOMATIC MEASUREMENTS SPLIT BY PHASE (IN MILLIMETRES).

phase	pre	4min	20	30	40	50	60
diff	-0.13	-0.09	-0.19	-0.15	-0.11	-0.07	-0.006
stddev	0.56	0.26	0.34	0.34	0.27	0.19	0.08

The automatic measurements are slightly larger than the manual ones. This is due to the fact that the operator might use her own method of marking the diameter, with endpoints lying closer to each other than in the automatic case. Hence, in case of the manual measurement there may be a systematic error. However, this doesn't matter, as the main metric used in the subsequent analysis is the FMD, defined by the formula $FMD = \frac{pdh-b}{b} \cdot 100\%$, where pdh is a peak diameter in hyperemia and b is a baseline value. The FMD is a ratio-type metric, so systematic error doesn't affect the differences between automatic and manual measurements given in terms of FMD.

VI. Concluding remarks

In this paper we presented an application for monitoring the changes in arterial diameter, based on ultrasound videos. It was validated by comparing its outcomes with those acquired manually. The tool turned out to be very accurate and outperforms manual measurement process. Except determining the artery diameter over time it is also able to retrieve ECG from an M-mode ultrasound. The results obtained for both signals are synchronized, therefore it is possible to obtain the artery diameters in R wave points, which is a novel approach introduced in our algorithm (see Fig. 7).

This is the first stage of the research in which we will build the cardiovascular predictive model to search for the new cardiovascular factors. Therefore, the following will be introduced during the further part of the work. First, both regions of interest shown in Fig. 1 will be selected automatically, together with the initial outlines of the arterial walls (see Fig. 3). Second, we will quantitatively describe the artery diameter over time in terms of its characteristic components, such as the slope of its initial part and length of the plateau. As the result, the software will be used to evaluate a distensibility of arterial wall (e.g. carotid).

References

[1] Bartoli G. et al. 2008. Model-based analysis of flow-mediated dilation and intima-media thickness. J Biomed Imaging 16, http://dx.doi.org/10.1155/2008/738545
[2] Celermajer D.S. 1997. Endothelial dysfunction: does it matter? Is it reversible? J Am Coll Cardiol 30:325–333, http://dx.doi.org/10.1016/S0735-1097(97)00189-7
[3] Dick H.J. et al. 2011. Assessment of flow-mediated dilation in humans: a methodological and physiological guideline. Am J Physiol-Heart Circ Physiol 300:H2–H12, http://dx.doi.org/10.1152/ajpheart.00471.2010
[4] Faita F. et al. 2008. Detection of artery interfaces: a real-time system and its clinical applications. Med Imaging 69200F–69200F, http://dx.doi.org/10.1117/12.770408

a)

b)

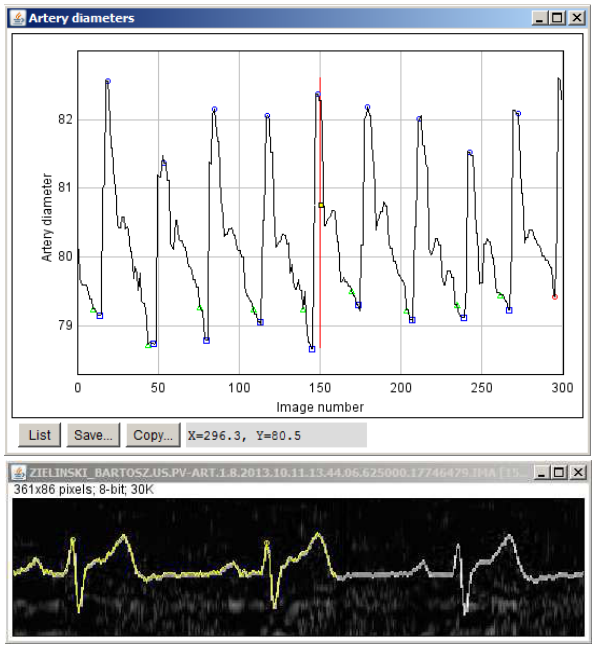


Fig. 7. The result of the artery diameter and ECG analysis – a and b, respectively. The R wave points are marked as green triangles in a and yellow circles in b. Blue circles and squares in a are local maximums and minimums of the artery diameter.

[5] Fan L. et al. 2000. Ultrasound measurement of brachial flow-mediated vasodilator response. IEEE Trans Med Imaging 19:621–631, http://dx.doi.org/10.1109/42.870669

[6] Gemignani V. et al. 2007. A system for real-time measurement of the brachial artery diameter in B-mode ultrasound images. IEEE Trans Med Imaging 26:393–404, http://dx.doi.org/10.1109/TMI.2006.891477

[7] Jazuli F., Pyke K.E.. 2011. The impact of baseline artery diameter on flow-mediated vasodilation: a comparison of brachial and radial artery responses to matched levels of shear stress. Am J Physiol-Heart Circ Physiol 301:H1667–H1677, http://dx.doi.org/10.1152/ajpheart.00487.2011

[8] Kaneko T. et al. 2007. Ultrasonic measurement of change in elasticity due to endothelium dependent relaxation response by accurate detection of artery-wall boundary. Jpn J Appl Phys 46:4881, http://dx.doi.org/10.1143/JJAP.46.4881

[9] Kass M. et al. 1988. Snakes: Active contour models. Int J Comput Vision 1:321–331

4th International Workshop on Advances in Semantic Information Retrieval

RECENT advances in semantic technologies form a solid basis for a variety of methods and instruments that support multimedia information retrieval, knowledge representation, discovery and analysis. They influence the way and form of representing documents in the memory of computers, approaches to analyze documents, techniques to mine and retrieve knowledge. The abundance of video, voice and speech data also raises new challenging problems to multimedia information retrieval systems.

We believe that our workshop will facilitate discussions of new research results in this area, and will serve as a meeting place for researchers from all over the world. Our aim is to create an atmosphere of friendship and cooperation for everyone, interested in computational linguistics and semantic information retrieval. The ASIR'14 workshop will continue to maintain high standards of quality and organization, set in the previous years. We welcome all the researchers, interested in semantic information retrieval, to join our event.

Topics

The workshop addresses semantic information retrieval theory and important matters, related to practical Web tools. The topics and areas include but not limited to:

- Domain-specific semantic applications.
- Evaluation methodologies for semantic search and retrieval.
- Models for document representation.
- Natural language semantic processing.
- Ontology for semantic information retrieval.
- Ontology alignment, mapping and merging.
- Query interfaces.
- Searching and ranking.
- Semantic multimedia retrieval.
- Visualization of retrieved results.

Event Chairs

Klyuev, Vitaly, University of Aizu, Japan
Mozgovoy, Maxim, University of Aizu, Japan

Program Committee

Borgo, Stefano, Laboratory for Applied Ontology, ISTC CNR, Italy

Budzyńska, Katarzyna, Institute of Philosophy and Sociology of the Polish Academy of Sciences, Poland

Carrara, Massimiliano, Universita di Padova, Italy

Cybulka, Jolanta, Poznan University of Technology, Poland

Dobrynin, Vladimir, Saint Petersburg State University, Russia

Goczyła, Krzysztof, Gdansk University of Technology, Poland

Haralambous, Yannis, Institut Telecom - Telecom Bretagne, France

Homenda, Wladyslaw, Warsaw University of Technology, Poland

Jin, Qun, Waseda University, Japan

Kaczmarek, Janusz, Lódz University, Poland

Kakkonen, Tuomo, University of Eastern Finland, Finland

Krawczyk, Bartosz, Wroclaw University of Technology, Poland

Kulicki, Piotr, John Paul II Catholic University of Lublin, Poland

Lai, Cristian, CRS4, Italy

Leonelli, Sabina, University of Exeter, United Kingdom

Ludwig, Simone, North Dakota State University, United States

Martinek, Jacek, Poznan University of Technology, Poland

Mirenkov, Nikolay, University of Aizu, Japan

Mozgovoy, Maxim, University of Aizu, Japan

Nalepa, Grzegorz J., AGH University of Science and Technology, Poland

Palma, Raúl, Poznan Supercomputing and Networking Center, Poland

Piasecki, Maciej, Wroclaw University of Technology, Poland

Pyshkin, Evgeny, St. Petersburg State Polytechnical University, Russia

Reformat, Marek, University of Alberta, Canada

Shtykh, Roman, CyberAgent Inc., Japan

Slezak, Dominik, University of Warsaw & Infobright Inc., Poland

Soldatova, Larisa, Brunel University, United Kingdom

Suárez-Figueroa, Mari Carmen, Ontology Engineering Group, Scool of Computer Science at Universidad Politécnica de Madrid, Spain

Tadeusiewicz, Ryszard, AGH University of Science and Technology, Poland

Vacura, Miroslav, University of Economics, Czech Republic

Vazhenin, Alexander, University of Aizu, Japan

Wang, Haofen, Shanghai Jiao Tong University, China

Wu, Shih-Hung, Chaoyang University of Technology, Taiwan

Zadrozny, Slawomir, Systems Research Institute, Poland

Ławrynowicz, Agnieszka, Poznan University of Technology, Poland

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 255–259
DOI: 10.15439/2014F343
ACSIS, Vol. 2

Semantic sentence structure search engine

Nikita Gerasimov
Nothern (Arctic) Federal University,
Severnaya Dvina Emb. 17,
Arkhangelsk, Russia; 163002;
Email: n.gerasimov@narfu.ru

Maxim Mozgovoy
The University of Aizu, Tsuruga,
Ikki-machi, Aizu-Wakamatsu,
Fukushima, 965-8580 Japan
Email: mozgovoy@u-aizu.ac.jp

Alexey Lagunov
Nothern (Arctic) Federal University,
Severnaya Dvina Emb. 17,
Arkhangelsk, Russia; 163002;
Email: a.lagunov@narfu.ru

***Abstract*—Many of current web search engines rely on inverted index-based data structures as document information store. Since and inverted index is a map from individual document words to their respective locations, such data structure destructs semantic links between the words, and thus does not support structural user queries. In other words, such systems can only find the documents that contain user-specified words. In this paper we propose to create semantic links between the terms contained in inverted index, and in such way create a semantic network. This network will preserve the internal structure of the stored documents, and will enable the users to perform structural queries. Both structural-saving indexation and structural user search query allow to save semantic speech meaning of the text while search process.**

I. Introduction

Today all popular search engines operate with inverted index [1],[2],[3], which provides the grounds for high-quality keyword-based search. The main idea is to create a mapping from every token to a list of its positions in the documents, indexed by the search engine. While both page ranking and linguistic algorithms can offer rather acceptable results for the users, the very idea of processing non-linked keywords, extracted from texts, imply non-semantic search only.

Thus, today's semantic networks, implemented both by commercial companies and and open communities are not fully utilized by the search engines. Powerful linguistic and statistic functions, implemented in modern search engines, are not used to their full extent.

A. Preserving semantic links

The main idea of the present work is to store not inverted index, but sentence structure with link to it's source page position. The base sentence structure consists of three elements: predicate, subject and object, called a triplet. This idea is presented in the Figure 1. Each page is parsed to get linked tokens, constituting the elements to be saved to the database with sentence links and source page positions. The tokens form an oriented graph or a semantic network. Subjects in such graph serve as objects for other subjects and vice versa. This structure is similar to RDF [5] (Resource Description Framework), which describes knowledge using a directed graph.

Search process is implemented with RDF queries over the semantic network. The user enters a triplet in form of three words, which is searched in database of linked documents (a semantic network). In the future, the user will be able to use natural language as a query language. In this case, the system will be able to process not only triplet words, but also other syntactic forms. This means that the indexer will have to process the source documents using an extended RDF scheme, which would contain also adjectives, adverbs, and other parts of speech (POS).

A query triplet can be searched in the database using a simple straightforward comparison or with the methods used in many popular search engines, such as synonyms dictionary and TF-IDF[1][4].

TF-IDF can be implemented as a coefficient of relevance, which influences the document position in the resulting list.

Our search algorithm is not intended to replace traditional inverted index search engines, and can be implemented within an additional module, or serve as a basis for a specialized fact search engine in a knowledge graph.

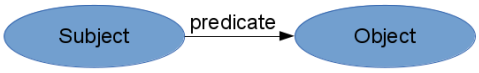


Fig 1: RDF

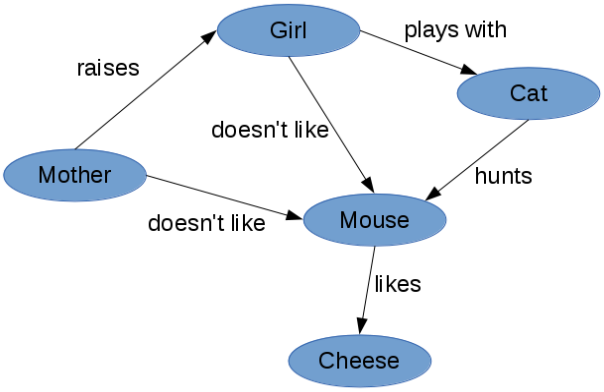


Fig 2: Semantic network

[1]TF-IDF is numerical statistic dimension defining document relevancy in document collection. Result depends on word frequency in current document and inverse frequency in other documents.

B. Basic search engine functions

Our search engine indexing mechanism (robot or spider) solves the following problems:

1. Detecting document external links
2. Useful content detection
3. Semantic structure parsing

C. Useful content detection

For useful content detection we used an artificial neural network, as suggested in [6]. Many of web-pages sources are divided into separate strings containing HTML markup and text, despite it doesn't influence to page rendering. Also we empirically devided HTML tags into two groups: simple and special. Simple tags collection contains text decorating tags like "<b><i><s>". Special tags set contains the others one. Our neural network detects whether a given string contains meaningful text or non-meaningful webpage elements.

While exploring HTML documents we found such regularities:

1. Useful content usually is absent at the beginning and ending of the article.
2. A string is probably useful if the presence of HTML tags inside the string is low.
3. Longer strings are most probably useful.

We used neural network with thr following input parameters:

1. Document string number expressed in percents.
2. A string length expressed in percents. 100% is the longest document string.
3. Relation between simple HTML tags and text chars.
4. Relation between special HTML tags and text chars.

Every parameter except the first one is repeated two more times: for the previous string, and for the next one. The characteristics of our neural networks are shown in Table I.

We trained the neural network using 50 English Wikipedia pages. This method allowed us to quickly get a content parser, having 83% decision accuracy. As a neural network engine we used Encog Java library.

D. Database

Our system uses two DBMS: a NoSQL graph-oriented Neo4J DBMS, and a NoSQL document-based MongoDB. In order to store tokens with minimal overhead, we employ tries (see Figure 3).

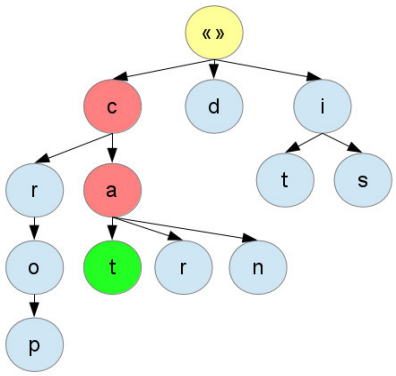


Fig 3: Trie

Tries are supported by Neo4J DBMS that stores all data as a graph, and provides handy ways to traverse graphs, and search and retrieve individual vertices.

To store RDF-like links we used a NoSQL document-based DBMS MongoDB to achieve structureless storage organization, and high speed. In our case, the web spider saves parsed sentences into the documents, containing document index of a predicate, an object, subject, and a link to a word trailing letter in a trie. This structureless organization allows us to add new part-of-speech elements without restructuring the database. Also this allows us to model any sentence structure with optional adjectives or participles.

II. Structure parsing algorithm

A. Common work algorithm

To get a parsed sentence, the system performs the following steps:

1. Anaphora resolution
2. Sentence segmentation
3. Token boundaries identification.
4. Part-of-speech tagging of the tokens array.
5. Syntactic parsing of the POS-tagged sequences.

During components selection we tried to use the subsystems, containing English language and preferentially Russian language model.

Table I.
Neural network characteristics

# Layer	Neurons count	Function
1 Output	1	TanH
2	4	TanH
3	7	TanH
4	11	TanH
5	12	TanH
6 Input	9	Linear

B. Anaphora resolution

Anaphora (coreference) resolution systems are less developed, but there are some systems available:

1. OpenNLP
2. CherryPicker
3. JavaRAP (pronoun coreference system)
4. BART
5. ARKref (rule-based)
6. ARS

According to the recommendations provided in [7], we have chosen ARKref as a main anaphora resolution module. ARKref is a deterministic, rule-based system that uses rich syntactic and semantic information to make antecedent selection decisions.

C. Identifying sentence and token borders; POS tagging

This spider system is implemented as a separate unit with a separate API. For sentence and token borders identification, there are many ready solutions available, and this topic is widely covered in the literature. For now, our system works mainly with the English language, but as we might want to extend the list of supported languages in the future, we so we selected an extensible open source Java TreeTagger system. TreeTagger is fast, and has low RAM and CPU consumption with availability of various language models. TreeTagger can be quickly replaced with any other tokenizer.

D. Dependency parsing

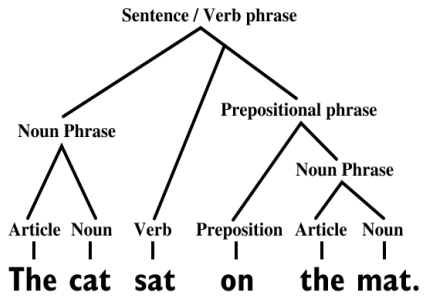


Fig 4: Constituency parsing

Firstly we tried to use Standford NLP Parser as a main sentence processing instrument, but we faced high RAM and CPU consumption. Furthermore, Stanford Parser uses constituency[2] grammars that do not reflect well the structure of languages with relaxed word order, such as Russian. For such languages dependency grammars are usually considered more appropriate.

Following the recommendations in [8], we have chosen MaltParser for best parsing quality from the list of available parsers.

To train the parsing system to recognize any particular language, a deeply annotated text corpus (a treebank) is needed. For each word in a Treebank, the following data is required:

1. Word position in the sentence
2. Word
3. Grammatical attributes
4. Head word position
5. Dependency type

MaltParser contains pre-trained models for English, French, and Swedish. For other languages, it is necessary to create a malttab training set. For the Russian language, the treebank is available as a part of "National Russian language corpus"

III. Implementation

A. The platform

Our search engine consists of two main parts: the search indexer (spider) and the web interface. As most of the NLP software is written in Java, the spider is also written in Java. Since some of the NLP systems operate with space-consuming language models, some heavy weight modules were separated from the base system and made available via RPC API. Thanks to this approach, the system has an ability to use several servers that process different languages (i.e., it is horizontally scalable). Such RPC-available modules are: the anaphora resolution system, the POS tagger, and the dependency parser. For easier development, we have chosen Apache Thrift RPC framework for every isolated component.

As mentioned above, the application stores data in two databases: graph-based Neo4J and document – based MongoDB. The web interface is written in JavaScript/JQuery and operates using Java Spring-based REST API. The component diagram is shown in Figure 5.

All components are implemented in similar ways, and each of them uses a multi-threaded RPC framework, and thus performs multi-threaded text processing.

B. Indexer component

The Indexer component's ("Spider" in the components diagram) aim is to get the next page from the list of links, to process it by calling other components' RPC API and to save the results into the database. Furthermore, this component extracts the links to the new documents to be analyzed, and adds them to the general links list.

This component works with other modules via RPC framework Apache Thrift, that is used due to the simplicity of cross-platform code generation, its lightweight protocol (as opposed to XML-RPC or SOAP), simple implementation and multithreading. At the present time, the system does not support language detection, but the system can operate via RPC with several other processing servers, handling different natural languages. To test the system, we used English Wikipedia as the data source.

C. Anaphora resolution

The anaphora resolution module operates with raw text (cleaned from HTML markup), and replaces pronoun or noun anaphors with their antecedents. As a result, the spi-

[2]Constituency grammar is based on Chomsky's generative grammar. Parsers based on constituency grammars try to divide sentences into smaller word groups until the individual tokens are identified. The example of phrase-structure (constituency) parsing is shown in Figure 4.

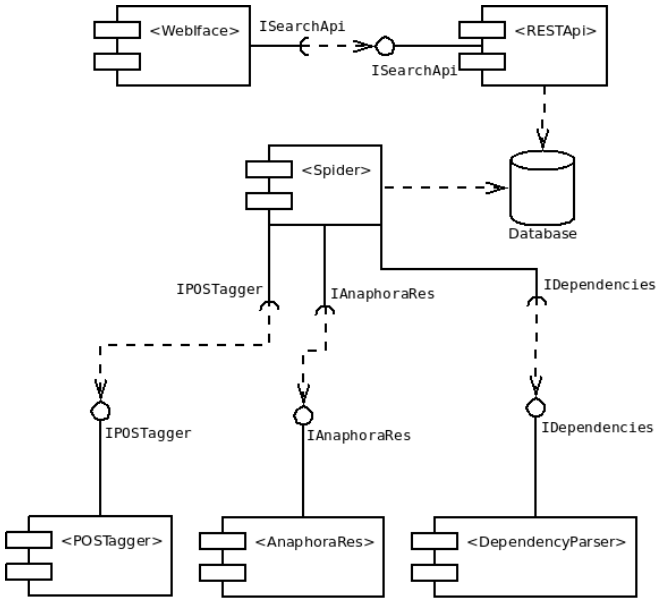


Fig 5: Components diagram

der gets two text versions: the raw text and the text with resolved coreferences. The latter document is being processed in other modules, but both are saved to the database. The anaphora resolution module is a multithreaded server. The number of threads is set up in the server configuration.

D. POS tagging and Dependency finder

These components are marked as "POSTagger" and "DependencyParser" in the components diagram.

Component class diagram for these modules is similar to the coreference resolution component it uses RPC-server classes, singleton configuration classes and other.

MaltParser makes output data in the CoNLL format, similar to the malttab format.

E. Web interface

The web interface is a web application, written in JavaScipt/JQuery. The current web interface allows the user to input three words: subject, predicate and object, to be sent to the server via the REST API. The system processes the query and finds the list of suitable sentences in the database.

REST API is implemented with Java Spring framework.

Using separate processing modules leads to ability of search query NLP processing. This would allow users to make queries as usual sentences.

IV. Related work

[9] also proposed similar semantic network storing approach. Author offers to store RDF structures like a graph using object-oriented databases.

[10] describes a system that processes automatic text sentences tagging for further text managing analyzing or searching.

Our system novelty essence is the approach to process, store and search text data. The method novelty lies in transformation text into RDF-like semantic network and following triplet search over the prepared semantic network index.

V. Conclusion

Our research aim was to try to create a semantic-powered search engine that uses NLP technologies. During the development we have analyzed different information retrieval and NLP instruments and methods, such as syntactic parsing, POS tagging, and coreference analysis.

As the result, we got a semantic sentence-structure search engine prototype. Currently, the system has the following limitations:

1. System processes English-language documents only.
2. The useful content extraction module reliably parses Wikipedia documents only.
3. The current system operates only with triplets. It cannot process adjectives or adverbs.
4. The system does not use any synonyms dictionary.

Finally, system searches triplet like three English words contacted with "AND" boolean operator and it is unusable in current state as providing search service is very poor concerning to internal Wikipedia search (cause Wikipedia is used as testing data source). Search engine returns result with given wittingly right query i.e. known triplet from known document.

Also sentence parsing system is very poor at the moment. We parsed part of Wikipedia for system test. The result is presented in Table 2.

Using synonyms dictionary, a more diverse knowledge-base as a data source, and coreference resolution improvements should make results better.

Table II.
Experiment results

Comment	Value
Total indexed documents	1401
Total parsed sentences	17226
Average parsed sentences in document	12
Average sentences in document	26
Result	54% of document information is lost

References

[1] S. Ilyinsky, M. Kuzmin, A. Melkov and I. Segalovich, “An efficient method to detect duplicates of Web documents with the use of inverted index”, WWW Conference, 2002

[2] S. Brin and L. Page, “The anatomy of a large-scale hypertextual Web search engine”, Proceedings of the seventh international conference on World Wide Web 7, 1998, pp. 107-117

[3] C. D. Manning, P. Raghavan and H. Schütze, “Introduction to Information Retrieval” Cambridge University Press. 2008 p. 6.

[4] A. Gulin, P. Karpovich, D. Raskovalov and I. Segalovich, “Ranking algorithms optimisation using machine-learning methods”, Romip proceedings, 2009

[5] A. Harth and S. Decker, “Optimized Index Structures for Querying RDF from the Web”, Digital Enterprise Research Institute (DERI), National University of Galway, Ireland, 2005, p. 2.

[6] S. Edunov, “How to extract useful content from HTML”, “http://www.algorithmist.ru/2010/11/html-2.html”

[7] Benjamin Chu Min Xian, F. Zahari and D. Lukose, “Benchmarking ARS: Anaphora Resolution System”, Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies, 2011, p. 39

[8] A. Gareyshina, M. Ionov, O. Lyashevskaya, D. Privoznov, E. Sokolova and S. Toldova, “RU-EVAL-2012: Evaluating dependency parsers for Russian” Proceedings of COLING 2012: Posters, IIT Bombay, Mumbai, India, pp. 349-360

[9] V. Bonstrom, A. Hinze, H. Schweppe, "Storing RDF as a graph", Web Congress, 2003. Proceedings. First Latin American , vol., no., pp.27,36, 10-12 Nov. 2003

[10] M. Kalender, Jiangbo Dang, "SKMT: A Semantic Knowledge Management Tool for Content Tagging, Search and Management," Semantics, Knowledge and Grids (SKG), 2012 Eighth International Conference on , vol., no., pp.112,119, 22-24 Oct. 2012

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 261–268
DOI: 10.15439/2014F253
ACSIS, Vol. 2

Extracting Semantic Prototypes and Factual Information from a Large Scale Corpus Using Variable Size Window Topic Modelling

Michał Korzycki, Wojciech Korczyński
AGH University of Science and Technology
in Kraków
ul. Mickiewicza 30, 30-962, Kraków, Poland
Email: {korzycki, wojciech.korczynski}@agh.edu.pl

***Abstract*—In this paper a model of textual events composed of a mixture of semantic stereotypes and factual information is proposed. A method is introduced that enables distinguishing automatically semantic prototypes of a general nature describing general categories of events from factual elements specific to a given event. Next, this paper presents the results of an experiment of unsupervised topic extraction performed on documents from a large-scale corpus with an additional temporal structure. This experiment was realized as a comparison of the nature of information provided by Latent Dirichlet Allocation and Vector Space modelling based on Log-Entropy weights. The impact of using different time windows of the corpus on the results of topic modelling is presented. Finally, a discussion is suggested on the issue if unsupervised topic modelling may reflect deeper semantic information, such as elements describing a given event or its causes and results, and discern it from pure factual data.**

I. Introduction

UNSUPERVISED probabilistic topic modelling is one of the most widely applied information retrieval techniques, in particular in researches on large-scale corpora. Its main assumption states that text documents are mixtures of topics which may be treated as a multinomial probability distribution over words. These distributions might be created with the use of a couple of methods [1], [2], [3].

In this paper we present the results of experiments in which we attempted to extract deeper semantic information from postprocessed results of unsupervised probabilistic topic modelling methods. It is worth emphasizing that unsupervised methods—despite their utility in information retrieval—cannot directly retrieve semantical relations from texts [4], [5]. Some introductory premises were presented by the authors in [6], such as the hypothesis, that although unsupervised topic modelling does not reflect directly semantic prototypes, those prototypes can be inferred from the extracted topics. This paper provides additional research in the discussed subject and presents new results that prove our hypothesis.

The size of the corpus we operated on was millions of documents. Moreover, it possessed an additional temporal structure.

The framework of the experiments presented in this paper is an integral part of a large scale project related to security and intelligence analysis. The suggested approach permits to find in an analyzed text elements related to the specific semantic prototypes (i.e. mental models) of the events described within and to discern them from pure factual information, given that a sufficiently large corpus is available. Discussed technique may be a first step towards creation of a method for automatic semantic prototype identification. Such methods are essential in a security and intelligence analysis as they can be applied in an automatic identification of objects and their properties in full-text sources.

We start our paper with a short overview of works related to the presented subject - it is included in Section II. Section III introduces and explains the concept of semantic prototypes, firstly described in [7]. Subsequent Sections focus on the experiments of unsupervised topic extraction performed in order to present a method of discerning semantic prototypes from factual information: Section IV introduces corpus used in our experiments, Section V describes the method itself. Results of the experiments are presented and discussed in Section VI. We conclude our paper in Section VII. At the end of the paper, in Section VIII, suggestions of future work are discoursed.

II. Related work

Latent Semantic Analysis (LSA) is an original word/document matrix rank reduction algorithm which extracts word co-occurrences in the frame of a text. As a result, each word in the corpus is related to all co-occurring words and to all texts in which it occurs. The LSA algorithm may be applied in various domains—from a text content comparison [8] to an analysis of human association norm [9]. Unfortunately, there is still little interest in studying the linguistic significance of LSA-made associations.

Latent Dirichlet Allocation (LDA), presented by David Blei, Andrew Ng, and Michael Jordan in 2003, is one of the best known generative model used for topic extraction. It assumes that a collection of documents may be represented by a mixture of latent topics, however words creating each topic are chosen according to a multinomial Dirichlet distribution of fixed dimensionality over the whole corpus. LDA is a technique based on the „bag of words" paradigm and it can infer

978-83-60810-58-3/$25.00

distributions of topics e.g. with the use of variational Bayes approximation [10], [11], Gibbs sampling [2] or expectation propagation [12].

Some recent research was focusing on finding if the relationships coming from the unsupervised topic extraction methods reflect semantic relationship reflected in human association norms. A comparison of human association norm and LSA-made association lists can be found in [4] and it should be the base of the study. Results of the other preliminary studies based on such a comparison: [5], [13], [14], show that the problem needs further investigation. It is worth noticing that all the types of research referred to, used a stimulus-response association strength to make a comparison. The results of the aforementioned research have shown that using unsupervised topic extraction methods one is able to create associations between words that are strongly divergent from the ones obtained by analysing the human generated associations.

As it has been already noticed, the methods mentioned above are not able to retrieve additional semantic information, however in this paper we introduce some postprocessing methods that may be useful in a semantic text classification.

III. Semantic prototypes

The notion of a semantic prototype comes from cognitive theory [7] where a notion is represented by its elements with their features. So, according to this model, a notion of a „bird" would be „composed" of such elements and features as „feathers", „beak" and „ability to fly". Semantic prototypes can also be discussed in the context of event descriptions that occur in texts. Prototype theory has also been applied in linguistics for mapping from phonological structure to semantics.

In a domain of natural language processing, this approach is reflected in so-called content theories. A content theory is used to determine a meaning of phrases and information they carry. One of the classic and most known elements of content theory is the Conceptual Dependency theory that was invented and developed by Robert Schank [15]. His main goal was to create a conceptual theory consistent in every natural language. The theory's main assumptions were: the generalization of representation language and inference about implicitly stated information. The first assumption means that two synonymous sentences should have identical conceptual representations. The second one states that all the implicit information has to be stated explicitly by inferring it from explicit information. Each sentence can be then represented in a form of conceptual dependency graph built of three types of elements: primitives, states and dependencies. Primitives are predicates that represent a type of an action, states specify the preconditions and results of actions and dependencies define conceptual relations between primitives, states and other objects [16]. Accordingly to the Conceptual Dependency theory we may represent the event of „tea drinking" by a sequence of events: „tea making", „cup operating", „tea sipping" and so on, that are composed of action, object etc.

The described event model proved itself to be very useful in many applications [17] and we found it very suitable to quantifiable comparisons to unsupervised topic modelling methods [18].

From that theory we deduce our model of an event - its prototype - a compound structure of actions, actors, states, and dependencies but also composed of preconditions and results, being events themselves. This event model reflects also very well the semantic structure of a text. If a document describes an event, it is almost always presented in the context of the causes of the event and the resulting consequences. This will be reflected in various topic models that will tend to reflect that most texts are represented as a linear combination of multiple topics. Determining from such combination which topic (event) can be classified as a cause and which is an effect would be very interesting, but that issue is beyond the scope of this paper.

On the other hand, the modelled text is composed not only of events, but also features of those events specific only to that instance of the event. As such, a text can be seen as having two aspects - the main event of the text intermingled with the elements of cause and result events and factual features that are specific to that single event. The latter aspect would relate to places, actors and contextual information. The former aspect would relate to generic elements that are common to similar events that occur in the corpus.

This paper focuses on unsupervised identification and retrieval of those two different event model aspects of texts.

IV. Corpus

The experiment was conducted on a subset of a 30-year worldwide news archive coming from the New York Times. That corpus has been chosen as it is interesting for a number of reasons:

- it is freely available,
- some interesting research results have been obtained based on it [19],
- it is quite comprehensive in terms of vocabulary and range of described event types,
- its relatively large size (approximately 90.000 documents per year, for a total of 2.092.828 documents spanning the years 1981-2012) gives an ample opportunity to experiment with various document time spans without impacting noticeably its event scope representiveness and lexicon balance.

After a set of trials with various time spans ranging from months to 15 years, the most pronounced effects of the experiment described below could be obtained by comparing two time spans - one covering 6 months and the other covering 4 years. Those sub-corpora contain over 45.000 and 350.000 documents, respectively.

V. Method

We based the data of our experiment on the term/document matrix populated with Log-Entropy weights [20]. More precisely, the value a_{ij} in the matrix corresponding to the i-th

word and j-th document can be expressed as the usual ratio of a local and a global weight:

$$a_{ij} = e_i \log(t_{ij} + 1)$$

where:

$$e_i = 1 - \sum_j \frac{p_{ij} \log p_{ij}}{\log n}$$

for n - the total number of documents, t_{ij} is the number of times the i-th word occurs in the j-th document and $p_{ij} = \frac{t_{ij}}{g_i}$, with g_i the total number of times the term i occurs in the whole corpus.

After building an LDA model (with d dimensions), we obtain a matrix v of size $n \times d$ composed of elements $v_j k$ describing how much the topic k impacts the document j. The v_{jk} matrix contains per design only non-negative values.

Each topic is in turn represented by a vector of probabilities describing how much a single word participates in this topic, in the form of a $N \times d$ matrix w (for N - the size of the lexicon).

As all the values in the matrices v and w are non-negative, their product contains the cumulative impact of each word for each document, summed over the range of all topics. For each document j and word i, what we will call the word model matrix m of size $n \times N$ composed of elements m_{ji}, is obtained by multiplying the document weights with the topic model $m = vw^\top$.

For a given document j we will now analyze the rank of words in the sorted vector m_{ji}, for each word i.

VI. Experiment

We based our experiment on a subset of the New York Times corpus described above. The two compared subsets were spanning 6 months and 4 years, respectively. We focused on the representation of the news item related to the accident of Kursk, the Russian submarine that sank on 12th August 2000. In order to affirm the results of our experiments, additional tests were performed, focusing on the information about the terrorist attacks launched upon New York City on the 11th September 2001. Their outcome is discussed in the second part of this Section.

The experiments were performed using 2 methods of topic modelling: LDA based on a term/document matrix populated with Log-Entropy weights and the pure Log-Entropy model based on the mentioned matrix. These models were computed basing on texts from a 4-year article span. Additionally, the Log-Entropy model was built on texts from a 6-month range in order to observe the changes resulting in considering different time windows. Finally, a ratio of Log-Entropy results from the 6-month and 4-year ranges was calculated so that we could better analyze changes that took place in models built in different time windows.

A. The accident of Kursk

Below is a fragment of the input text used for the primary experiments focusing on the accident of the Russian submarine Kursk:

A Russian submarine plunged to the seabed in the Barents Sea on Sunday during a naval exercise, possibly after an explosion on board, officials said today. They said the submarine was badly damaged, and was trapped at least 450 feet below the surface. They said they did not know how many of the more than 100 crew members on board were alive or how long they could survive. Tonight the navy began preparing a desperate attempt to rescue the crew. But navy officials said the odds of saving the men were slim. The submarine, called the Kursk, was not carrying nuclear weapons, the navy said, but was powered by two nuclear reactors, raising concerns about possible radioactive contamination. But Russian officials said the reactors had been turned off, and officials in Norway said a scientific vessel in the area had detected no signs of a radioactive leak. A flotilla of ships and rescue vessels was on the scene off Russia's northern coast in rough weather tonight, frantically searching for ways to reach the men. Navy officials said they intended to mount a rescue attempt on Tuesday. News reports said rescue workers had been trying to hook lines to the submarine to bring it air and fuel. If the sub lost power, the men could suffocate and the submarine's compartments could turn unbearably cold in the frigid waters. The navy's commander, Adm. Vladimir Kuroyedov, said, „Despite all the efforts being taken, the probability of a successful outcome from the situation with the Kursk is not very high." The White House spokesman, Joe Lockhart, said that President Clinton had been briefed about the accident, and that his national security adviser, Samuel R. Berger, had told Foreign Minister Igor S. Ivanov that the United States was willing to help.

Results presented below are very similar to the ones described in [6], with the exception of unnecessary terms that carry no importance and that we were able to exclude.

In order to properly understand the results of LDA-based modelling, one has to look at the analyzed event - the sinking of the Kursk submarine - in a more general way as an accident of a naval vehicle that happened in Russia.

After analysing which words are the highest ranked in our model (30 words with the highest score are presented in Table I), it may be observed that LDA model distinguished words that may be somehow connected with:

- vehicles: ship (ranked 2nd with score 0.00418), vessel (7th, 0.00246), plane (8th, 0.00238), boat (11th, 0.00229)
- transport: port (1st, 0.00434), airline (3rd, 0.00304), flight (4th, 0.00276), airport (5th 0.00271), tunnel (15th, 0.00200), pilot (25th, 0.00169), passenger (26th, 0.00168)
- Russia: Russia (6th, 0.00252), Russian (9th, 0.00236), Moscow (28th, 0.00157)
- sea (except for the already mentioned port, ship, vessel, boat): navy (10th, 0.00232), sea (13th, 0.00209), harbor

TABLE I
TOP 30 WORDS BASED ON THE LDA MODEL (THE KURSK ACCIDENT)

no.	Word	LDA model score
1	port	0.00434231934161
2	ship	0.00417804388954
3	airline	0.00304061756597
4	flight	0.00275556563789
5	airport	0.00271146829417
6	Russia	0.00251820497727
7	vessel	0.00246363118477
8	plane	0.0023790659542
9	Russian	0.00236238563557
10	navy	0.00231835409854
11	boat	0.00228900596583
12	mile	0.0021881530589
13	sea	0.00208970459425
14	harbor	0.0020105874839
15	tunnel	0.00200078016362
16	minister	0.00196687109163
17	authority	0.00186827318326
18	profitability	0.00180866361081
19	crew	0.00180463741216
20	air	0.00178208129395
21	official	0.00175995025308
22	treaty	0.00171964506823
23	hart	0.00170921811033
24	united	0.00169356866222
25	pilot	0.00169325060507
26	passenger	0.00167712980293
27	naval	0.0016418101694
28	Moscow	0.00156616846396
29	shipping	0.00153479445129
30	state	0.00145645985742

(14th, 0.00201), naval (27th, 0.00164), shipping (29th, 0.00153), water (49th, 0.00121), sailor (54th, 0.00119)

- accidents: besides many of the words already mentioned, the word crash (33rd, 0.00139) is significant.

These words are very general and are common terms used while describing some event. It has to be emphasized that there is no word specific for a given event. They were filtered out in accordance with the nature of LDA that rejects words characteristic for just a narrow set of documents and promotes words that are specific to extracted topics. Therefore, we cannot expect highly ranked terms that would be strictly connected with the accident of the Kursk submarine but rather words related generally to accidents or vehicles, transport, sea and Russia.

These words are very general and descriptive. Using them, it is not possible to state anything specific („factual") about the nature of a given event, its causes or consequences.

Analysing the results of the Log-Entropy model calculations, we are able to see that the highest ranked words are more specific than in the case of the LDA model.

TABLE II
TOP 30 WORDS BASED ON THE LOG-ENTROPY MODEL IN A 4-YEAR TIME WINDOW (THE KURSK ACCIDENT)

no.	Word	Log-Entropy model score
1	submarine	0.22720800350779063
2	Kursk	0.21964992269828088
3	minisub	0.2072304031428077
4	Barents	0.18764922042161675
5	navy	0.14100811315980294
6	reactor	0.1318971983154962
7	thresher	0.1257656918120723
8	Russian	0.12420286868127497
9	vessel	0.11552029612400631
10	rescue	0.11231821882564066
11	naval	0.11103796756237282
12	crew	0.10625421439440351
13	nuclear	0.10489886205812088
14	diving	0.1021749945646489
15	fleet	0.10093483510873145
16	accident	0.09974534960726877
17	hatche	0.09903148933253046
18	pressurized	0.0973239282607489
19	Russia	0.09658403552800936
20	ship	0.09603201115598564
21	baker	0.08955212916344289
22	Kuroyedov	0.08620535346215855
23	sank	0.0847013514975405
24	sea	0.08276943964465774
25	Lockhart	0.08162721527202152
26	radioactive	0.0787603000930047
27	Nilsen	0.07615887534697596
28	breathable	0.07448263140205012
29	Komsomolet	0.07436551614837245
30	stricken	0.07292194644563503

Table II presents the 30 highest ranked words according to the Log-Entropy model in 4-year time window. Among them there are ones that are related to the causes of the main event:

- *reactor* (6th, 0.13190), *nuclear* (13th, 0.10490), *radioactive* (26th, 0.07876): despite the fact that in case of Kursk accident reactors shutting down is rather a consequence, many news described also some previous submarine accidents caused by malfunction of nuclear reactors
- *accident* (16th, 0.09975): some „accident" as a reason of submarine sinking
- *pressurized* (18th, 0.09732): media reported that the lack of pressurized escape chambers was the reason why the crew was not able to get out of a submarine
- *sank* (23rd, 0.08470): „the submarine sank" as a the central event
- *stricken* (30th, 0.07292): „submarine was stricken" as a reason of the accident

Words that can also be found as related to the consequences of the discussed accident:

- *minisub* (ranked 3rd with score 0.20723): a minisub was sent with a rescue mission
- *Thresher* (7th, 0.12577): USS Thresher was a submarine, which sinking was frequently compared to the accident of Kursk
- *rescue* (10th, 0.11232) and *crew* (12th, 0.10625): rescue crew was sent in order to help sailors
- *Kuroyedov* (22nd, 0.08621): Fleet Admiral Vladimir Kuroyedov was in charge of navy when Kursk sank and therefore after the accident spoke with the media very often
- *Lockhart* (25th, 0.08163): Joe Lockhart was the White House spokesman that talked to the media after the accident of Kursk and informed about the American president's offer of help
- *Nilsen* (27th, 0.07616): Thomas Nilsen is a Norwegian researcher that wrote a report on Russian fleet. He was also interviewed by media after the accident
- *Komsomolet* (29th, 0.07437): K-278 Komsomolet was a Soviet nuclear-power submarine that was mentioned frequently in many reports on Soviet/Russian fleet after the accident of Kursk
- *stricken* (30th, 0.07292): „stricken submarine" as a consequence of the accident

These words are much more specific than in the case of those extracted by the LDA model. They strictly concern this event and describe its causes and consequences.

At first sight, the results of Log-Entropy model calculations in 6-month time window are very similar to the previous ones. We can see the same elements that we identified as the cause of the accident (*sank, stricken, reactor, nuclear*) and its consequences (eg.: *minisub, Thresher, rescue, crew, Kuroyedov*). However, the results yielded in these two time windows differ in scores. In order to analyze how the rank of particular words changed, we calculated a ratio of each word's score in two time span windows - spanning 4 years and 6 months. Having in mind that changing the time window practically does not change the local weight of a given term but changes its global weight, this ratio would emphasize these changes as a comparison of each word's global weights while similar local weights would become irrelevant.

As the Table III presents, it turned out that the words that could be used in describing causes and consequences of the Kursk sinking are now much more emphasized. Moreover, the most specific for this particular event words are stressed, while terms that could be used in descriptions of other, similar accidents (e.g. reactor, nuclear, radioactive, rescue, crew) have lower rank. Besides, new interesting words appeared when considering a ratio-based ranked list of words:

- *Kuroyedov* (ranked 3rd), *minisub* (ranked 4th), *Nilsen* (ranked 6th): they are still high ranked as the most specific words for this particular event
- *seabed* (ranked 5th): as a consequence of the accident, the submarine was plunged to the seabed

TABLE III
TOP 30 WORDS BASED ON THE LOG-ENTROPY MODELS RATIO (THE KURSK ACCIDENT)

no.	Word	Log-Entropy ratio
1	Kursk	1.24871629725
2	Barents	1.19631168918
3	Kuroyedov	1.17117438796
4	minisub	1.1623897656
5	seabed	1.08585777559
6	Nilsen	1.07438727067
7	Vladimir	1.0649817326
8	torpedo	1.06238346439
9	flotilla	1.06017391338
10	Thresher	1.05610904592
11	photo	1.05231801755
12	certified	1.05071293879
13	outcome	1.04815882266
14	avalon	1.0380988501
15	sailor	1.03431234091
16	periscope	1.02936384958
17	fuel	1.02728401954
18	site	1.02714550054
19	hatche	1.02449383024
20	submarine	1.023577913
21	underwater	1.02209583656
22	torpedoes	1.01904285092
23	Ivanov	1.01647340538
24	doubtful	1.01608381017
25	hull	1.01556031012
26	naval	1.01495656826
27	Joe	1.01339486585
28	Baker	1.0120830858
29	sunk	1.01187373966
30	sunken	1.01048072834

- *torpedo* (ranked 8th), *torpedoes* (ranked 22nd): an explosion of one of torpedoes that the Kursk was carrying, has been recognized as the main reason of the accident
- *Ivanov* (ranked 23rd): in time of the Kursk sinking Sergei Ivanov was the head of the Russian Security Council, therefore was highly involved in this case, so his name was often mentioned as a consequence of this accident
- *hull* (ranked 25th): after the accident, the rescue crew tried to get into the submarine through its hull
- *slim* (ranked 35th): day after day the chances of saving sailors were slimmer

It seems very interesting how calculating of the ratio helped with finding new words describing causes and consequences and how it distinguished terms that are specific for a given event. It also emphasized changes that occurred in different time windows.

B. September 11 terrorist attacks

Some additional tests needed to be launched in order to affirm results obtained in the previously performed experiments.

A fragment below exemplifies the input text used in the subsequent experiments, focused on the September 11 terrorist attacks on New York City:

> *Hijackers rammed jetliners into each of New York's World Trade Center towers yesterday, toppling both in a hellish storm of ash, glass, smoke and leaping victims, while a third jetliner crashed into the Pentagon in Virginia. There was no official count, but President Bush said thousands had perished, and in the immediate aftermath the calamity was already being ranked the worst and most audacious terror attack in American history. The attacks seemed carefully coordinated. The hijacked planes were all en route to California, and therefore gorged with fuel, and their departures were spaced within an hour and 40 minutes. The first, American Airlines Flight 11, a Boeing 767 out of Boston for Los Angeles, crashed into the north tower at 8:48 a.m. Eighteen minutes later, United Airlines Flight 175, also headed from Boston to Los Angeles, plowed into the south tower. Then an American Airlines Boeing 757, Flight 77, left Washington's Dulles International Airport bound for Los Angeles, but instead hit the western part of the Pentagon, the military headquarters where 24,000 people work, at 9:40 a.m. Finally, United Airlines Flight 93, a Boeing 757 flying from Newark to San Francisco, crashed near Pittsburgh, raising the possibility that its hijackers had failed in whatever their mission was. There were indications that the hijackers on at least two of the planes were armed with knives. Attorney General John Ashcroft told reporters in the evening that the suspects on Flight 11 were armed that way.*

Table IV presents top 30 results of LDA-based modelling performed in a 4-year time span window. As previously, we are able to perceive some groups of words, linked together by a certain topic:

- war: attack (ranked 1st with score 0.00270), war (8th, 0.00176), military (13th, 0.00163), force (29th, 0.00113)
- United Stated of America: Bush (2nd, 0.00266), American (6th, 0.00184), York (24th, 0.00121)
- terrorism (except for already mentioned attack, force): anthrax (3rd, 0.00203), terrorist (18th, 0.00143)
- public service: police (4th, 0.00196), security (20th, 0.00139), firefighter (38th, 0.00099)
- Afghanistan: Afghanistan (5th, 0.00190), Taliban (7th, 0.00183)
- aircraft: airline (9th, 0.00175), plane (23rd, 0.00123), airport (25th, 0.00118), flight (28th, 0.00114)
- society: state (14th, 0.00161), government (17th, 0.00146), president (22nd, 0.00124), nation (26th, 0.00115), administration (32nd, 0.00104), country (42nd, 0.00098)

As it was observed previously, LDA-based modelling rejects words that are specific for a given event. The highest ranked terms are general and cannot be linked with any factual information. As one can see, there are no words that are characteristic for the September 11 terrorist attacks but rather words that could be related to any document focused on the subject of war, terrorism, United States of America and so on.

TABLE IV
TOP 30 WORDS BASED ON THE LDA MODEL (THE SEPTEMBER 11TH TERRORIST ATTACKS)

no.	Word	LDA model score
1	attack	0.00270258155503
2	Bush	0.00265822334759
3	anthrax	0.00202525476511
4	police	0.00196184349059
5	Afghanistan	0.00190457356009
6	American	0.00183689824487
7	Taliban	0.00182571967641
8	war	0.00176326965287
9	airline	0.00174704614254
10	bin	0.00171653103426
11	official	0.00171409002157
12	united	0.00165387405236
13	military	0.0016267486976
14	state	0.00160529481878
15	Laden	0.00148723727284
16	people	0.00147777652497
17	government	0.00145696458969
18	terrorist	0.00142711844383
19	city	0.00139962463902
20	security	0.00139337982458
21	world	0.00136066739187
22	president	0.00124062013179
23	plane	0.00122558021987
24	York	0.00120866113945
25	airport	0.00118486349022
26	nation	0.00115097052471
27	center	0.0011502315111
28	flight	0.00113663752416
29	force	0.00112785897603
30	time	0.00107041659097

These conclusions are very similar to the ones drawn in case of the previous experiments.

The results of Log-Entropy model calculations are also analogous to the case of documents related to Kursk accident, including the possibility of distinguishing causes and consequences of a given event, however we decided not to present them for the reason of shortening the paper.

Bigger expressiveness might be attributed to the ratio of each word's Log-Entropy model score in two time span windows - spanning, as previously, 4 years and 6 months.

Table V presents 30 words with the highest ratio of Log-Entropy model score in two forementioned time span windows. As it might be noticed, there are much more terms that are related to the particular event - September 11 terrorist

TABLE V
TOP 30 WORDS BASED ON THE LOG-ENTROPY MODELS RATIO (THE SEPTEMBER 11 TERRORIST ATTACKS)

no.	Word	Log-Entropy ratio
1	terrorist	1.57893538599
2	attack	1.42460445953
3	Afghanistan	1.42152193493
4	Osama	1.40325332485
5	Taliban	1.36639135448
6	Afghan	1.3476264106
7	bin	1.3345158298
8	hijacker	1.32412486848
9	Kabul	1.31277718096
10	Laden	1.31109812221
11	hijacked	1.26364453579
12	terror	1.25068637547
13	Pentagon	1.22267231541
14	hijacking	1.22261266319
15	trade	1.22143888629
16	Bush	1.19449256298
17	aftermath	1.19022203646
18	Islamic	1.18133136868
19	firefighter	1.16997652186
20	tower	1.16270222181
21	Ashcroft	1.16229315062
22	jetliner	1.14621366889
23	rubble	1.13927570543
24	inhalation	1.13484053252
25	plane	1.12086607749
26	disaster	1.09775233687
27	twin	1.08233678639
28	airline	1.08152641944
29	Vesey	1.08150236466
30	militant	1.07469496007

attacks. In this case, more general words that could be used in any other description of attack, war, etc. are less stressed.

Moreover, we are able to distinguish words that could be considered as causes and consequences of the given event:

- *Osama* (ranked 4th with score 1.40325), *bin* (7th, 1.33451), *Laden* (10th, 1.31110): Osama bin Laden was the founder of terrorist organisation al-Qaeda which was responsible for launching the attacks
- *Afghanistan* (3rd, 1.42152), *Afghan* (6th, 1.34763), Kabul (9th, 1.31278): the war in Afghanistan (with the capital in Kabul) was one of the consequences of terrorist attacks on 11th of September 2001
- *hijacker* (8th, 1.32412), *hijacked* (11th, 1.26364453579), *hijacking* (14th, 1.22261): the planes were used as destructive weapons, because they were hijacked
- *aftermath* (17th, 1.19022): the usage of this word indicates an introduction of a given event's consequences
- *Ashcroft* (21st, 1.16229): on 11th of September 2001 John Ashcroft was an Attorney General who, in consequence of the terrorist attacks, was a supporter of passage of one of the main antiterrorism acts (USA Patriot Act)
- *rubble* (23rd, 1.13928), *disaster* (26th, 1.09775), *crashed* (31st, 1.07208), *perished* (36th, 1.06022): these are some words used to describe the consequences of an attack
- *rescue* (33rd, 1.07091), *rescuer* (35th, 1.06506), *evacuated* (39th, 1.05352): in a consequence of the terrorist attack, rescue crews tried to help people and evacuate then from the World Trade Center

Again, it proved that Log-Entropy model discerns more factual data than LDA-based model. Moreover, calculating of the ratio of score obtained in two time span windows helped us to find new interesting terms and stress the changes of results in different time windows.

VII. CONCLUSION

In this paper we extended our work introduced in [6] where we introduced the concept of the text being a structure consisting of a mixture of event descriptions and factual information. Additional experiments performed on the second subset of the large-scale corpus proved again that some methods of postprocessing the results of unsupervised methods could help model an event in a semantically meaningful way, reflecting its semantic structure. Moreover, by comparing the results of Latent Dirichlet Allocation (LDA) and Vector Space Model methods we were able once more to observe how the former distinguished descriptive and general information, while the latter emphasized more specific terms. This specific information could be useful in description of event's causes and consequences.

However, it has to be stressed that the method of discerning causes and consequences of a given event is not a subject of our work and would be an interesting topic of future work. In our paper we tried to distinguish causes and consequences more or less accurately without any advanced method.

VIII. FUTURE WORK

This paper presents the new experiments and affirms the authors' hypothesis discussed already in [6] that by comparing the results of topic modelling and vector modelling coming from different subsets of a corpus, varying by time scope and size, the obtained information can be additionally graded by the level of its generality or specificity. That in turn can show us a way to create a method for discerning semantic prototypes (general description of events) from factual information (specific to events).

However, as seen in the preliminary results above, this hypothesis is supported by manually verified examples that do not scale to a more generic case. Thus, the current ongoing research focuses on creating a metric being able to assess automatically the level of generality or the amount of facts in a specific result. Such a metric takes into consideration factors related to the relative amount of Named Entities in the results, the distance from various text clusters obtained via topic modelling etc. By defining such a metric, crucial parameters for a correct fact versus semantic prototype extraction method

can be automatically determined. Some of the parameters currently considered are: the chosen time window relative and absolute sizes (the analyzed corpus covers over 30 years of press notes), the time shift of the window time frame relative to the analyzed event (preceding, succinct or just surrounding), the topic modelling methods settings.

The long term goal of this research is the creation of a method for automatic semantic prototype identification. Pure unsupervised methods, as those presented in this paper, are not the only venue of approach considered. A parallel research is conducted, based on human based association networks, as presented in [4]. We expect to obtain valuable results coming from the convergence of both approaches.

ACKNOWLEDGMENTS

The research reported in the paper was partially supported by grants "Advanced IT techniques supporting data processing in criminal analysis" (No. 0008/R/ID1/2011/01) and "Information management and decision support system for the Government Protection Bureau" (No. DOBR-BIO4/060/13423/2013) from the Polish National Centre for Research and Development.

REFERENCES

[1] M. Steyvers and T. Griffiths, "Probabilistic topic models," in *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, 2005. [Online]. Available: http://psiexp.ss.uci.edu/research/papers/SteyversGriffithsLSABookFormatted.pdf

[2] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. Suppl. 1, pp. 5228–5235, April 2004. doi: 10.1073/pnas.0307752101. [Online]. Available: http://dx.doi.org/10.1073/pnas.0307752101

[3] J. Boyd-Graber, J. Chang, S. Gerrish, C. Wang, and D. Blei, "Reading tea leaves: How humans interpret topic models," in *Neural Information Processing Systems (NIPS)*, 2009.

[4] I. Gatkowska, M. Korzycki, and W. Lubaszewski, "Can human association norm evaluate latent semantic analysis?" in *Proceedings of the 10th NLPCS Workshop*, 2013, pp. 92–104.

[5] T. Wandmacher, "How semantic is latent semantic analysis?" in *Proceedings of TALN/RECITAL*, 2005.

[6] M. Korzycki and W. Korczyński, "Does topic modelling reflect semantic prototypes?" in *New Research in Multimedia and Internet Systems*, ser. Advances in Intelligent Systems and Computing, A. Zgrzywa, K. Choroś, and A. Siemiński, Eds. Springer International Publishing, 2015, vol. 314, pp. 113–122. ISBN 978-3-319-10382-2. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-10383-9_11

[7] E. Rosch, "Principles of categorization," in *Cognition and categorization*, E. Rosch and B. Lloyd, Eds. Hillsdale, New Jersey: Erlbaum, 1978, pp. 27–48.

[8] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990. doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9. [Online]. Available: http://dx.doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

[9] D. Ortega-Pacheco, N. Arias-Trejo, and J. B. B. Martinez, "Latent semantic analysis model as a representation of free-association word norms." in *MICAI (Special Sessions)*. IEEE, 2012. doi: 10.1109/MICAI.2012.13. ISBN 978-1-4673-4731-0 pp. 21–25. [Online]. Available: http://dx.doi.org/10.1109/MICAI.2012.13

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944937

[11] R. Řehuřek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[12] T. Minka and J. Lafferty, "Expectation-propagation for the generative aspect model," in *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002. ISBN 1-55860-897-4 pp. 352–359. [Online]. Available: http://dl.acm.org/citation.cfm?id=2073876.2073918

[13] M. Wettler, R. Rapp, and P. Sedlmeier, "Free word associations correspond to contiguities between words in texts." *Journal of Quantitative Linguistics*, vol. 12, no. 2-3, pp. 111–122, 2005. doi: 10.1080/09296170500172403. [Online]. Available: http://dx.doi.org/10.1080/09296170500172403

[14] T. Wandmacher, E. Ovchinnikova, and T. Alexandrov, "Does latent semantic analysis reflect human associations?" in *Proceedings of the Lexical Semantics workshop at ESSLLI'08*, 2008.

[15] R. C. Schank, "Conceptual dependency: A theory of natural language understanding," *Cognitive Psychology*, vol. 3, no. 4, pp. pages 532–631, 1972. doi: 10.1016/0010-0285(72)90022-9. [Online]. Available: http://dx.doi.org/10.1016/0010-0285(72)90022-9

[16] S. L. Lytinen, "Conceptual dependency and its descendants." *Computers and Mathematics with Applications*, vol. 23, pp. 51–73, 1992. doi: 10.1016/0898-1221(92)90136-6. [Online]. Available: http://dx.doi.org/10.1016/0898-1221(92)90136-6

[17] W. Lubaszewski, K. Dorosz, and M. Korzycki, "System for web information monitoring," in *Computer Applications Technology (ICCAT), 2013 International Conference on*, Jan 2013. doi: 10.1109/ICCAT.2013.6522053 pp. 1–6. [Online]. Available: http://dx.doi.org/10.1109/ICCAT.2013.6522053

[18] K. Dorosz and M. Korzycki, "Latent semantic analysis evaluation of conceptual dependency driven focused crawling," in *Multimedia Communications, Services and Security*, ser. Communications in Computer and Information Science. Springer Berlin Heidelberg, 2012, vol. 287, pp. 77–84. ISBN 978-3-642-30720-1. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-30721-8_8

[19] K. Leetaru, "Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space," *First Monday*, vol. 16, no. 9, 2011.

[20] T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, Eds., *Handbook of Latent Semantic Analysis*, ser. University of Colorado Institute of Cognitive Science Series. Mahwah, New Jersey, USA: Lawrence Erlbaum Associates, 2007. ISBN 9780805854183

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 269–272
DOI: 10.15439/2014F25
ACSIS, Vol. 2

An approach to service-oriented information systems architecture development based on semantic closure measure

Viktor Mokerov
D. Serikbayev East Kazakhstan state technical university
69 A.K. Protozanov Street, Ust-Kamenogorsk, The Republic of Kazakhstan
Email: Viktor.O.Mokerov@ieee.org

***Abstract*—The goal of this research is to design the approach to automated development of service-oriented architecture of information systems based on ontology. The approach presented in the paper is based on methodology of system and ontology design. Novelty of the approach is in application of semantic closure of system functions as clustering criteria. The results of the research were used in practice for design of e-university knowledge database. This paper is a part of the author's PhD thesis.**

I. Introduction

APPLICATION of ontology in information systems design tasks became widely used mostly in recent years. On one hand, ontology is based on sophisticated mathematical tool - formal logic. On the other hand, expressive graphical notations developed for it make its application not more complicated than other modeling languages. In papers [1], [2], [3], [4] approaches to generation of information system structures based on system ontology and structure samples are presented. Generation of system ontology based on descriptive models is a critical task as for existing system modification as well as one of steps for system design. Issues of system ontological description development based on project documentation and views in modeling languages are described in papers [5], [6], [7], [8].

Possible solution for tasks of systems modernization and integration using knowledge database is suggested in [9], [10], [11], [12], [13]. Papers [14], [15], [16], [17], [18] are devoted to issues of knowledge databases design for information systems based on ontological approach.

One of the main characteristics of service-oriented architecture is a loose coupling of systems components that provides for support of system modifiability and scalability. The task of the architect designing a system based on service-oriented architecture is to develop system structure that provides for minimum relations between different services. Architect decomposes functions into services in accordance with functions semantic. Functions working mostly with the same fragment of subject area will be related to the same service, while functions that are not related by semantic will be related to different services. Possessing knowledge of functions semantic, this task can be automated to a large degree. This paper suggests approach to solving automation task based on system ontology. Services structure development automation allows, on one hand, partially reduce load on the architect in solving routine tasks. On the other hand, given large quantity of initial data the application of suggested methods allows creating system prototype that can be used by the architect.

II. Ontology of service-oriented system structure

To solve task of design automation, it is required to reflect system elements in ontology. The main structural elements for service-oriented system are services and their operations (functions). Services and functions cannot exist on its own, they always relate to some business process that they implement, as well as possess information on physical objects of a system - contents. Knowledge of system structure, as well as of semantic of processed information, is required during system modernization. This is why it is feasible to perform ontological engineering of subject area with the purpose of developing separate ontology containing subject area semantic. Information system structure ontology in notation IDEF5 is given in Fig. 1.

Concept $Entity$ describes all contents of a subject area that are described in subject ontology. $Entity$ serves as super-concept for all contents of the subject area ontology. Concept $Process$ presents business process of the subject area. Samples of this concept are developed based on the results of system analysis of the subject area. In accordance with Zachman's model, business processes are reflected in concept view. To model business processes the visual modeling languages can be used: UML [11], BPMN [12], IDEF0 [13]. Description of business processes given in notations can be brought to ontology as suggested in [10], [14], [15]. Concept $ProcessFunction$ describes functions that provide for completion of business processes. Concept $Service$ describes information systems services. Concept $ServiceOperation$ describes operations of a service. Service operation is used to perform one or more functions in business process.

The following assumptions are used as limitations:

- each business process consists of at least one function;
- each function can write or read information from at least one content;

978-83-60810-58-3/$25.00

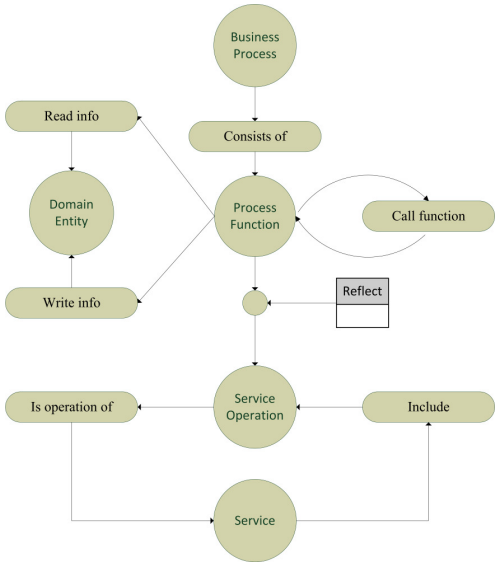


Fig. 1. Ontology of service-oriented system structure

- business process function can call for another function of this or another business process;
- service consists of at least one operation.

Role $ConsistsOf$ determines membership of function in business process. Role $Call$ is used to express call of one function from another. Role $ReadInfo$ sets reading relations between a function and a content of subject area, role $WriteInfo$ sets writing relations. Role $Include$ determines inclusion of operation into service composition. Reverse role $isOperationOf$ sets functional relation of function membership in service. Role $Reflect$ is introduced to reflect service operations at business processes functions.

III. Formalization of service-oriented system structure for automation of its development

Task of generation of services set S comes to clustering task for system functions set $F = \{f_1, \ldots, f_n\}$ with unknown quantity of clusters.

Functional maximizing semantic closure between functions of one service is used as criteria of clustering quality:

$$G = \sum_{i}^{k} \sum_{f_x \in S_i} SI(f_x, f_i) \rightarrow \max \qquad (1)$$

Where k is the number of clusters, f_i is the center of cluster S_i, SI is the similarity functions between f_x and f_i business process functions.

Function of cluster elements similarity (SI) calculates measure of semantic closure for two functions f_x and f_y of subject area in accordance with the following relations:

$$SI(f_x, f_y) = \sum_{i=1}^{n} w_i SI_i \qquad (2)$$

$$\sum w_i = 1$$

$$SI_{BP}(f_x, f_y) = \frac{|\exists Process.ConsistsOf.f_x \sqcap \exists Process.ConsistsOf.f_y|}{|\exists Process.ConsistsOf.f_x \sqcup \exists Process.ConsistsOf.f_y|}$$

$$SI_{ER}(f_x, f_y) = \frac{|\exists f_x.ReadInfo.Entity \sqcap \exists f_y.ReadInfo.Entity|}{|\exists f_x.ReadInfo.Entity \sqcup \exists f_y.ReadInfo.Entity|}$$

$$SI_{ER}(f_x, f_y) = \frac{\exists |f_x.WriteInfo.Entity \sqcap \exists f_y.WriteInfo.Entity|}{|\exists f_x.WriteInfo.Entity \sqcup \exists f_y.WriteInfo.Entity|}$$

Where $SI_{BP}(f_x, f_y)$ defines semantic closeness as a ratio of the cardinality of the subset of business processes, which connected by $ConsistsOf$ role with both functions f_x, f_y, to the cardinality of the subset of business processes, which connected by $ConsistsOf$ role with any of the functions f_x, f_y.
$SI_{ER}(f_x, f_y)$ - defines semantic closeness as a ratio of the cardinality of the subset of entities, which connected by $ReadInfo$ role with both functions f_x, f_y, to the cardinality of the subset of entities, which connected with $ReadInfo$ role with any of the functions f_x, f_y.
$SI_{ER}(f_x, f_y)$ - defines semantic closeness as a ratio of the cardinality of the subset of entities, which connected by $WriteInfo$ role with both functions f_x, f_y, to the cardinality of the subset of entities, which connected with $WriteInfo$ role with any of the functions f_x, f_y.

Service structure obtained as a result of algorithm performance meets the requirement of weak system relatedness.

IV. Experimental evaluation of the method

Experimental evaluation of the suggested method was performed using processes of educational portal of East Kazakhstan State Technical University developed within the program of grants financing No. state registration 0112PK01674 (2012-2014), as application of e-university knowledge database. The fragment of electronic portal chosen for method evaluation included:

- processes: Curriculum development (P1), Education program expertise (P2), Choice of individual educational path (P3), Selection of graduates by qualification (P4);
- contents: Goal of educational program (E1), Results of educational program learning (E2); Educational program (E3), Module (E4), Course (E5), Competence (E6), Student (E7);
- functions: Search of module prerequisites (F1), Evaluation of completeness of educational program (F2), Evaluation of correctness of educational program (F3), Evaluation of goals and results of educational programs

TABLE I
SEMANTIC CLOSURE OF FUNCTIONS

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
F1	0.80	0.35	0.35	0.05	0.40	0.25	0.35	0.56	0.43	0.08	0.12	0.08
F2	0.35	0.80	0.60	0.25	0.05	0.18	0.10	0.45	0.10	0.09	0.17	0.09
F3	0.35	0.60	0.80	0.12	0.00	0.13	0.10	0.40	0.10	0.08	0.12	0.00
F4	0.05	0.25	0.12	0.30	0.00	0.06	0.12	0.24	0.05	0.10	0.13	0.04
F5	0.40	0.05	0.00	0.00	1.00	0.13	0.50	0.25	0.32	0.00	0.06	0.10
F6	0.25	0.18	0.13	0.06	0.13	0.80	0.28	0.25	0.38	0.13	0.19	0.13
F7	0.35	0.10	0.10	0.12	0.50	0.28	1.00	0.31	0.27	0.08	0.12	0.00
F8	0.56	0.45	0.40	0.24	0.25	0.25	0.31	0.80	0.19	0.12	0.09	0.05
F9	0.43	0.10	0.10	0.05	0.32	0.38	0.27	0.19	0.80	0.24	0.29	0.24
F10	0.08	0.09	0.08	0.10	0.00	0.13	0.08	0.12	0.24	0.80	0.60	0.15
F11	0.12	0.17	0.12	0.13	0.06	0.19	0.12	0.09	0.29	0.60	0.80	0.10
F12	0.08	0.09	0.00	0.04	0.10	0.13	0.00	0.05	0.24	0.15	0.10	0.80

TABLE II
MATRIX OF RELATIONS OF FUNCTIONS AND PROCESSES OF EDUCATIONAL PORTAL

Function	P1	P2	P3	P4
F1	1	1		
F2		1		
F3		1		
F4	1	1		
F5	1			
F6	1	1	1	1
F7	1			
F8	1	1		
F9	1		1	1
F10			1	
F11			1	
F12				1

TABLE III
MATRIX OF RELATIONS OF FUNCTIONS AND CONTENTS OF SUBJECT AREA BY ROLES $WriteInfo$ / $ReadInfo$

Function	E1	E2	E3	E4	E5	E6	E7
F1				/1		/1	
F2	/1	/1	/1	/1	/1	/1	
F3			/1	/1			
F4	/1	/1	/1	/1	/1		
F5				1/	1/	1/	
F6						/1	
F7	1/	1/	1/	/1	/1		
F8	/1	/1	/1	/1			
F9				/1		/1	
F10		/1		/1			/1
F11			/1	/1	/1	/1	/1
F12		/1				1/	/1

(F4), Generation of courses module (F5), Course search (F6), Module curriculum generation (F7), Evaluation of curriculum (F8), Search of alternative modules (F9), Evaluation of influence of individual curriculum of a students on results of education (F10), Generation of individual student curriculum (F11), Search of graduates by required competencies (F12).

Relations of functions and processes by role $ConsistsOf$ were determined based on process model of educational portal. They are presented in table II.

Data of table II shows that some system functions are involved in few processes at the same time, which makes it impossible to use processes as main criteria for functions uniting.

Based on the results of the analysis of information calls to database, functions relations with subject area contents by roles $ReadInfo$ and $WriteInfo$ were chosen. They are presented in table III.

Results of calculating measure of semantic closure of functions SI in accordance with (2) are given in table I.

Obtained evaluation of semantic closure was used for functions clustering. Algorithm FOREL was used to perform clustering. Experiment was conducted for neighbor elements search radius of 0.76 - average value of distance matrix composed from one complements of semantic closure matrix elements. The following clusters were obtained for search radius:

- F1;F5;F6;F7;F9;F10;F11;F12;
- F2;F3;F4;F8;

There are no linear call sequences in obtained clusters, therefore, all obtained functions can be presented as service operations.

As the results of experimental evaluation showed, set of developed services conforms to services, which were designed empirically without using automation means.

Results showed:

1) functions close by semantic entered into one cluster;
2) functions with weak relations, not depending on used neighbor elements radius search, were chosen into separate cluster (service);
3) distribution of functions with high degree of relatedness depends on clustering parameters that allows identify those functions during further attempts varying neighbor elements search radius;
4) when radius close to average of distance matrix is used, formed clusters to a larger degree conform to services formed during empirical designing without using automatic generation of system structure.

The following services can be conditionally highlighted based on clusters formed using search radius 0.76:

- data search and input;

- analysis and expertise.

Identification of functions with high degree of relatedness allows designer to perform optimization of process model, in order to reduce relatedness of system elements.

V. Conclusion

Accumulated experience in information systems design can be formalized in form of knowledge database that will allow automating routine operations related to designing. Generation of architecture elements and realization of system as well as documentation and interface, allows designers to concentrate on solving application tasks. Semantic of functions relations with business-processes and subject area fragments can be used to solve task of developing information system services. Clear indication of such semantic in form of ontology allows automating process of system functions clustering into services.

References

[1] C. Pahl, "Semantic model-driven architecting of service-based software systems,"Information and Software Technology, vol. 49, pp. 838-850, 2007.

[2] J. Davies, D. Faitelson, and J. Welch, "Domain-specific Semantics and Data Refinement of Object Models,"Electronic Notes in Theoretical Computer Science, vol. 195, pp. 151-170, 2008.

[3] Yu. A. Orlova, "Analysis of models and methods of increasing effeciency of software design,"IZVESTIYA VolgGTU, vol. 11, no. 9, pp. 137-141, 2010.

[4] D. Fogli and L. P. Provenza, "A meta-design approach to the development of e-government services,"Journal of Visual Language and Computing, vol. 23, no. 2, pp. 47-62, 2012.

[5] V. Anaya, G. Berio, M. Harzallah, P. Heymans, A. L. Opdahl, and M. Jose, "The Unified Enterprise Modelling Language âĂŤ Overview and further work,"Computers in Industry, vol. 61, pp. 99-111, 2010.

[6] C. Lopez, V. Codocedo, H. Astudillo, and L. Marcio, "Bridging the gap between software architecture rationale formalisms and actual architecture documents: An ontology-driven approach,"Science of Computer Programming, vol. 77, no. 1, pp. 66-80, 2012.

[7] K. Robles, A. Fraga, J. Morato, and J. Llorens, "Towards an ontology-based retrieval of UML Class Diagrams,"Information and Software Technology, vol. 54, no. 1, pp. 72-86, 2012.

[8] R. C. De Boer and H. Van Vliet, "Architectural knowledge discovery with latent semantic analysisâĂŻ: Constructing a reading guide for software product audits,"The Journal of Systems & Software, vol. 81, pp. 1456-1469, 2008.

[9] Y. Ma, "Dynamic evolutions based on ontologies,"Knowledge-Based Systems, vol. 20, pp. 98-109, 2007.

[10] R. Valencia-garci, R. Marti, and F. Garci, "An ontology , intelligent agent-based framework for the provision of semantic web services,"Expert Systems with Applications, vol. 36, pp. 3167-3187, 2009.

[11] C. Zanni-merk and D. Cavallucci, "Use of formal ontologies as a foundation for inventive design studies,"Computers in Industry, vol. 62, pp. 323-336, 2011.

[12] L. Rao, G. Mansingh, and K. Osei-bryson, "Building ontology based knowledge maps to assist business process re-engineering,"Decision Support Systems, vol. 52, no. 3, pp. 577-589, 2012.

[13] D. Strasunskas and S. E. Hakkarainen, "Domain model-driven software engineeringâĂŻ: A method for discovery of dependency links,"Information and Software Technology, vol. 54, no. 11, pp. 1239-1249, 2012.

[14] T. Massoni, "A Framework for Establishing Formal Conformance between Object Models and Object-Oriented Programs,"Electronic Notes in Theoretical Computer Science, vol. 195, pp. 189-209, 2008.

[15] L. Thiry and B. Thirion, "Functional metamodels for systems and software,"The Journal of Systems & Software, vol. 82, no. 7, pp. 1125-1136, 2009.

[16] N. Bolloju, C. Schneider, and V. Sugumaran, "A knowledge-based system for improving the consistency between object models and use case narratives,"Expert Systems With Applications, vol. 39, no. 10, pp. 9398-9410, 2012.

[17] B. Henderson-Seller, "Bridging metamodels and ontologies in software engineering,"The Journal of Systems & Software, vol. 84, no. 2, pp. 301-313, 2011.

[18] A. De Nicola, M. Missikoff, and R. Navigli, "A software engineering approach to ontology building,"Information Systems, vol. 34, pp. 258-275, 2009.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 273–280
DOI: 10.15439/2014F106
ACSIS, Vol. 2

Learning History with Timelines: Use Cases, Requirements and Design

Evgeny Pyshkin
Institute of Computing and Control
St. Petersburg State
Polytechnical University
St. Petersburg, Russia, 195251
Email: pyshkin@icc.spbstu.ru

Nikita Bogdanov
Institute of Computing and Control
St. Petersburg State
Polytechnical University
St. Petersburg, Russia, 195251
Email: nik.see.7@gmail.com

***Abstract*—This paper is focused on using computer timeline based interfaces in the domain of history learning. The paper explains what kind of problems do historians and learners face when they deal with chronologically ordered historical information. We review existing approaches and software tools that support timelines. Based on the timeline metaphor ontological model, we examine major features of existing software tools and introduce some novel elements that might be considered as requirements for further implementations.**

I. Introduction

AWARENESS technology and pervasive nature of current software transform the university curriculum and computer-assisted learning and teaching environments dramatically. Today software has a strong impact both on technology sensitive disciplines and on liberal arts [1]. Specifically, for the domain of computer-assisted language learning (CALL), Beatty mentioned that current CALL is *"an amorphous or unstructured discipline, constantly evolving both in terms of pedagogy and technological advances in hardware and software"* [2]. Vice versa, the latter observation can be applied not only to language learning but to various areas of technology-driven education as well.

Let's consider history learning. One definition of history is *"a chronological record of significant events (as affecting a nation or institution) often including an explanation of their causes"* [3]. Hence, above all others, learning history means learning events, their causes and dependencies. Chronology tables, or timeline charts, are used traditionally in numerous history monographs, biography books, science reviews, and so on. As the number of events and contexts increases, one feels hard to manage all the related information. Timelines created by using special software tools don't simply provide a way to record and store event-related information by using computer databases (and this is kind of computer assistance too). Furthermore, they are designed to provide a specific interface to deal with chronology information in a way that is fast impossible or hardly implementable without computers. Existing solutions include such features as timeline zooming, group editing, nesting timelines, attaching multimedia information, managing references to geography maps, association with information about history artifacts, 3-d visualization, to cite a few. This albeit incomplete list illustrates a real technology-driven transformation of the *active learning* space achieved with help of the computer and software technology of the day.

The focus of this work is to examine problems that historians and history lovers face when they deal with chronologically ordered information. We try to analyze the timeline metaphor from the ontological perspective in order to discover concepts and features which aren't supported by existing timeline visual interfaces and software solutions.

The remaining text of the paper is organized as follows. In section II we analyze the timeline metaphor in the most common sense. Section III lists existing tools using the timeline metaphor to model time related data and processes. We review basic user interface layout types and major stereotypes and features implemented in selected software tools. In section IV we pay special attention to several approaches used as a kind of formal foundation in timeline processing software. In section V the timeline ontology is introduced with deeper analysis of modeling timeline and event associations. We explain main elements of the map of timeline related concepts and examine requirements for novel features that are missed in the existing applications including event associations, alternate time scaling with respect to different chronology styles, and regional zooming. In section VI we describe user interface elements supporting timeline and event associations and discuss their implementation in the prototype application.

II. A Timeline metaphor

Time is an immanent attribute of information. Whatever we have as a subject context (e.g. history, literature, music, computer programs, linguistics, etc.), events, concepts, documents, art and engineering artifacts appear and develop in time. Even text semantics and word relatedness often change in time. Attention to temporal information attributes and to possible changes of word usage over time can affect the degree of semantic relatedness [4]. Thus, capturing time related information is an important and complex issue in semantic information retrieval.

In many engineering areas (including software design), visualization is one of known ways to decrease system com-

978-83-60810-58-3/$25.00

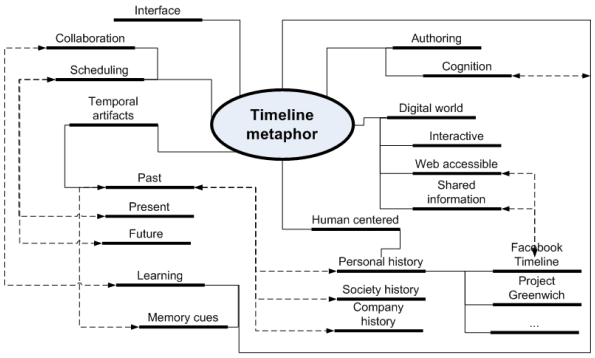


Fig. 1. Timeline metaphor as a map of concepts

plexity and to improve system perception by learners and collaborators. Now let's go back to the history learning issues. Timelines is one of visual ways to organize, manage and display information about events. Wikipedia defines a timeline as *"a way of displaying a list of events in chronological order, sometimes described as a project artifact"*. There might be different kind of events: historical events, cultural events, project milestones or tasks, scheduled appointments, transportation time marks, etymology study points or even research paper publication dates. Thus, on the one hand, timelines are really multidisciplinary visual formalism. On the other hand, it is one of ways to support systematic learning of history related disciplines (like history itself, history of the art, natural science, astronomy, and others).

Despite timelines are usually linked to a graphic design with using bars labelled with dates and associated with event descriptions, Wikipedia timelines (however strange it might be) are presented mostly in textual or table based forms. This makes difficult to zoom events representation to make more emphasis on more important events, to recognize connections between different timelines, to visualize difference of events density in different periods of history, or to highlight different contexts related to the same timeline. Figure 1 shows the sketch of the timeline metaphor in form of a brainstorming map of concepts.

Timelines are more than an attractive and intuitive way to visualize historical information. Project management and collaborative work organization is currently a vast area of using a metaphor of the timeline. It allows combining the broader project view of planned and completed activities with coordinated end user oriented facilities focused upon the specific tasks and events related to the team members [5], [6].

More data people tend to collect, more interests grow in finding out new models to organize information retrieved from the data. Timelines became one of human-computer ubiquitous interfaces used for recording, structuring and browsing personal history [7]. Furthermore, a metaphor of the timeline provides a special searching interface which might be considered as an obvious implementations of the concept of temporally-oriented non-navigational searching: users are open

TABLE I
TIMELINE IMPLEMENTATIONS (PART I)

Name	*Main focus*	*Description*
SmartDraw[a]	Workflow, business processes	Commercial business chart software supporting flowcharts, decision trees, cause and effect diagrams, timelines, etc. Timelines are focused on modeling company workflow, office workers management, information flow, reporting and documentation.
Asana[b]	Team and project management	Commercial packages for teamwork organization based on task and responsibility centered model of the project process flow.
Timelime[c]	Workflow, planning	Supports creating timelines for daily activities, team works, calendar related events and task grouping.
Matchware Education[d]	Mind mapping, project management	Timeline component is a part of the MindView application. Support fixed selection of time periods scaling (standard, daily, weekly, historic and geological). Supports look and fell interface.

[a] http://www.smartdraw.com/examples/timelines
[b] http://www.asana.com
[c] http://timelimeapp.com
[d] http://www.matchware.com/en/products/mindview/education/timelines.htm

for suggestions since they might have no clear preliminary understanding what document they are trying to locate [8]. What is more, people often prefer browsing over direct search even if the search target is known: *"the interface intelligently emphasizes potentially relevant items on a timeline, so that an item can easily be recognized and selected for further inspection"* [9].

III. STATE OF THE ART

Let us introduce existing tools that use the timeline metaphor for the needs of history learning and explain their focus, current features, and layout models they are based on.

A. Tools

Table II cites examples of solutions that interest us at most since they can be used as elements of computer-assisted history learning space. Note that a British Library project seems to be a little bit apart of other history chronology based solutions since it represents an approach to expose the *museum collection artifacts* by using the timeline view, without direct relation to the historical event-based context-dependent editable timelines. Table I lists examples of timeline tools used in business oriented applications like project scheduling, staff management, or task planning.

There are three basic types of timeline layouts used by existing implementations:

- **Time centered layout** uses the event snippets positioned along the timeline bar with detailed descriptions and links to the external resources appeared as popup elements. The *Vistorica* web site is an example (see Figure 2). In the

TABLE II
TIMELINE IMPLEMENTATIONS (PARTII)

Name	*Main focus*	*Description*
Timeline Maker[e]	Presentations	The solution is integrated with PowerPoint and supports timeline diagramming for better presentations.
TimeGlider[f]	History, project planning	Text based representation with abilities to attach graphics. Support several configurable event categories and map tags.
TimelineJS[g]	History, education, personal planning	Web based application for automatic timeline generation from Google spreadsheets input data. Support event layout by using multi-row timelines.
TimeToast[h]	History, education, personal history	Web based application supporting simple timeline construction able to be shared share and categorized by using predefined categories (music, film, science and technology, business, politics, biography, art and culture, personal, history). Popular timelines are listed on the web site.
Tiki-toki[i]	History, education, presentations, personal history	Web based application for creating timelines in very visual way (which include 3d-visualization).
Vistorica[j]	History, education	Historical persons centered application. Names and events are managed by using a set of predefined contexts (events in Europe, science and technology, European works, mathematics and engineering, humanities, culture, economy, politics, military). Integration with geography information is supported.
Chronozoom[k]	History, education	Huge open API and history learning experimental platform for managing user-defined timelines with rich GUI supporting zooming, creating nested and shared timelines linked to a big variety of media resources. Education perspective is one of dominating reason.
British Library[l]	Education, featuring museum collections	The application uses timelines to support exploring the library artifact collection represented collection items chronologically ranging from medieval times to the present day.

[e] http://www.timelinemaker.com/
[f] http://timeglider.com
[g] http://timeline.knightlab.com
[h] http://www.timetoast.com
[i] http://www.tiki=toki.com
[j] http://vistorica.com
[k] http://www.chronozoom.com
[l] http://www.bl.uk/learning/histcitizen/timeline/acceessvers/index.html

Vistorica there is also a special area to represent event related geographic locations.

- **Topic centered layout** shifts focus to the event detailed description while timeline bar serves as a navigation interface. The *TimelineJS* illustrates an idea (see Figure 3).
- **Container model** shown in Figure 4 is introduced in the *Chronozoom* project where a timeline is considered as a container of exhibits with one or more artifacts illustrating the event related topic.

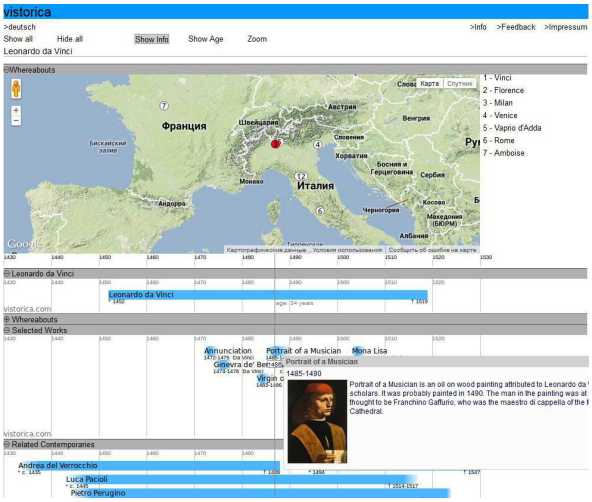


Fig. 2. Time centered layout in *Vistorica*

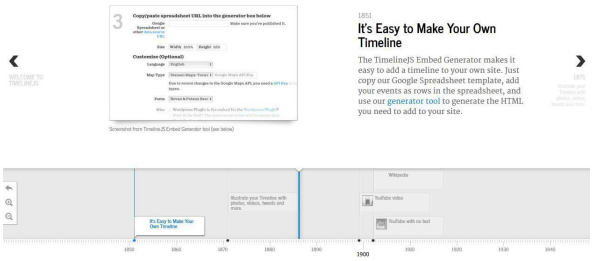


Fig. 3. Topic centered layout in *TimelineJS*

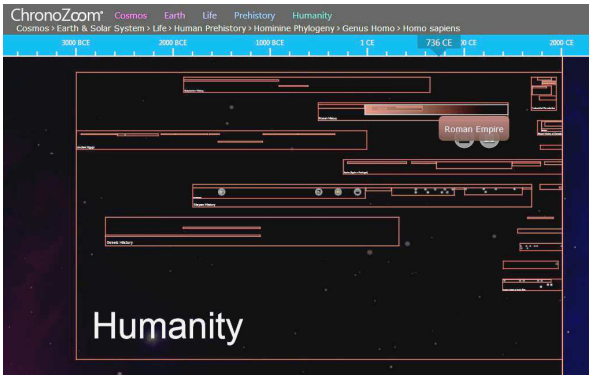


Fig. 4. *Chronozoom*'s container model

B. Features

Table III lists major features that general purpose timeline tools support. Through learning and educational perspectives, Chronozoom seems to be potentially the most powerful open-source platform for history visualization [10].

Despite many interesting features and layout concepts can be found in existing implementations, there is a space for further improvements with respect to human centered and history learning centered views. In the following sections we propose some feature concepts aimed to improve requirements for further implementations.

IV. Related Work

There is a variety of research works on generating document summaries to be included into the list of retrieved documents while searching [11], [12]. The main idea is to evaluate the semantic meaning of the document paragraphs and choose the paragraphs with the best meanings for the summary. In the area of timeline visualization, there are aspects of constructing relevant summaries for using them as timeline event headlines. In the work [13] the authors investigate possibilities to construct timeline summaries from collections of news articles available on the web. The main difficulty concerns the problem of choosing the most meaningful news which relate to the event date: the relevant information can be found not only in news appeared exactly on some certain date. They may be published *after* or *before* the analyzed date and have references to this date. The complexity of this task grows if we take into account huge amount of news reports received from numerous websites. Users are often interested only to catch the main idea of the news from such a flooding news stream and to discover how do the news appeared on different dates relate to each other. News summarization in the form of timelines may help to reach this goal. Xu et al. introduced a cross-media evolutionary summarization approach which states the formal model that helps deciding whether the news (texts and images) obtained from a certain media or posted by a certain author is the best candidate to be included into the final timeline [14]. They apply an idea of collecting recommendations to the domain of news selection: each candidate news can *"recommend"* the others and in turn can be "recommended". The candidate's authority increases if s/he gets more positive recommendations. Furthermore, recommendations from a more authoritative candidate make other candidates' authority increasing (thus, in a certain sense it is similar to the model used to evaluate research work impact and productivity by using the H-index). This approach can be used as a foundation to automate the process of selecting authoritative and reliable event descriptions to be included into the timeline.

V. Timeline Concept

As a rule, an event has temporal, spatial and schematic attributes associated with it. That's why investigating relationships, dependencies and correlations between events organized in the chronological order is of much importance [15].

In computer assisted learning systems, the content dependent facilities are necessary in order to express better event semantics and associations. Jouault and Seta cite an example of some military battle description [16]. In addition to timeline positioned events like the battle, the armistice, the treatment

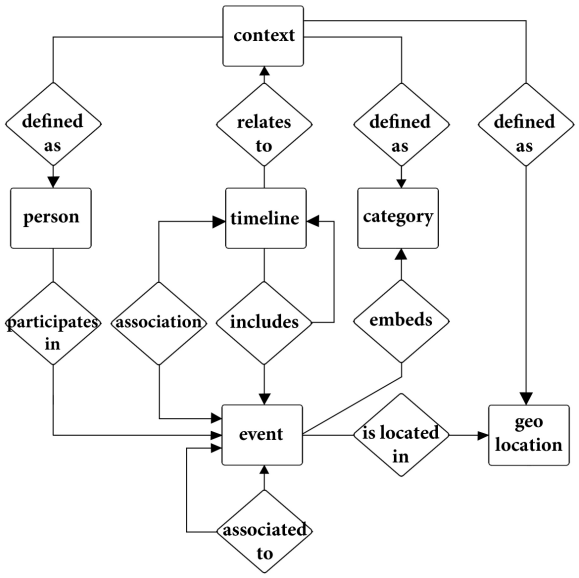


Fig. 5. Core of the timeline ontology

and so on, there are things which are not events but which should be represented in coordination with the timeline (combatant armies, countries that signed a treaty, persons involved to the events, etc.). Every such artifact may have a link to the related timeline. Thus, the associations are as much important as the events.

Here is the list of main timeline ontology entities shown in Figure 5:

- **Event (historical event, artifact)**. Event description is usually represented by the headline summary and associated with date or date range information as well as with links to related resources.
- **Category**. A category represents the event or timeline specific domain (e.g. history, architecture, countries, music, arts, etc.).
- **Timeline context**. A context may have similar sense as an event category, but may also relate to geographic locations, persons or other user defines contexts.
- **Geographical location**. It includes data about geographic coordinates, links to economy, political or cultural contexts bound to geography.
- **Person**. A person indication may be important due to his or her participation in some event. A person may define a timeline context too.
- **Event associations**. Associations are ontology elements that are poorly supported in software tools surveyed in Section III, except the implicitly existing temporal associations. However within the framework of history learning and analysis, associations are very important properties.

When teaching history, an instructor might require shifting learner's focus from the timelines or events themselves to the associations, for example:

TABLE III
SUPPORT FOR MAJOR FEATURES IN EXISTING GENERAL PURPOSE TIMELINE IMPLEMENTATIONS

Features	*Major implementations*						
	Timeline Maker	*TimeGlider*	*TimelineJS*	*TimeToast*	*Tiki-Toki*	*Vistorica*	*ChronoZoom*
Zoom	+	+	+	–	+	+	+
Event headlines	+	+	+	+	+	+	+
Event details	+	+	+	+	+	+	+
Event importance	–	+	–	–	–	–	+
Tags	–	+	–	+	+	–	–
Images	+	+	+	+	+	+	+
Multimedia	–	–	+	+	+	–	+
External links	+	+	+	+	+	+	+
Mapping to geography	–	–	+	–	–	+	Partially
Multiple categories	–	–	–	+	+	+	+
User defined categories	via event category	–	+	–	+	–	+
Comparing timelines	–	–	–	–	–	+	+
Nested timelines	–	–	–	–	–	–	+
User defined timelines	+	+	+	+	+	–	+
Web access	–	+	+	+	+	+	+
Sharing	Partially	+	+	+	+	–	+
Group editing	Partially	–	–	–	+	–	+
Multiple rows	+	+	+	+	+	+	+
Look and feel	+	–	–	–	+	–	+
Multiple views	Table	–	+	Table	+	–	–
3D-view	–	–	–	–	+	–	–
Search	+	+	–	–	+	–	+
Presentation centered	+	–	+	–	+	–	–
Information centered	–	+	+	+	+	+	+
Integration	PowerPoint	–	–	–	–	–	–
Open API	–	–	JSON	–	JSON	–	+
Embedding	MS Office	Web	Web	Web	Web	–	–
Import Data	+	CSV/JSON	Google spreadsheet	–	Partially from YouTube/RSS	–	Partially
Export Data	Text	CSV/JSON	–	–	CSV/PDF	–	Partially

- How (and why) two (or more) events are related to each other;
- For what reasons an event can be considered as a cause or a consequence of another one;
- How timelines are related, or how events do affect certain timelines;
- What is the similarity between sets of events in different timelines.

Figure 6 introduces possible associations to be considered as a part of the user interface. We borrowed the *Chronozoom*'s concept of a timeline container and added some supplementary constructs to the picture.

The task of comparing *industrial revolution* periods in Russian and Japanese history serves us as an example. When a teacher (or a learner) considers the industrial transformations in Japan after the long era of Tokugawa family dominance, the events that occurred in Japanese industry, its transportation system, or naval building couldn't be analyzed without paying attention to the *Meiji reconstruction* period, so these two timelines are deeply related. For the case of Russia we can cite an example of *constructivist* trends in architecture which are in strong relation with *industrialization* processes. From the other point of view, the constructivism timeline is affected by the certain event from *Kazimir Malevich*'s biography: Malevich is generally thought as an inventor of the term *"constructivism"*, so there is a dependency between the event and the timeline. Some events may be related despite they are relatively distant: the Russian cruiser *"Aurora"* was a battleship participated in the *Tsushima battle* and later won renown with the blank shot that symbolically started the *October revolution*.

The Figure 6 highlights the issue of implicit conversions that humans often do. An entity considered as an event in one timeline context (e.g. the *Tsushima battle* within the boundaries of *1904-05 Russo-Japanese war* timeline) may be transformed to a timeline if we analyze the battle in details. Unfortunately, there is no direct interface feature that takes this aspect into account.

The next aspect we would like to mention is timeline scales. In all the implementations that we surveyed, the only used time scale refer to traditional (western christian) chronology. However, other traditions to deal with chronology exist and often required to represent timelines in better correspondence to national and cultural contexts and to individual preferences of a historian, to cite a few:

- Russian orthodox church calendar;
- Japanese chronology which refers to their own history pe-

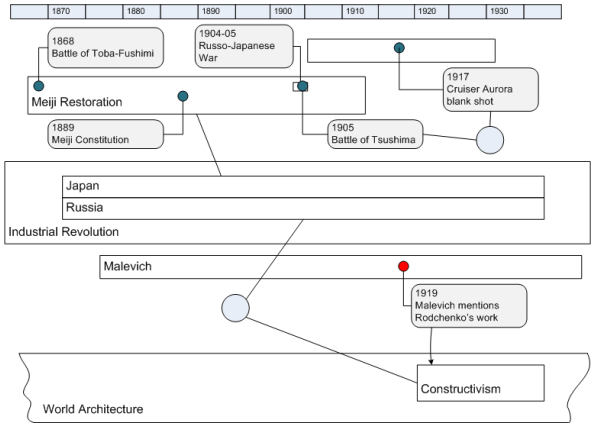


Fig. 6. Artifact associations

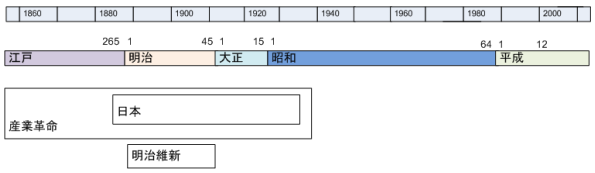


Fig. 7. Example of Japanese chronology timeline bar

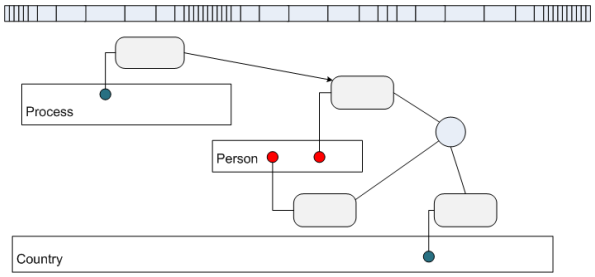


Fig. 8. Flexible time bar and distant associated events

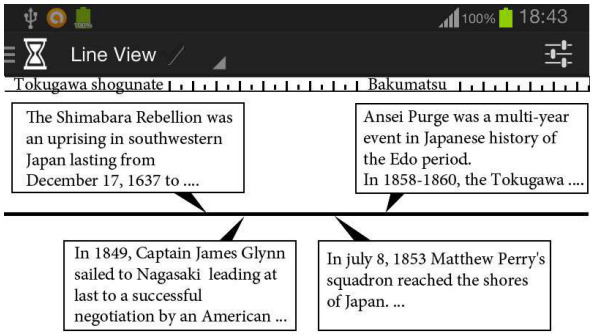


Fig. 9. Headlines and a timeline bar focused view

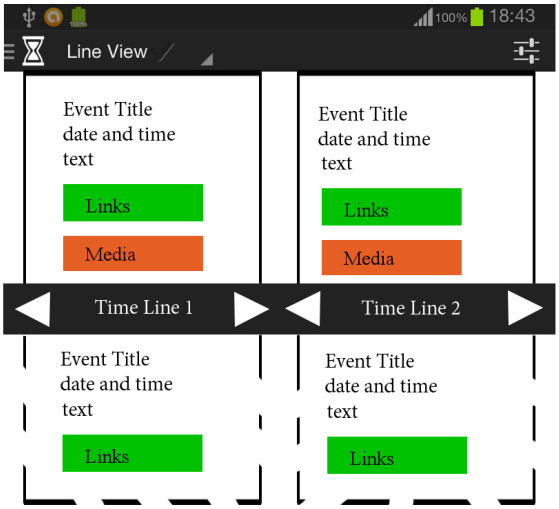


Fig. 10. Descriptions and comparing focused views

riodization like Jōmon, Yayoi, Kofun and to the Japanese era calendar scheme (nengō) referred to the Imperial practice;

- Chinese Han calendar;
- Non-traditional chronologies.

For certain timelines, historians might require possibility to use parallel subchronologies like for a case of the French Republican Calendar used by the French government for about 12 years from late 1793 to 1805 during the French Revolution. Therefore, an alternate timeline might inherit the core chronology but use parallel time scales as Figure 7 illustrates for the selected events represented earlier in Figure 6.

There are speculative feature concepts that are subjects of further discussions. While learning history people sometimes tend to analyse relationships between very distant events. In this case the possibility can be useful to zoom in or out only some selected time regions allowing focusing on event relationships rather than on the chronology. Figure 8 illustrates this issue.

At last, historians often learn parallels between events in quite different time periods (for example, comparing French and Russian Revolutions). In this case a *"shift and fix"* concept can probably serve the idea: virtually moving one timeline over time with fixing another "in place". It seems that digital timelines can assist such kind of history analysis quite easy and therefore provide a space for a process similar to natural science experiments.

VI. Android Prototype Application

A. Interfaces

We started implementing timeline views described in the previous section while developing an Android prototype application. There are two traditional views. The view focused on event headlines allows showing events related to one or more categories attached to the time bar. The view focused on event descriptions allows managing detailed event information and comparing detailed descriptions related to the selected timelines as Figure 9 and Figure 10 illustrate.

In addition to traditional views we design the interfaces for map and grid views (shown in Figure 11) which are timeline

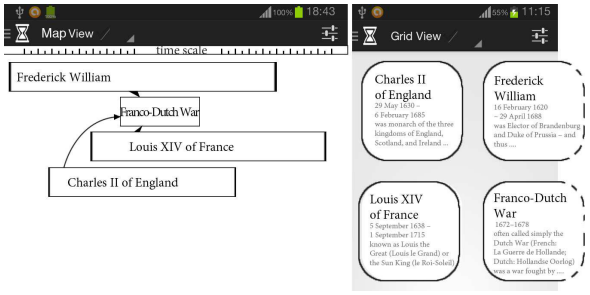


Fig. 11. Map and grid timeline views

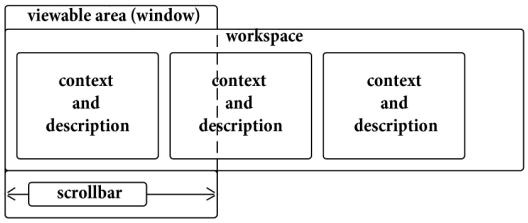


Fig. 12. Sliding window model

association centered. The map view (inspired by a concept of a semantic network) highlights possible relationships between timelines. The map view is aimed to facilitate navigation between related timelines. Evidently, a number of associated timelines can also be considered as a context. The grid view inherits a general concept of a sliding window (see Figure 12) and serves as a kind of an organizational view. It supports timeline arrangements from the user's point of view (similar to application icons) with no direct implications to the timeline relations. For portable devices it is essential to have switching between different views.

B. Database

Currently the application uses its own database aimed to store two types of events. The first group includes events created manually by the user. The second one corresponds to the events grabbed from the Wikipedia used as a history knowledge source. The database structure shown in Figure 13 is developed according to the timeline ontology model presented earlier.

Below the selected database tables are listed with brief descriptions.

All_Keywords table contains keywords used for searching within the *Author*, *Category*, *Geo*, *Person* and *TimeLine* tables.

Synonym table contains keyword synonyms with respect to the context of tables where the data are to be searched. The latter option is necessary since two keywords may be synonyms within the context of one table while keeping independence for another table.

Author_Keyword, *Category_Keyword* and *Person_Keyword* are auxiliary tables used for searching.

Author corresponds to the event description origins: who did create the event description and what was the source.

The *Timeline*, *Context*, *Category*, *Persons*, *Event*, *Association* and *Geo* tables relate to the respective timeline ontological entities. Specifically, the *Geo* table is connected with other tables (*Continent*, *Country*, *Nationality*, *City*, etc.) representing different aspects related to the geographical location, political or social associations.

The *Person_geo* table contains information about geographical locations related to a specific person.

The *Event_and_Person* table connects events with persons involved.

The *Links* table contains links to the external resources related to the event.

Despite the database structure may change in the future revisions, it is nevertheless useful to explain better timeline related entities and their associations.

C. Future work

Except implementing modules depending on designed interfaces, further steps have to be taken to communicate with external services like *Chronozoom* by using their open API. The idea is to consider portable application as an interface which would fit better the requirements and stereotypes of a mobile device user. Another challenging problem (which refers strongly to teh domain of semantic information retrieval) is how to extract relevant information about event associations automatically.

VII. Conclusion

Increasing interests to timeline modeling can serve as an example of how new tasks and user interfaces appear as consequence of computing and web technology and applications development. People create timeline based visual interfaces and visual representations of temporal data to improve knowledge acquisition. Information retrieval algorithms and related information processing techniques are not only about accessing some content rapidly and precisely but also about enabling better human or society understanding of explored phenomena, their relations and their mutual dependency with other phenomena and artifacts.

Historians can use timeline based tools while learning, researching or teaching historical periods discoverable with respect to different national and cultural contexts. Indeed, there is a good reason explaining why do many books on history contain chronology tables. Similar to natural sciences where setting up an experiment is a usual way of study, timelines created with computer tools make arranging learning experiments for history education easier. We may collate similar periods in different cultures, or, conversely, analyze comparable events that took place in different epochs and in different places. We are able even to model prefigured events and historical hypotheses.

In so doing, developers are able create a better framework implementing a concept of active learning in tight cooperation with recent achievements of information retrieval methods and software technology.

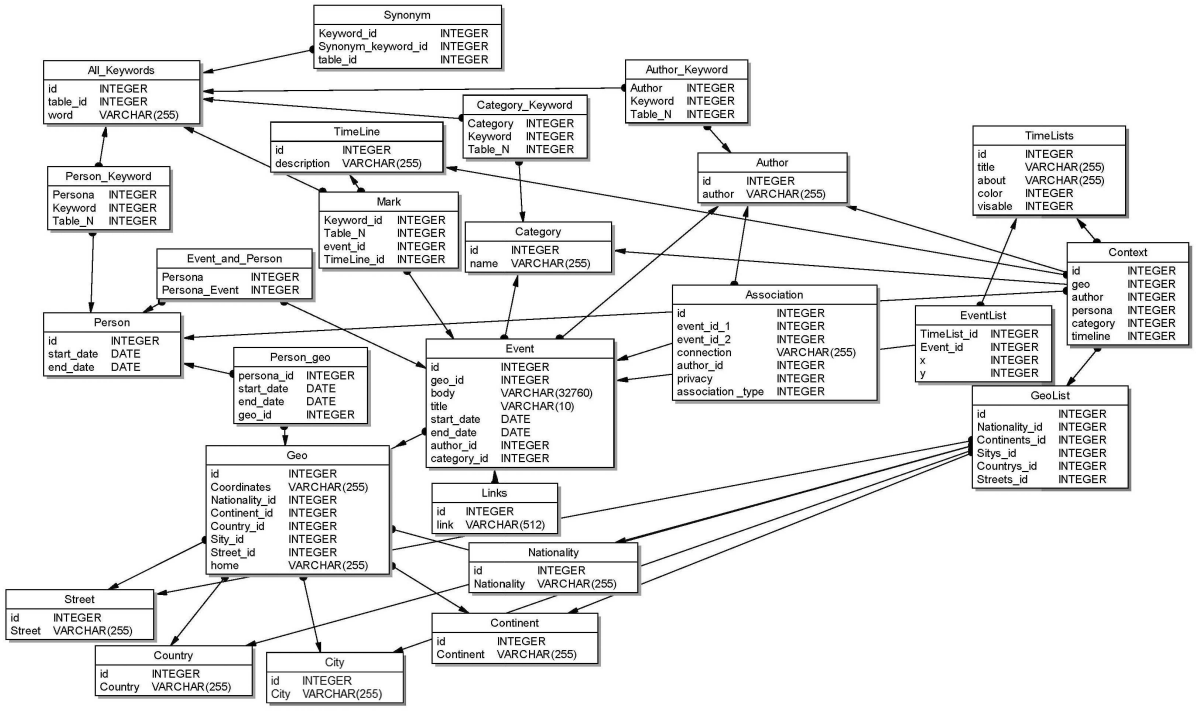


Fig. 13. Database structure

REFERENCES

[1] E. Pyshkin, "In the right order of brush strokes: a sketch of a software philosophy retrospective," *SpringerPlus*, no. 186, p. 3:186, 2014.

[2] K. Beatty, *Teaching and Researching Computer-assisted Language Learning. 2nd edition*. Pearson Education, 2010.

[3] "Merriam-webster dictionary: History." [Online]. Available: http://www.merriam-webster.com/dictionary/history

[4] K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch, "A word at a time: Computing word relatedness using temporal semantic analysis," in *Proceedings of the 20th International Conference on World Wide Web*, ser. WWW '11. New York, NY, USA: ACM, 2011, pp. 337–346. [Online]. Available: http://doi.acm.org/10.1145/1963405.1963455

[5] M. Bohøj, N. G. Borchorst, N. O. Bouvin, S. Bødker, and P.-O. Zander, "Timeline collaboration," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10. New York, NY, USA: ACM, 2010, pp. 523–532. [Online]. Available: http://doi.acm.org/10.1145/1753326.1753404

[6] L. Batrinca, M. T. Khan, D. Billman, B. Aydemir, and G. Convertino, "A timeline visualization for multi-team collaborative planning," in *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '13. New York, NY, USA: ACM, 2013, pp. 157–162. [Online]. Available: http://doi.acm.org/10.1145/2468356.2468385

[7] E. Thiry, S. Lindley, R. Banks, and T. Regan, "Authoring personal histories: Exploring the timeline as a framework for meaning making," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '13. New York, NY, USA: ACM, 2013, pp. 1619–1628. [Online]. Available: http://doi.acm.org/10.1145/2470654.2466215

[8] K. Church, "How many multiword expressions do people know?" in *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, ser. MWE '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 137–144. [Online]. Available: http://dl.acm.org/citation.cfm?id=2021121.2021152

[9] A. Ajanki, M. Koskela, J. Laaksonen, and S. Kaski, "Adaptive timeline interface to personal history data," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ser. ICMI '13. New York, NY, USA: ACM, 2013, pp. 229–236. [Online]. Available: http://doi.acm.org/10.1145/2522848.2522870

[10] R. L. Walter, S. Berezin, and A. Teredesai, "Chronozoom: Travel through time for education, exploration, and information technology research," in *Proceedings of the 2Nd Annual Conference on Research in Information Technology*, ser. RIIT '13. New York, NY, USA: ACM, 2013, pp. 31–36. [Online]. Available: http://doi.acm.org/10.1145/2512209.2512217

[11] V. Oleshchuk and V. Klyuev, "Context-aware summary generation for web pages," in *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 2009. IDAACS 2009. IEEE International Workshop on*, Sept 2009, pp. 561–564.

[12] S. Satake, S. Ishikawa, D. Vazhenin, and V. Klyuev, "A semantic-based summarizing tool for improving the usability of general purpose search engines," in *Pervasive Computing (JCPC), 2009 Joint Conferences on*, Dec 2009, pp. 235–238.

[13] G. Binh Tran, M. Alrifai, and D. Quoc Nguyen, "Predicting relevant news events for timeline summaries," in *Proceedings of the 22Nd International Conference on World Wide Web Companion*, ser. WWW '13 Companion. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013, pp. 91–92. [Online]. Available: http://dl.acm.org/citation.cfm?id=2487788.2487829

[14] S. Xu, L. Kong, and Y. Zhang, "A cross-media evolutionary timeline generation framework based on iterative recommendation," in *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval*, ser. ICMR '13. New York, NY, USA: ACM, 2013, pp. 73–80. [Online]. Available: http://doi.acm.org/10.1145/2461466.2461480

[15] A. Biswas, B. V. Sagar, and J. Srinivasan, "Managing and correlating historical events using an event timeline datatype," in *Proceedings of the 13th International Conference on Database Systems for Advanced Applications*, ser. DASFAA'08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 604–612. [Online]. Available: http://dl.acm.org/citation.cfm?id=1802514.1802580

[16] C. Jouault and K. Seta, "Building a semantic open learning space with adaptive question generation support," in *Proceedings of the 21st International Conference on Computers in Education*, ser. ICCE'13. Asia-Pacific Society for Computers in Education, 2011, pp. 41–50. [Online]. Available: http://icce2013bali.org/datacenter/mainconferenceproceedingsforindividualdownload/c1/C1-f-206.pdf

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 281–288

DOI: 10.15439/2014F323
ACSIS, Vol. 2

LELA - A natural language processing system for Romanian tourism

Bernadette Varga*, Alina Dia Trambitas-Miron*, Andrei Roth*,
Anca Marginean†, Radu Razvan Slavescu†, Adrian Groza†
*Semantic Web Department, Recognos Romania
{bernadette.varga, dia.miron, andrei.roth}@recognos.ro
†Intelligent Systems Group
Department of Computer Science, Technical University of Cluj-Napoca
{Anca.Marginean, Radu.Razvan.Slavescu,Adrian.Groza}@cs.utcluj.ro

***Abstract*—This paper presents a commercial semantic-based system for the Romanian tourism. The Lela system exploits both open linked data from Romanian and international sources, and also proprietary databases in the tourism domain. We present the process of creating the linked data set, based on: i) engineering the LELA Romanian tourism ontology, and ii) populating the ontology by linking open data. The system also provides a natural language interface for the Romanian language. The queries are automatically translated into SPARQL based on a controlled vocabulary derived from the Lela ontology.**

***Index Terms*—Semantic information retrieval, Query interfaces, Natural language processing, Linked Data, Tourism ontology**

I. Introduction

LELA is an intelligent blogging-platform designed for providing personalized information about Romanian touristic places. The user can query both subjective and objective information about places of interest. This is possible because Lela uses a custom made semantic annotation tool for blog posts, that identifies points of interest(POIs) and extracts their features and the sentiments expressed about them. The extracted data is used to annotate posts thus allowing their semantic indexing. Lela also provides a Natural Language Question Answering mechanism that allows users to express queries in Romanian language.

II. System architecture

The Lela system relies on the Lela ontology that we engineered for the Romanian touristic domain. The ontology is automatically populated using two methods: i) linking structured data from various sources in the touristic domain, and ii) using natural language processing of available touristic blogs. In the architecture of the system (figure 1) the *Data Collector* module is responsible for the first task, while the *Data Extractor* structures information from blogs in Romanian language. The *Question Answering* module handles queries in natural language against the assertions in the Lela ontology.

The *Data Collector* module identifies and imports relevant information related to Romanian points of interests by linking touristic information from open data provided by the Romanian agencies, complemented with relevant knowledge from Wikipedia, DBpedia, or Freebase. Data is collected using

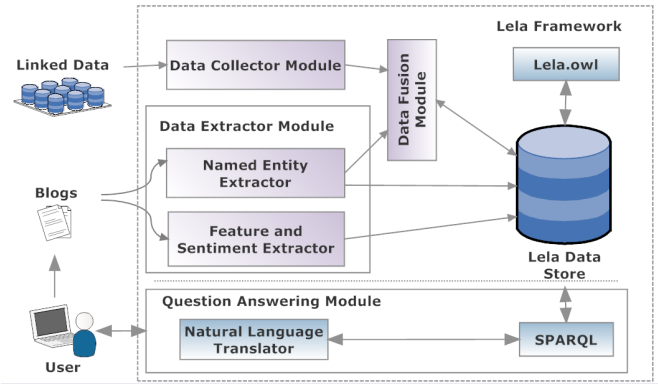


Fig. 1. LELA system architecture.

available SPARQL endpoints or source specific APIs. As information related to the same POI is usually available in more than one data source, we have developed a customized *Data Fusion* module, aiming to identify individuals described in different data sets. Based on information such as geospatial location and textual description, the equivalent individuals are linked via the *same-as* relationship and the various information asserted is fused.

Lela framework provides also a blogging platform where Romanian bloggers can write their stories about the places they have visited. In order to respond to the specific needs of bloggers that already built a readers community and an online reputation via their own blogs built on platforms such as WordPress, Drupal, Blogger, Tumblr, *etc.*, we will provide custom plug-ins for each one of the systems mentioned above, that will allow the semantic indexing of blog posts. Even though the content published on those blogs will not be copied on our platform, it will be available for Lela users to query and explore. This way we will acquire subjective information (stories, opinions) related to POIs.

The *Data Extractor* module analyses the available blogs in order to: i) perform Name Entity Recognition for the main concepts in the Lela ontology: points of interest, accommodation, restaurants or touristic activities; ii) assert different relationships between individuals in the ontology, as they appear in

the blogs, and iii) identify sentiments expressed in relation to each feature of a specific concept in the ontology. For instance, for the `Accommodation` concept, we are interested in opinions regarding features like `location`, `view`, `comfort`, `furniture`, `service` or `value for money`. The facts corresponding to these features (e.g., `Hotel X` serves `good food`) are stored into the ABox using the `Match` concept, which references the blog post that was analyzed, and the position inside the text for the POIs and opinions that were identified(”matched”) by the system.

For opinion detection, we used a machine learning technique based on Conditional Random Fields (CRFs) [1]. We employ this technique in order to find an appropriate labeling for blog sentences, regarded as sequences of words. The labels we try to detect describe the position in a sentence of a word which refers a specific instance of a concept (e.g. ”Grand Hotel Italia”), a specific feature we are interested in (e.g. ”cazare (accommodation)”) and the associated opinion (e.g. ”bun (good)”). For opinion polarity, we used the WordNet-Affect for Romanian [2].

The module which labels the text uses a model generated in the training phase, starting from a set of 200 manually labeled phrases. This set is further expanded by replacing some words of interest (especially the opinion adjectives like ”good”) with their synonyms, thus obtaining more training examples. A set of attributes has been selected for describing each word in the training set, and among them are the word’s Part-Of-Speech, whether the word belongs to an entity of interest and the type of entity. For example, let us consider the sentence ”Am fost la Transilvania International Film Festival si mi-a placut” (I was at the Transilvania International Film Festival and liked it). The attributes associated with the word ”Transilvania” will have the following values: ”NNP” for the Part-Of-Speech, ”B” for the attribute which specifies the name of the entity starts here and ”EVB” for the attribute specifying the word starts the name of an event.

The model generated based on the training example could be improved by expanding the set of examples and/or attributes. A separate module allows adding new attributes (”features” in CRF terminology) and computing their corresponding values before generating the new model. Once this is done, the new model is used to detect the entities the text is talking about, their specific features and the opinion on them. The opinion information gets stored in the A-Box as explained above and can refer either a feature of a concept instance or a pair Activity-Location (e.g., ”skiing” in ”Predeal”).

The opinions concerning each feature of a specific instance are aggregated into a quality score for that particular feature. The function which performs this takes into account both the detected opinion polarities (on a scale from -2/very bad to +2/very good) and the weights specified by the user for each feature s/he might be interested in, according to their importance from his/her point of view. When the discovered Named Entities are not recognized as Romanian POIs available within the Lela Data Store, they are added to the data store as new instances.

```
11. (define-role fromBlogPost :domain Match :range BlogPost)
12. (define-role hasSubject :domain Match :range LelaAxis)
13. (define-concrete-domain-attribute hasScore :domain Match
:type real)
14. (define-concrete-domain-attribute hasText :domain Match
:type string)
15. (define-role speaksAbout :domain BlogPost :range
LelaAxis)
16. (instance m1 Match)
17. (instance b100 BlogPost)
18. (instance mateicorvin POI)
19. (attribute-filler m1 "casa matei corvin atrage multi
turisti" hasText)
20. (related m1 b100 fromBlogPost)
21. (related m1 mateicorvin hasSubject)
22. (attribute-filler m1 0.8)
```

Fig. 3. Relating information about a blog with the *n-ary design* pattern.

III. Linked data creation process

The process of creating the Lela linked data set consists of three main steps: i) engineering a Romanian tourism ontology, ii) developing of data collection and data fusing modules, iii) publishing the resulting data sets.

A. Definition of a Romanian tourism ontology

To develop the Lela ontology, we follow the methodology in [3] and we also enact various ontology design patterns [4].

The later is described in KRSS syntax[1]. The four axes of the Lela-core ontology are `Accommodation`, `Activity`, `EatingAndDrinking` and `PointsOfInterest`, denoted by `POI` (line 1 in figure 2). Apart from those, Lela ontology also offers special classes for describing events, price, infrastructure, contact details, facilities of each point of interest, etc. The main properties defined in our ontology have restricted domains and ranges (figure 2 lines 3-6) which are used to facilitate reasoning among the top level concepts. The partition design pattern [7] was used to partition the top level of the ontology.

Beside the top level concepts, we also introduces the concept `Match` for representing the relations between the touristic places and the blog posts that POIs appeared in. This concept was modelled by enacting the *n-ary ontology design pattern* [7]. The goal was to combine several information about a tourism blog (see fig. 3) regarding: subject of the blog according to the concepts in Lela (line 12), computed score about an instance in the ontology (axiom 13), or provenance information like author, starting and ending text index (text position) which relates to an identified instance in our ontology. As an example, the individual `m1` of type `Match` is related to the blog `b100` via the role `fromBlogPost`.

The point of interest `mateicorvin` is related to the same match `m1` by the relation `hasSubject`. The positive score

[1]For a detailed explanation about families of description logics, the reader is referred to [5], while for the complete KRSS syntax to [6].

```
1. (define-concept LelaAxis (or Accommodation Activity EatingAndDrinking POI Location))
2. (disjoint Accommodation Activity EatingAndDrinking POI Location)
3. (define--role hasAccommodation :domain (or Activity EatingAndDrinking POI):range Accommodation)
4. (define-role hasActivity :domain (or Accommodation EatingAndDrinking POI) :range Activity)
5. (define-role hasEatDrink :domain (or Accommodation Activity POI) :range EatingAndDrinking)
6. (define-role hasPOI :domain (or Accommodation Activity EatingAndDrinking) :range POI)
7. (define-role hasLoc :domain LelaAxis :range Location :transitive t :inverse LocatedIn)
```

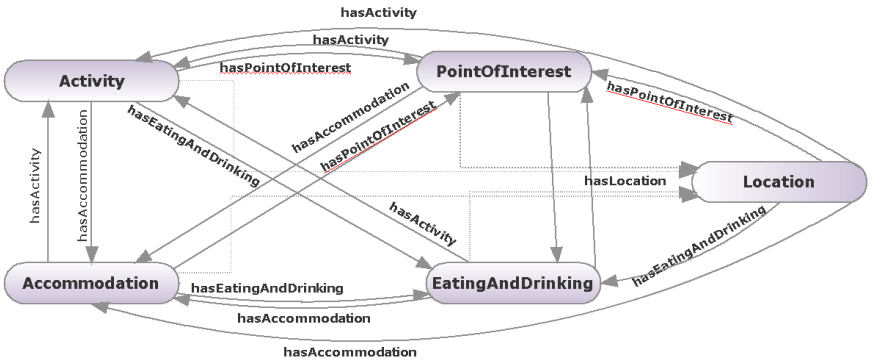


Fig. 2. Top level of the Lela-core ontology module.

of 0.8 in line 22 is computed with a basic opinion mining algorithm from the blog post.

B. Data collection and fusion

We focused on discovering and selecting the relevant open data sets for the Romanian tourism domain. Touristic points of interest are collected from tow sources: i) the available *POI data* sources (Wikipedia, Freebase, DBPedia, Geonames, Wikisherpa, Wikitravel) and ii) Named Entity recognition from touristic blogs. Data fusion is performed in AllegroGraph and saved as a triple store[2], while RacerPro server is used for reasoning on the Lela ontology.

The following sources were exploited: *Wikipedia*, *DBpedia*, *Geonames*, *Freebase*, *Wikisherpa*, *WikiTravel* (table I). A proprietary dataset from Cluj4All (www.cluj4all.com - Recognos' own database about Cluj-Napoca with over 7000 described objectives, from which 1000 relevant touristic points) was also included. The data sets provided by various Romanian governmental agencies were used containing information for Romanian museums, churches, and historical points.

Wikipedia has categorized pages, and some of it's textual content tagged. There are very few consistent patterns followed by the content generators or authors (an exception would be the infobox content in the right side). However we observed that similar tags were used for describing the Romanian touristic objectives, and similar naming conventions for pages. For example, we retrieved values from page that respected the pattern "List_of_places_from_Cityname" from (http://ro.wikipedia.org/wiki/Lista_locurilor_in_Cluj-Napoca).

[2]Available at http://www.recognos.ro/lela/LelaLinkedDataSet.nq

TABLE I
LINKING AVAILABLE DATASETS.

Data set	Available at	Description
Romania Museum Guides	http://data.gov.ro/dataset	Descriptive data and geolocations of 967 museums in Romania
Wikipedia	http://wikipedia.ro	Various categories about Romanian touristic places
Freebase	http://www.freebase.com/	Community-curated database of well-known people, places and things - some about Romania
Geonames	http://www.geonames.org/	Covers all countries and contains over eight million place names - some about Romania
Wikisherpa	http://www.wikisherpa.com/	Data from wikiTravel in a more structured way
DBpedia	http://dbpedia.org/	Structured data from wiki to other external resources
Cluj4All	cluj4all.com	Around 7000 objectives about Cluj-Napoca

TABLE II
LINKING LELA ONTOLOGY WITH DBPEDIA.

Lela concepts	DBpedia concepts
POI	Museums, Castles, Towers, Churches, Cathedrals, Monuments, OutdoorSculptures, Bridges, Parks, Zoos
Activity	Cinemas, Theater, Activity, Shopping
Accommodation	Hotel
EatingAndDrinking	Restaurant

DBpedia organizes its data into triples, and data is linked to external data sets [8].

We queried the DBpedia database for 5 main cities (Bucharest, Cluj-Napoca, Timisoara, Brasov and Sibiu) following a predefined mapping of the 4 main Lela classes to the DBpedia specific classes (table II). A simple example for such

```
SELECT distinct
?subject ?latd ?longd ?about ?image ?category
?sameAs ?abstract ?wikipedia ?label
WHERE {
   ?subject <http://purl.org/dc/terms/subject>
            <http://dbpedia.org/resource:Category:
             Museums_in_#placeName\#>.}
   OPTIONAL {?subject dbpedia-owl:thumbnail ?image.}
   OPTIONAL {?subject rdfs:label ?label.}
   OPTIONAL {{?subject foaf:homepage ?about.}}
   OPTIONAL {{ ?subject geo:lat ?latd. ?subject
               geo:long ?longd.} union
            {?subject dbpprop:latitude ?latd.?subject
               dbpprop:longitude ?longd.}}
   OPTIONAL {?subject owl:sameAs ?sameAs.
            FILTER contains(str(?sameAs), "freebase").}
   OPTIONAL {?subject foaf:isPrimaryTopicOf ?wikipedia.
            FILTER contains(str(?wikipedia), "wikipedia").}
   OPTIONAL {?subject dbpedia-owl:abstract ?abstract.
            FILTER (LANG(?abstract)='ro' ||
                    LANG(?abstract)='en')}
   BIND('Museum' AS ?category).
            FILTER (!contains(str(?subject), "List_of")).
}}
```

Fig. 4. Querying DBpedia for the `Museum` category.

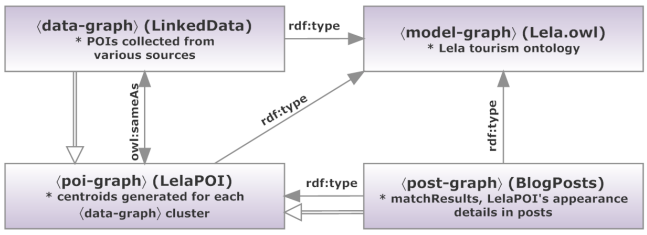


Fig. 5. LELA graph relations.

a query is shown in figure III-B, where where we extracted geospatial data, image, category, abstract and Wikipedia link for all instances of the Wikipedia' `Museum` category. *Geonames* has it's own category and ranking system for objectives. We retrieved the Hotels, Restaurants and Point of Interest for the above mentioned cities. *Freebase* and *Wikisherpa* have similar data collection processes. We accessed these resources through their APIs, and stored the obtained values in the Lela data store. Finally, we imported the xml and xls data from the government provided sources, and the Cluj4All private database.

The above cited data sources have many POIs that are relevant for the tourism domain or for Romania in general. However we only considered those POIs that satisfied both constraints at the same time, for the 5 most significant Romanian cities. As a result, the import process collected around 5.000 POIs, but the data was still noisy, as it contained many overlaps in the form of several instances of POIs collected from different sources semantically describing the same object. In order to overcome this issue, a fusion algorithm was applied. In order to make this process easier, we stored the data in several independent graphs, as the triple store solution we used - AllegroGraph - offers the possibility to partition data for an easier management. The reasoning mechanism can also be applied on the graph-level, not taking into account the several other graphs that might exist in a triple store. This proved to be helpful, when dealing with a bigger amounts of data.

In figure 5 the graph partitioning is shown. We created four separate graphs. The ⟨*model-graph*⟩ contains the TBox details of the Lela ontology. It includes the tourism taxonomy concepts and relations between them, as well as the object and data properties. The ⟨*data-graph*⟩ contains the external data, imported as triples. This graph uses some properties defined by the ⟨*model-graph*⟩ (concepts, label properties for names etc). In the fusion process, in the ⟨*data-graph*⟩ a representative element is created for each group of objectives semantically describing the same element. These groups are called by us POI clusters, and the representative element is the centroid POI. These centroids are saved separately in the ⟨*poi-graph*⟩, and we call them LelaPOIs. They have their derived properties from the cluster data, they also have a `owl:sameAs` property referring their original sources (instances from ⟨*data-graph*⟩). This process supports more efficient query plans, that exploit data from remote sources only when additional information related to a given POI is explicitly requested. Because of the `owl:sameAs` property, AllegroGraph allows us to get any details from data graph. The fourth graph is the ⟨*post-graph*⟩, which contains the blog post related information, like the a blog post's content or the matching details, after the post analysis.

The corresponding algorithm is summarized in Algorithm 1. In a first step, the algorithm finds all the instances that have the same wikipediaUrl, and link them with `owl:sameAs` property. The wikipedia urls are not ambiguous, so the operation will be correct. In a second step, it finds all the instances that have the same freebaseUrl, and if there is `owl:sameAs` between them, then add it.

Thirdly, it checks equality of label values(compare lela:hasName properties). If perfect match found and the objects are located in the same place, and no equality still reported, then link them with `owl:sameAs`.

Fourthly, the algorithm finds all the instances that have the same Web page. If there is no `owl:sameAs` link between them, adds it. Two additional steps have been also applied:

1) *Fusion cities* that have been read and imported from an xls document with cities and instances that have been imported from other resources like Geonames, Freebase.
2) *Fusion counties* that have been read and imported from an xls document with counties/instances that have been imported from other resources (Geonames, Freebase or others) - based on the previously mentioned information and same `lela:hasName` property.

Finally, based on the previously generated groups (a group is considered as a series of elements related by the `owl:sameAs` property) a special instance for each cluster is created in ⟨*poi-graph*⟩ (recall figure 5). The centroid of the group is asserted as an instance of the corresponding most specific concept from the Lela ontology.

Data: KB, the LELA Knowledge Base;
$xlsCities, otherCities$, lists of cities;
$xlsCounties, otherCounties$, lists of counties;
poi, the LELA POI graph;
Result: an augmented LELA Knowledge Base
foreach $i \in instances(KB)$ **do**
  foreach $j \in instances(KB)$ **do**
    if $i \neq j$ **then**
      if $wikipUrl(i) = wikipUrl(j) \vee freebaseUrl(i) = freebaseUrl(j)$ **then**
        $assert$(owl:sameAs$(i, j), KB)$
      end
    end
  end
end
foreach $i \in instances(KB)$ **do**
  foreach $j \in instances(KB)$ **do**
    if $i \neq j$ **then**
      if *lela:hasName*(i) = *lela:hasName*$(j) \wedge$ *loc*(i) = *loc*(j) **then**
        $assert$(owl:sameAs$(i, j), KB)$
      end
    end
  end
end
foreach $xc \in xlsCities$ **do**
  foreach $oc \in otherCities$ **do**
    if $xc \neq oc$ **then**
      if $wikipUrl(xc) = wikipUrl(oc) \vee freebaseUrl(xc) = freebaseUrl(oc)$ **then**
        $assert$(owl:sameAs$(xc, oc), KB)$
      end
    end
  end
end
foreach $xc \in xlsCounties$ **do**
  foreach $oc \in otherCounties$ **do**
    if $xc \neq oc$ **then**
      if (*lela:hasName*(xc) = *lela:hasName*$(oc)) \wedge (wikipUrl(xc) = wikipUrl(oc) \vee freebaseUrl(xc) = freebaseUrl(oc))$ **then**
        $assert$(owl:sameAs$(xc, oc), KB)$
      end
    end
  end
end
$clusters \leftarrow Partition(instances$,owl:sameAs)
foreach $c \in clusters$ **do**
  $i \leftarrow selectSpecialInstance(c)$
  $addToGraph(i, poi)$
end

Algorithm 1: LELA fusion algorithm

```
31. cat: EatingandDrinking, Location,
        Accommodation, POI, Activity,....;
32. fun ActivityhasLocation :
    EatingandDrinking -> Location -> PropertyCl;
33. Pizza : Object;
34. VSki: ActivityVerbPhrase;
35. VDrink, VEat : EatingandDrinkingVerbPhrase ;
36. V2Eat : Object ->
     EatingandDrinkingVerbPhrase;
37. QWhereModVerbPhrase :
      Modality -> VerbPhrase  -> Question;
```

Fig. 6. Abstract grammar derived from the Lela ontology.

C. Saving and publishing the resulting data sets

The data collection process resulted in the import of approximately 5.000 instances, some of them semantically describing the same point of interest, without any flag pointing out their equality. To eliminate this issue, the instances were grouped into clusters, based on characteristics such as their names, wikipedia pages, spatial coordinates, etc. For each cluster, the centroid was selected to became an instance of the `LelaPOI` concept. The centroid and the other individuals in the cluster are linked via a specific `similarity` relationship asserted in the LELA ontology. The unified data set is stored in a local AlegroGraph triplestore [9]. Currently the triplestore contains around 3.200 unique tourism objectives collected for five cities. The points of interest are described by 40.697 of RDF triples.

IV. Querying the linked data set in controlled language

To explore the linked dataset we provide a natural language query interface. The queries can be expressed in a controlled vocabulary for the Romanian language. The queries in natural language are automatically translated into SPARQL. The translation is based on three grammars that we developed in the Grammatical Framework [10], [11]:

1) one abstract grammar, derived from the Lela ontology;
2) one concrete grammar for the Romanian language
3) one concrete grammar for the SPARQL.

First, the abstract grammar in figure 6 is based on the main concepts and roles of the Lela ontology. The concepts in Lela are represented as categories in the grammatical framework, while roles as functions (lines 31-32). Individuals in the ontologies are modelled as instances of generic type *Object* (line 33). Activities are encapsulated as VerbPhrases (i.e., the verb VSki for the ski activity in line 35). Various eating and drinking activities are modelled with a specific verb phase (i.e., *EatingandDrinkingVerbPhrase* in line 36). The function introduces in line 36 is used to represent eating and drinking activities with parameters (i.e., eating pizza). The query template in line 37 is used to match against queries which include modal verbs (i.e., where can I eat pizza?).

Second, the concrete grammar for the Romanian language (figure 8 defines the controlled natural language used to query the system. The relevant verbs in the tourism domain are

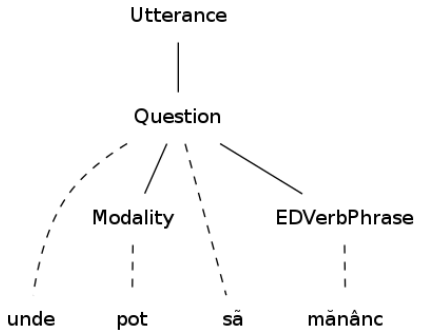


Fig. 7. Parse tree for ”Where can I eat?”. The pronoun is missing because the flexional form of the predicate `EDVerbPhrase` is enough to deduce the agent of the query.

specified (like drink in line 1, eat in line 2, sky in line 3) v_beschX are functions for smart paradigms of a language that provide different inflexion of verbs for different persons, numbers and tenses for the Romanian language. Romanian specific parsing rules are used here to define equivalence between related queries. For instance, specific to the Romanian language is to commit the pronoun in questions like ”Unde pot să mănânc?” (Where can I/We/You eat?), instead of ”Unde pot eu să mănânc?” (Where can I eat?) we define new forms for pronouns `ifemAbsent_Pron` (lines 44-48). The corresponding parse tree is depicted in figure 7, where the personal pronoun does not appear for the Romanian version of the query, as it can be deduce from the verb's flexional form.

Thirdl, the concrete grammar for SPARQL was develop to automatically translate the controlled natural language into a formal query. GF uses the grammars both for parsing and linearizing, therefore a translation from a Romanian phrase to a SPARQL query is done by (1) parsing in the grammar of Romanian language followed by (2) linearising the obtained parsed tree in SPARQL concrete grammar. For each verbal phrase, we give the corresponding SPARQL statement together with the name of the variable that is to be included in the SELECT clause of the query. The grammar in figure 9 is used to translate questions related to eating, drinking or other touristic activities (such as `SkiActivity`), queries which include modal verbs. The SPARQL grammar is exemplified in figure 10. The three queries in Romanian language illustrated in figure 10 are: ”Where can we play badminton?”, ”Where can I ski?”, ”Where is the Agape restaurant?”. The resulted SPARQL code can be used to directly query the Lela ontology

The system allows *qualitative queries* i.e., ”What is the atmosphere at Pizzeria Napoli?”, ”Which is the best restaurant with Romanian cousin?”. Answering to these queries exploits the knowledge provided by the opinion analyser module. The qualitative query ”Which restaurants have good food?” in figure 11 is matched against two concepts in the Lela ontology: i) intersection between `Food` and `Good` and ii) the concept defined by those instances whose role `QuisineQuality`

```
> p "unde putem sa jucam  badminton"
    |l -lang=triple
SELECT ?where, ?activity where {
 ?activity rdf:type ex:BadmintonActivity.
 ?where ex:hasActivity ?activity .
 ?match rdf:type Match . ?match hasSubject
 ?where. ?match hasScore ?score }
ORDER by Desc(?score)

>p "unde pot sa schiez" |l -lang=triple
SELECT ?where, ?activity
WHERE { ?activity rdf:type ex:SkiActivity.
      ?where ex:hasActivity ?activity .
      ?match rdf:type Match .
      ?match hasSubject ?where.
      ?match hasScore ?score }
ORDER by Desc(?score)

>p "unde este restaurantul Agape"
    |l -lang=triple
SELECT ?location where
    { ex:id342451 ex:hasLocation ?location}
```

Fig. 10. Queries in Romanian language automatically translated in SPARQL.

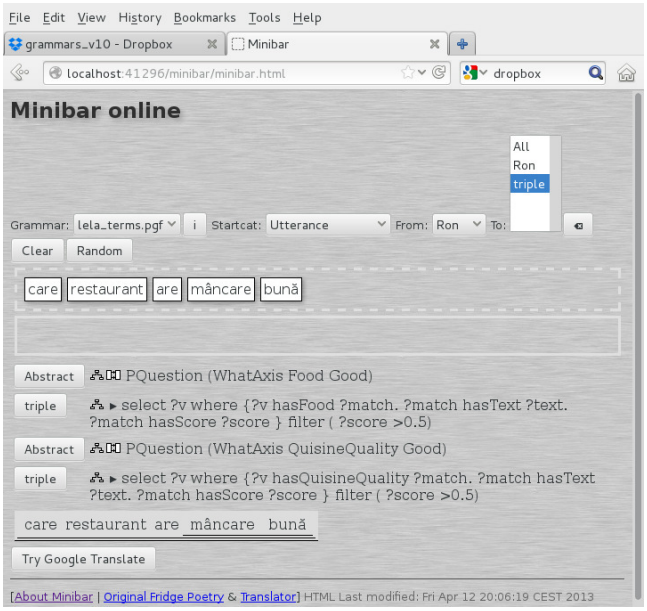


Fig. 11. Qualitative queries filter the results based on the available opinions on the topic of the query.

points towards the concept `Good`. The corresponding SPARQL queries filter the results to those instances classified by the opinion analyser module as positive (that is $?score > 0.5$).

The Romanian grammar is used to generate all the flexional forms of the vocabulary in order to guide the user to generate grammatically correct queries (12). In our case, the vocabulary is restricted to touristic terms from the Lela ontology. After typing a word, the system displays all the possibilities to complete the question in the defined controlled natural language.

To sum up, the translator of from Romanian language to SPARQL is able to handle the following types of queries in

```
41. VDrink = mkVP (v_besch73 "bea");
42. VEat   = mkVP (v_besch52 "manca");
43. VSki   = mkVP (v_besch10 "schia");
44. QWhere&ModVerbPhrase m vp =
45.   mkQS (mkQCl where_IAdv (mkCl (mkNP ifemAbsent_Pron) (mkVP m  vp)))
46.  |mkQS(mkQCl where_IAdv (mkCl (mkNPweAbsent_Pron) (mkVP m vp)))
47. ifemAb&sent_Pron =
48. P.mkPronoun [] "mine" "mie" [] [] "meu" "mea" "mei" "mele" Fem Sg P1 ;
```

Fig. 8. Concrete grammar for the Romanian language.

```
VDrink = {v="?eatdrink"; body="?eatdrink rdf:type ex:EatingandDrinking."};
VEat =   {v="?eatdrink"; body="?eatdrink rdf:type ex:EatingandDrinking."};
VSki =   {v="?activity"; body="?activity rdf:type ex:SkiActivity."};
QWhere&EatingandDrinkingDModVerbPhrase  x y ="select "++ y.v ++ "where {"
       ++ y.body ++
       ++ ``?match rdf:type Match . ?match hasSubject "++ y.v ++"."
       ++ "?match  hasScore  ?score }"
       ++"ORDER by Desc(?score)";
```

Fig. 9. Part of the grammar developed to translate a query into SPARQL.

Fig. 12. Guiding the process of constructing queries: (top) after the word ”ce” (what/how) is types only the grammatically correct flexional forms remain (bottom) the SPARQL version for the query ”How is the food at Agape restaurant?”.

which for each type several linguistic patterns are modelled: 1) Retrieving location of various elements from the Lela ontology (Accommodation, Eating and Drinking, Activities, POIs, etc.); 2) Identifying and describing simple activities (swim, walk) and compound activities (play badminton); 3) Handling queries containing reflexive or verbs with direct object; 4) Handling questions in which the subject is not explicitly expressed; 5) Enhancing verbs with modalities (can, should, may, etc.); 6) Qualitative queries.

V. Discussion and related work

Natural Language to SPARQL. To our knowledge, this is the first system which translates queries from the Romanian language into SPARQL syntax. The system relies on a domain-dependent controlled vocabulary in the tourism domain. For the English language, various systems do exist [12].

The *QuestIO* system [13] is open-domain, with the vocabulary automatically derived from the data existing in the knowledge. The system was designed to handle language ambiguities and incomplete or syntactically ill-formed queries by enacting fuzzy string matching and ontology similarity metrics. We focused on several specific difficulties for the Romanian language like: i) the inconsistent use of diacritics and special symbols, or ii) the flexibility of the sentence structure, which allows questions with or without pronouns.

The ONLI+ system [14] is a portable ontology-driven question answering system for English language. Similar to our work, the RacerPro system was used to reason on the ontology and to retrieve data. Differently, the translation is between English and nRQL, while in our case between Romanian and SPARQL, where both nRQL and SPARQL are recognized by RacerPro.

Another notable effort in the context of Semantic Web is the combination between between ACE and GF from [15]. Approaches for verbalization based on ontology is introduced in [16] for English and Greek languages. A controlled natural language for editing ontology is presented in [17] based on Attempto Controlled English (ACE) language.

Linking tourism data. Regarding the link data component, a similar approach is the tourism linked data set in [18], based on the European statistics data from 1985 about 150 cities in Europe. The Lela system complements linked open data with information extracted from blogs to offer both subjective impressions about places and objective data. Besides DBpedia, YAGO2 [19] focuses on automatically extracting and publishing structured knowledge from Wikipedia. While the DBpedia taxonomy is manually developed and maintained, YAGO integrates the WordNet taxonomy, which leads to a

higher number of classes in YAGO. Expanding our system to manage this richer taxonomy is one of the directions we intend to pursue as future work.

Romanian language processing. For the Romanian language several large annotated corpora do exist (George Orwell's novel 1984, Plato's Republic, ROCO), lexicons (WEB-DEX, CONCEDE, EUROVOC) [20] with the corresponding tools for exploiting these dictionaries (http://dexonline.ro/unelte) None of these resources deal with translation between a natural language and a formal language. We argue that such a translator can trigger various practical development at the application level. A Romanian grammar was developed by [21] that includes 866 grammatical rules and 320 affixes, which have been used for the development of a morphological vocabulary of cca. 30,000 words. For the natural language part of our work we based on the resource library for Romanian developed in [22]. Our morphological vocabulary was generated only for the tourism domain, with the goal to translate natural language queries into SPARQL. Our system for translating Romanian language queries into SPARQL syntax fills, in our view, an important gap among the existing linguistic resources for Romanian language [20].

VI. Conclusions

This paper introduced the Lela commercial product, which intends to be a semantic-based info-point for touristic information in Romania, offering both objective information and subjective impressions about places of interest. It provides data for the 4 main axes: accommodations, eating and drinkings, destinations and activities, with a special focus on the latter one. In order to provide these data, the system integrates Open Linked Data with subjective opinions expressed in articles to generate added value. The system also offers semantic search functionality through the Romanian natural language query interface, which translates the Romanian questions into SPARQL based on a controlled vocabulary derived from the developed LELA touristic ontology.

We are currently applying the natural language processing module to the task of populating the touristic objectives in Lela ontology with specific features identified in Romanian blogs.

The system is intended to be available for public use on http://www.lela.ro by the end of 2014.

Acknowledgements

Part of this work was supported by the PN-II Innovation Checks of the Romanian Ministry of Research for supporting research in small and medium enterprises, through the project "LELA - Collaborative Recommendation System in the Tourism Domain Using Semantic Web Technologies and Text Analysis in Romanian Language".

References

[1] Sutton, Charles and McCallum, Andrew, "An introduction to conditional random fields," vol., no., p., 267373, 2012. doi: 10.1561/2200000013. Available: http://dx.doi.org/10.1561/2200000013

[2] Bobicev, V. and Maxim, V and Prodan, T and Burciu, N. and Anghelus, V., "Emotions in words: developing a multilingual wordnet-affect," in *Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics, Iasi, Romania*, 2010. doi: 10.1007/978-3-642-12116-6_31. Available: {http://dx.doi.org/10.1007/978-3-642-12116-6_31}

[3] Noy, Natalya F and McGuinness, Deborah L and others, "Ontology development 101: A guide to creating your first ontology," 2001.

[4] Pollock, Jeffrey T and Hodgson, Ralph, "Ontology design patterns," doi: 10.1002/0471714216.ch7. Available: http://dx.doi.org/10.1002/0471714216.ch7

[5] Baader, Franz, *The description logic handbook: theory, implementation, and applications*. Cambridge university press, 2003. Available: http://dx.doi.org/10.2277/0521781760

[6] Haarslev, Volker and Hidde, Kay and Möller, Ralf and Wessel, Michael, "The racerpro knowledge representation and reasoning system," vol., no., 2012. doi: 10.3233/SW-2011-0032. Available: http://dx.doi.org/10.3233/SW-2011-0032

[7] Presutti, Valentina and Gangemi, Aldo, "Content ontology design patterns as practical building blocks for web ontologies," in *Conceptual Modeling-ER 2008*. Available: http://dx.doi.org/10.1007/978-3-540-87877-3_11

[8] Jens Lehmann and Robert Isele and Max Jakob and Anja Jentzsch and Dimitris Kontokostas and Pablo N. Mendes and Sebastian Hellmann and Mohamed Morsey and Patrick van Kleef and Sören Auer and Christian Bizer, "DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia," *Semantic Web Journal*, 2014.

[9] Watson, Mark, *Practical Semantic Web and Linked Data Applications - Common Lisp Edition*, 2010.

[10] Aarne Ranta, *Grammatical Framework: Programming with Multilingual Grammars*. Stanford: CSLI Publications, 2011, iSBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth).

[11] Aarne Ranta, "Gf: A multilingual grammar formalism," vol., no., 2009. doi: 10.1111/j.1749-818X.2009.00155.x. Available: http://dx.doi.org/10.1111/j.1749-818X.2009.00155.x

[12] Lopez, Vanessa and Uren, Victoria and Sabou, Marta and Motta, Enrico, "Is question answering fit for the semantic web?: a survey," vol., no., 2011. doi: 10.3233/SW-2011-0041. Available: http://dx.doi.org/10.3233/SW-2011-0041

[13] Tablan, Valentin and Damljanovic, Danica and Bontcheva, Kalina, "A natural language query interface to structured information," in *The Semantic Web: Research and Applications*. Available: http://dx.doi.org/10.1007/978-3-540-68234-9_28

[14] Mithun, Shamima and Kosseim, Leila and Haarslev, Volker, "Resolving quantifier and number restriction to question owl ontologies," in *Semantics, Knowledge and Grid, Third International Conference on*. IEEE, 2007. Available: http://dx.doi.org/10.1109/SKG.2007.255

[15] Kaarel Kaljurand and Tobias Kuhn, "A multilingual semantic wiki based on attempto controlled english and grammatical framework," vol., abs/1303.4293, 2013. doi: 10.1007/978-3-642-38288-8_29. Available: http://dx.doi.org/10.1007/978-3-642-38288-8_29

[16] Androutsopoulos, Ion and Lampouras, Gerasimos and Galanis, Dimitrios, "Generating natural language descriptions from owl ontologies: the naturalowl system," vol., 2013. doi: 10.1613/jair.4017

[17] Kaarel Kaljurand, "ACE View — an ontology and rule editor based on Attempto Controlled English," in *5th OWL Experiences and Directions Workshop (OWLED 2008)*, Karlsruhe, Germany, 26–27 October 2008. doi: 10.5167/uzh-8822 12 pages. Available: http://dx.doi.org/10.5167/uzh-8822

[18] Sabou, Marta and Arsal, Irem and Braşoveanu, Adrian MP, "Tourmislod: A tourism linked data set," vol., no., 2013. doi: 10.3233/SW-2012-0087. Available: http://dx.doi.org/10.3233/SW-2012-0087

[19] Hoffart, Johannes and Suchanek, Fabian M. and Berberich, Klaus and Weikum, Gerhard, "YAGO2: A spatially and temporally enhanced knowledge base from wikipedia," vol., 2013. doi: 10.1016/j.artint.2012.06.001. Available: http://dx.doi.org/10.1016/j.artint.2012.06.001

[20] Cristea, Dan and Forăscu, Corina, "Linguistic resources and technologies for romanian language," vol., no., p., 40, 2006. doi: 10.1.1.414.9781

[21] Boian, E and Ciubotaru, C and Cojocaru, S and Colesnicov, A and Demidova, V and Malahova, L, "Lexical resources for romanian-a project overview," in *Symposium on Intelligent Systems and Applications, September*, 2003. Available: http://dx.doi.org/10.2218/jls.v1i1.824

[22] Enache, Ramona and Ranta, Aarne and Angelov, Krasimir, "An open-source computational grammar for romanian," in *Computational Linguistics and Intelligent Text Processing*, vol., ISBN 978-3-642-12115-9. Available: {http://dx.doi.org/10.1007/978-3-642-12116-6_14}

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 289–296
DOI: 10.15439/2014F273
ACSIS, Vol. 2

Ontology-based Concept Similarity Integrating Image Semantic and Visual Information

Mengyun Wang, Xianglong Liu, Lei Huang, Bo Lang, Hailiang Yu
State Key Laboratory of Software Development Environment
Beihang University, Beijing 100191, China
Email:{jesuisenvie, xlong_liu, huanglei, langbo, yhl}@nlsde.buaa.edu.cn

Abstract—In recent years, the concept similarity measure has received wide attention in many applications, such as ontology construction, text analysis, image retrieval, etc. Currently, the concept similarity measure depends on the information mining in various knowledge bases, like dictionaries, ontologies, image annotation labels, and search engines. However, these knowledge bases usually only contain semantic information. With the development of the Internet and the popularity of the digital imaging devices, a lot of images and related texts have appeared, which help us to further mine the concept similarity relationships. The concept similarity is the outcome of human subjective perception. In addition to analysis of semantic information, the content of image itself precisely provides the visual perception information, which also plays an important role in the access of concept similarity relationships. To integrate both image semantic and visual information, in this paper we propose an ontology concept similarity measure that simultaneously utilizes the image semantic annotations and visual features to optimize the ontology-based metrics. The experiment result on the Corel dataset demonstrates the effectiveness of our proposed method.

I. Introduction

THE concept similarity plays a critical role in ontology construction and multimedia analysis. It is widely used in different researches and applications, such as ontology learning, semantic disambiguation, text clustering, and information annotation and retrieval.

Generally speaking, semantic similarity quantitatively describes the similarity degree between concepts. Traditional metrics assess the concept similarity relationships by exploiting one or several knowledge bases like corpus and ontologies. With the quick development of the Internet and the popularity of the digital imaging devices such as webcams, phone cameras, and digital cameras, a lot of images and related texts have emerged and formed a much rich knowledge base. For instance, Flickr provides free services for uploading and sharing images, where some necessary information such as titles, descriptions, and labels is usually required. As a large knowledge base, the text information can be used to evaluate the semantic similarity between labels [1]. In this paper, we consider the words 'label' and 'concept' refer to the same thing. However, such way only considers the correlation of text information around images, ignoring the indispensable effect of image visual information. Google search engine stores a wealth of network text resources, which can also help define the semantic similarity of concepts [2]. Compared with ontologies, there are several problems with concepts similarity measure only depending on the above knowledge bases: there is a lot of noisy data resulting inaccurate similarity measure; the concept definition is inexplicit, and cannot be used to distinguish synonyms or antonyms.

To address these problems, we propose an ontology concept similarity measure named OVS that simultaneously integrate both image semantic and visual information. Based on the ontology semantic similarity metrics, OVS exploits the images' visual features and related semantic annotations to optimize the semantic similarity relationships, which are more consistent with human cognition. Compared with traditional methods, the OVS integrates the semantic annotations around images with a variety of semantic knowledge in ontology such as hierarchical structure and semantic relationships, forming a richer semantic knowledge base. Meanwhile, we take images' visual information into consideration, utilizing visual knowledge base to optimize the concept similarity relationships. Integrating multiple knowledge bases together enables our approach to comprehensively express the semantic similarity relationships between concepts.

The main contributions of this paper are summarized as follows:

1. We propose a concept similarity measure by taking both images' visual features and semantic annotations into consideration.
2. We further propose an ontology-based concept similarity measure, which integrates both ontology semantic relationship and visual similarity measure.

The rest of this paper is organized as follows: Sec. II summarizes the existing ontology-based concept similarity measures. Sec. III introduces a novel ntology concept similarity measure named OVS based on the integration of image semantic and visual information. Sec. IV introduces the experiment and the last section is summarization of this thesis.

978-83-60810-58-3/$25.00

II. ONTOLOGY-BASED CONCEPT SIMILARITY MEASURES

An ontology is a formal, explicit specification of a shared conceptualization [3]. 'Formal' refers to the fact that the ontology should be machine readable, while 'Shared' reflects the notion that an ontology captures consensual knowledge. The characteristics mentioned above make an ontology be a reliable structured knowledge base. With the rapid development of the Semantic Web, a large number of universal ontologies and domain ontologies are generated and widely applied to knowledge-based systems, particularly, the measure of concept similarity.

One of the most widely used ontologies is Wordnet. It is an English semantic dictionary which is domain-independent. A synset corresponds to a concepts. Wordnet describes more than 100,000 English concepts and multiple semantic relationships between those concepts, such as hyponymy, part-of, synonymy, and antonymy, of which the hyponymy occupies nearly 80%. By connecting the related concepts together with multiple semantic relationships, Wordnet becomes a hierarchical structure or network structure. Based on this, researchers can mine the semantic similarity relationships of concepts with the methods like graph model.

Ontology-based concept similarity measures can be divided into four categories: the path-based method [4]-[7], the feature-based method [8], the IC(Information Content)-based method [9]-[12], and the gloss-based method [13][14]. In this section, we will introduce these methods and analyze the advantages and disadvantages of them.

Ontology can be modeled as a directed graph, in which a vertex represents a concept and an edge represents the hyponymy relation between two concepts. To calculate the similarity of concepts, the most direct method which is proposed by Rada et al. [4] is to compute the shortest path between two concepts, which follows the assumption that the concepts are more similar as the path is shorter. However, this method only considers the path between concepts, which cannot convey the similarity relationships of concepts precisely. Thus, several researches have taken the depth of the concept in the ontology into consideration [5][6]. Furthermore, multiple semantic relationships of concepts in the ontology are exploited, in addition to the hyponymy relations, to measure the semantic similarity relationships of concepts [7].

The advantage of the path-based method is the briefness in computation with the graph model, while faces the disadvantages that:

1. It only considers the shortest path between concepts within the ontology, ignoring the rich semantic knowledge of it.
2. The weight of every edge is identical in this approach. However, in the real sense, the semantic distance of each edge may not the same, which depends on the hierarchy granularity and degree of details described by concepts.

When two concepts are from different ontologies, the path-based method cannot calculate the semantic similarity of the concepts. Luckily, feature-based method can adapt to it. Feature-based method count the common part of property features between two concepts. If the common part is large, the concepts are similar, otherwise not. The property features can be extracted from a variety of semantic information like hypernym. This method describes the similarity degree of concepts more precisely by taking the common characteristic and the difference of concept properties into consideration. However, it usually depends on a large ontology like Wordnet.

In order to make up the lack of path-based measure, Resnik [9] proposed an IC-based method. He tried to utilize the IC shared by concepts to calculate the semantic similarity. IC can be derived from the frequency of concept appearance in a corpus, and the shared part of two concepts is represented by the LCS(Last Common Subsumer). But there is a problem: in terms of two concept-pairs with the same LCS, they will get the same similarity score. To address this problem, Lin [10] and Jiang & Conrath [11] proposed methods to improve Resnik's method. They both considers the IC of two concepts as well as the IC of their LCS, to represent the concepts more comprehensively. The key point of this method is the access of IC, which depends on the corpus after text processing or the ontology with rich concepts. The accuracy of IC-based measure, to some extent, is influenced by the analysis of these knowledge bases.

The gloss-based method was first applied to semantic disambiguation. Lesk [15] compared the glosses of a word in a phrase with the glosses of others, finding the most similar sense as the sense of the word in this phrase. The glosses were described in a dictionary. Then Banerjee & Pedersen [13] replaced it with Wordnet and introduced the ontology in. The premise of this method is the existence of a perfect corpus, including detailed gloss of words. Wordnet can satisfy the requirement which provides the hierarchical structure of concepts and abundant semantic glosses.

III. ONTOLOGY CONCEPT SIMILARITY METHOD BASED ON IMAGE SEMANTIC ANNOTATIONS AND VISUAL FEATURES

In Sec. II, we summarize different categories of the ontology-based concept similarity measures. In general terms, the measure of concept similarity is extracting a variety of semantic information based on different knowledge bases, to find the semantic similarity relationships between concepts. Ontology provides a variety of knowledge sources for semantic similarity measure, which contains a wealth of information on the concept gloss information and other semantic relations with a hierarchical structure based on the hyponymy relations. However, they are all semantic information based on text whether the glosses or the multiple semantic relations of concepts. As mentioned above, there are a lot of images and related semantic annotations on Flickr. Some researchers have tried to use the image semantic

annotations to obtain the similarity relationship of labels. With the fast growing of images on the Internet and the research and application of ontology in image analysis, annotation and retrieval, the visual information of images is richer. And from a visual point of view, we are more likely to perceive whether two concepts are similar. Therefore, we propose an ontology concept similarity measure named OVS based on the integration of image semantic and visual information, which optimize the ontology-based method with images' semantic annotations and visual features, to obtain more consistent concept similarity relationships with human cognition.

A. Semantic and Visual Relations based on Labeled Images

The images in the web usually carry some labels expressing certain semantic concepts. Making use of annotation relationships between images and concepts in terms both semantic and visual correlations, we can obtain more powerful concept similarity measure.

Suppose there are N images, each of which is annotated by one or several labels from M concepts. Then we denote the pairwise similarity relationships between multiple concepts from both semantic and visual aspects using a symmetric matrix R with each element r_{ij} representing the similarity between concept c_i and c_j. The similarity matrix R consists of a semantic relation matrix and a visual relation matrix respectively characterizing the correlations between concepts and images based on the semantic annotations and visual features.

The semantic relations matrix S is a matrix of $M \times N$. Each element of S is expressed as followed.

$$s_{ij} = \begin{cases} 1 & \text{if image j is annotated by concept i} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Visual relations matrix V is also a matrix of $M \times N$. Each element of V represents the similarity between the visual features of concepts and images. Each concept i corresponds to a subset of images annotated by it. The visual features of concept i then can be described by the subset of images. We take the average of all these images' visual features from the subset as the visual feature of concept i. With the visual features of concept i by f_i and that of image j represented by f_j, we calculate the element v_{ij} of V as followed. In this paper, we choose SIFT to be the visual features of images, as it has demonstrated its out-performance in some benchmark evaluation comparing with various global features and local feature descriptors.

$$v_{ij} = \cos(f_i, f_j) = \frac{\langle f_i, f_j \rangle}{\|f_i\| \|f_j\|} \quad (2)$$

In practice, not every image shows strong similarity to its concept due to the visual variation of each concept, which will assign a lot of undesired images with small similarity to each concept in the dense matrix V. In order to make the visual relations matrix V robust to the variations, we only select the top k concepts with highest similarity for each image, thus forming a sparse visual relations matrix V. The smaller the value of k, the more sparse the visual relations matrix V (the value of k is determined in Sec. IV.C.1).

B. Concept Similarity based on Image Semantic and Visual Relations

To incorporate both semantic and visual relations in the labeled image sets, we adopt a linear combination of V and S to get a comprehensive relations matrix VS,

$$VS = \mu * V + (1 - \mu) * S \quad (3)$$

where μ in (0,1) is the weighting factor, which balances the contribution of the semantic and visual information (the value of μ is determined in Sec. IV.C.2) in the concept similarity measure.

Based on the fused relation matrix $VS = [y_1, \ldots, y_m]^T$, where the m-th row y_m is the comprehensive relations vector of concept m. If two concepts share close semantics, then their comprehensive relations vectors should be very similar. Therefore, we employ the cosine similarity between the comprehensive relation vectors as the semantic similarity between two concepts.

$$r_{ij}^{VS} = \cos(y_i, y_j) = \frac{\langle y_i, y_j \rangle}{\|y_i\| \|y_j\|} \quad (4)$$

By calculating the pairwise similarities between all concepts, we can obtain the similarity matrix R based on image semantic annotations and visual features. The above method for the concept similarity measure is named VS method.

In practice, the semantic annotations of the image are often incomplete, which leads to the sparsity of the semantic relations matrix S. In this case, the method only relying on image semantic annotations hardly obtains an accurate semantic similarity relationships of concepts. However, the VS method introduces the image visual information in Sec. III.A which hopefully compensates for the lack of semantic information.

One problem in the process of calculating the visual relations matrix V when only a few limited labels are available is that, the incompleteness of the semantic information will result in the bias of the visual feature center of each concept, which will adversely affect the visual relations matrix V. To solve the problem, attempts(like EM algorithms) that iteratively find the near-optimal visual feature center of each concept can be adopted.

C. Ontology Metric with the Concept Similarity

We have presented a concept similarity measure VS based on image semantic annotations and visual features. In this section, we will utilize it to optimize the metrics based on ontology. Ontology provides a wealth of semantic knowledge, especially the universal ontology (e.g. Wordnet), which does

not depend on specific areas and is more consistent with the human conception. On the other hand, the human visual perception is also the important factor that affects the judgment of the concept similarity. The visual feature of the image is one general way to precisely express the concept of human visual perception. Therefore, fusing the combination of both semantic and visual relations into the traditional ontology metric can effectively compensate for the lack of the rich and complex relations in ontology.

Specifically, we adopt the parameter weighting method [16]-[18] to calibrate the ontology similarity metric using the concept similarity R based on the semantic relations and visual relations extracted from a number of labeled images. Since the sparse semantic annotations in images usually introduce certain bias in the similarity measure, it is better to fuse the ontology metric and concept similarity measure in a robust way. In this paper, we propose a novel method named OVS in short, which adopt the popular exponential production to robustly incorporate the visual-semantic relations in the ontology metric. Namely, for any two concepts c_i and c_j, their final ontology-based concept similarity is as followed.

$$r_{ij}^{OVS} = r_{ij}^{O} \exp(r_{ij}^{VS}) \quad (5)$$

In the above formula, r_o is the concept similarity based on ontology while r_{vs} is the concept similarity based on image semantic and visual information.

IV. Experiment

A. Methods for Comparison

In order to verify the effectiveness of the method OVS we propose, we conduct several comparative experiments with two kinds of metrics, which are metrics based on the ontology and metrics based on the image semantic and visual information, as shown in Table I.

TABLE I.
METHODS FOR COMPARISON

Knowledge base	Method	Type	Published in	Formula
Ontology	WUP [5]	Path	1994	$sim_{WUP} = \frac{2*N_3}{N_1+N_2+2*N_3}$
	LCH [6]	Path	1998	$sim_{LCH} = -\log(\frac{N}{2*D})$
	HSO [7]	Path	1998	$sim_{HSO} = C - full\_path(c_1,c_2) - k*turns(c_1,c_2)$
	RAD [4]	Path	1989	$dis_{RAD} = \min_{\forall i} \lvert p_i(c_1,c_2) \rvert$
	SAN [8]	Feature	2012	$dis_{SAN} = \log_2(1+(\lvert \phi(c_1) \setminus \phi(c_2) \rvert + \lvert \phi(c_2) \setminus \phi(c_1) \rvert) / (\lvert \phi(c_1) \setminus \phi(c_2) \rvert + \lvert \phi(c_2) \setminus \phi(c_1) \rvert + \lvert \phi(c_1) \cap \phi(c_2) \rvert))$
	JCN [11]	IC	1997	$dis_{JCN}(c_1,c_2) = (IC(c_1)+IC(c_2)) - 2*sim_{RES}(c_1,c_2)$
	RES [9]	IC	1995	$sim_{RES}(c_1,c_2) = IC(LCS(c_1,c_2))$
	LIN[10]	IC	1998	$sim_{LIN}(c_1,c_2) = \frac{2*sim_{RES}(c_1,c_2)}{IC(c_1)+IC(c_2)}$
	JCN_SAN [12]	IC	2013	$IC_{SAN}(c) = -\log p(c) \cong -\log\left(\frac{\lvert leaves(c) \rvert}{\lvert subsumers(c) \rvert}+1\right) / (max\_leaves+1)$
	RES_SAN [12]	IC		
	LIN_SAN [12]	IC		
	LESK [13]	Gloss	2002	-
	VECTOR [14]	Gloss	2006	-
	VECTOR_PAIRS [14]	Gloss	2006	-
Image	VS	-	Sec. III.B	-

B. Evaluation Benchmark and Indicator

The objective evaluation of concept similarity is very difficult as the concept similarity is a human's subjective perception. To compare the methods fairly, some scholars have constructed artificial evaluation datasets as the groundtruth. This dataset contains a number of concept-pair, judge by a group of people with a similarity score. The average of valid data is the similarity of the concept-pair. In early, Rubenstein & Goodenough [19] and Miller & Charles [20] constructed those datasets, which are widely used to evaluate and compare the similarity measure. The current popular datasets are WordSim-353 [21], MEN [22], RWS [23], etc. MEN is constructed by Elia Bruni, containing 3000 concept-pairs with high frequency of appearance, in which the similarity is in the range of 0-50. We choose Corel [24] to calculate the concept similarity, which contains 260 concepts. Based on the consideration of the above two aspects, we selected overlap portions of MEN and Corel as the target concept, which contains a total of 96 concepts, composing 118 concept-pairs.

Considering the artificial evaluation of concept similarity and the metric result as two sequences, their relevance is the indicator to judge semantic similarity metrics. The correlation is 1 if the scores are both exactly the same , which means that the results of the semantic metrics is consistent with human perception, while correlation of 0 means that the result of semantic metrics and the result of human perception are completely irrelevant. Currently the most widely used methods are the Pearson and Spearman correlation coefficient. Pearson correlation coefficient is calculated as

$$\rho_p = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{N\sum x_i^2 - (\sum x_i)^2}\sqrt{N\sum y_i^2 - (\sum y_i)^2}} \quad (6)$$

In the above formula, $x_1,x_2,\ldots,x_n$ and $y_1,y_2,\ldots,y_n$ are the result of human judgment and semantic metric. Spearman correlation coefficient is usually considered to be the Pearson correlation coefficient between the variables after ranking. We need to rank the sequence first. If the rank is identical, the Pearson and Spearman correlation coefficients are equivalent according to the equation (6). Otherwise it is calculated according to the following formula,

$$\rho_s = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \quad (7)$$

where d_i is the difference of x_i and y_i.

C. Results and Discussion

In order to verify the effectiveness of the method OVS, we conduct several comparative experiments with two kinds of metrics, which are metrics based on the ontology and metrics based on the image semantic and visual information.

1) The Impact of the Sparse Degree of Visual Relations Matrix

In Sec. III.A, we consider removing the interference data of visual relations matrix V, and only selecting the first k concepts with highest similarity of the image, thus forming a sparse visual relations matrix V. The smaller the value of k is, the higher the sparse degree of V is. In this section, we take the experiment to evaluate the value of k.

Parameter μ is the weight factor for the image semantic and visual information. $\mu = 0$ means to calculate the semantic similarity matrix R with only image semantic annotations, while $\mu = 1$ means to calculate R with only image visual features. We take the case of $\mu = 0$ as our basis, to calculate the top 10 similar concepts of each concept on Corel. And then we calculate the top 10 similar concepts with different values of k in the case of $\mu = 1$. After all, the accuracy of the top 10 similar concepts with different values of k, with respect to the case of $\mu = 0$ is seen as the index to judge the sparse degree of V. As the number of labels belonged to a image is limited, we set the range of k as 10 or less. The result is shown in Fig. 1.

As Fig. 1 shows, with the increasing of k , the sparse degree of V decreases and the accuracy of top 10 similar concepts decreases accordingly. It is obvious that the value of k should be as small as possible. But if the value of k is too small, it will cause the over lost of the information in V. So in the following experiment, we take k = 5.

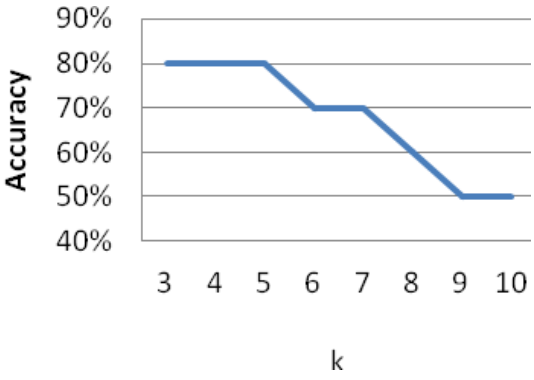


Fig. 1 The accuracy of top10 similar concepts with different values of k

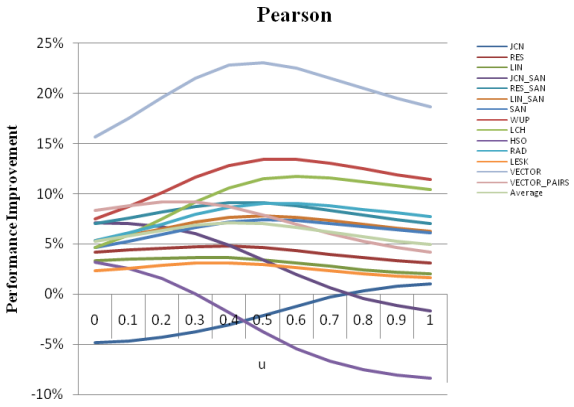


Fig. 2 Performance Improvement of Pearson correlation coefficients

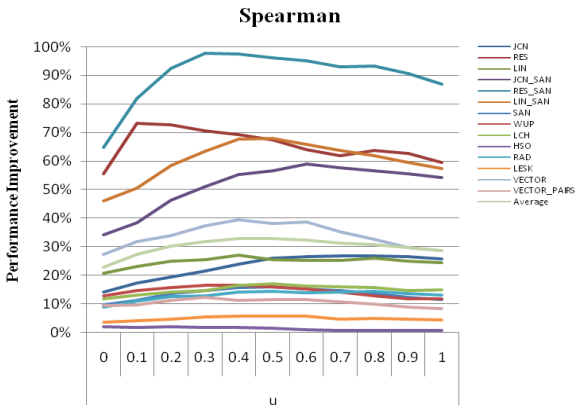


Fig. 3 Performance Improvement of Spearman correlation coefficients

2) The Choice of Weighting Factor μ

In Sec. IV.B, we propose the evaluation benchmark and index of semantic similarity metrics. First, we calculate the semantic similarities of the target concept-pairs with the ontology-based metrics listed in Table I, and then calculate the correlation coefficients of every single result of those metrics with human judgment. Secondly, based on those ontology metrics in Table 1, we utilize OVS method proposed in Sec. III.C we calculate the semantic similarities of the target concept-pairs separately, and then calculate the correlation coefficients of these results with human judgment. Finally, we calculate the correlation coefficients promotion of results based on OVS with respect to the results of ontology-based metrics, and the result is shown in Fig. 2 and Fig. 3.

As Fig. 2 and Fig. 3 shows, each curve represents the change trend of the correlation coefficients promotion of the

TABLE II.
COMPARISON OF OVS WITH METHODS BASED ON ONTOLOGY AND IMAGE SEMANTIC AND VISUAL INFORMATION OVS

Method	Type	Pearson		Spearman	
		Ontology	OVS	Ontology	OVS
WUP	Path	0.3847	**0.4363**	0.3678	**0.4271**
LCH	Path	0.4778	**0.5327**	0.4600	**0.5382**
HSO	Path	**0.4353**	0.4189	0.4759	**0.4833**
RAD	Path	0.4093	**0.4461**	0.4600	**0.5269**
SAN	Feature	0.3800	**0.4081**	0.3662	**0.4244**
JCN	IC	**0.1254**	0.1227	0.3329	**0.4200**
RES	IC	0.3122	**0.3266**	0.1454	**0.2435**
LIN	IC	0.3216	**0.3325**	0.2080	**0.2613**
JCN_SAN	IC	0.2164	**0.2239**	0.2639	**0.4134**
RES_SAN	IC	0.3081	**0.3361**	0.1495	**0.2933**
LIN_SAN	IC	0.3262	**0.3518**	0.1718	**0.2883**
LESK	Gloss	0.3711	**0.3821**	0.5562	**0.5879**
VECTOR	Gloss	0.3387	**0.4167**	0.2536	**0.3507**
VECTOR_PAIRS	Gloss	0.3514	**0.3792**	0.3624	**0.4045**
VS	Image	-	0.2919	-	0.3533

OVS method based on every ontology-based metrics with different value of μ. Due to the different methods to get the concept similarity, which leads to different rank, Pearson and Spearman correlation coefficients are not the same. But the changing trend of promotion with different value of μ is almost the same. In Fig. 2, except that the Pearson promotions of OVS based on JCN, HSO, JCN_SAN metrics vary monotonically with different value of μ, the promotions of OVS based on other ontology-based metrics get the maximum value with the value of $\mu \in (0.3,0.6)$. We also can find the similar conclusion for Spearman promotion. Based on the analysis above, in the sequent experiments, we take the optimal value $\mu = 0.5$.

3) Evaluation of Concept Similarity Method

As Fig. 2 and Fig. 3 shows in Sec. IV.C.2, the concept similarities based on OVS get a promotion with respect to ontology-based metrics, except JCN, HSO and JCN_SAN. We also can get the same conclusion from Table II.

In addition, we can also get the idea from Table II that the path-based and gloss-based methods among the ontology-based methods have the best effect whether they are based on ontologies alone or based on the OVS, especially for the LCH metric, while IC-based methods have the worst. However, the IC-based methods get a promotion after the optimization of OVS. It is clear that the OVS method integrating the image semantic annotation and visual features together can achieve better performance than ontology-based methods. We also can find that in most cases the OVS method based on ontology metrics have a better effect than the VS method we proposed in Sec. III.A. Thus, we believe that the concept similarity relationships derived from ontologies and image semantic and visual information together is superior to the methods of ontologies or image semantic and visual information.

V. CONCLUSION

In this paper, we summarize the concept similarity methods based on ontology, and presents an ontology-based concept similarity method OVS integrating the image semantic and visual information. Firstly, OVS get concept similarities with different ontology-based metrics, then compute another concept similarity with the image semantic annotations and visual features as a knowledge base, and finally integrate the two kinds of measurement, constituting a new concept similarity measure. In the experiments, we discuss and determine the OVS method parameter value and compare the methods based on ontology and image semantic and visual information with the evaluation benchmark of human judgment. The result verifies the effectiveness of the proposed method OVS.

In this paper, we measure the concept similarity relationships on the Corel dataset, which contains less visual information than the images in the network. As future work, we plan to apply our method to large datasets like Imagenet and applications like image annotation and retrieval.

REFERENCES

[1] Xu H, Zhou X, Wang M, et al. Exploring Flickr's related tags for semantic annotation of web images. Proceedings of the ACM International Conference on Image and Video Retrieval. ACM, 2009: 46. DOI: http://dx.doi.org/10.1145/1646396.1646450

[2] Cilibrasi R L, Vitanyi P M B. The google similarity distance. Knowledge and Data Engineering, IEEE Transactions on, 2007, 19(3): 370-383. DOI: http://dx.doi.org/10.1109/TKDE.2007.48

[3] Studer R, Benjamins V R, Fensel D. Knowledge engineering: principles and methods. Data & knowledge engineering, 1998, 25(1): 161-197. DOI: http://dx.doi.org/10.1016/S0169-023X(97)00056-6

[4] Rada R, Mili H, Bicknell E, et al. Development and application of a metric on semantic nets. Systems, Man and Cybernetics, IEEE Transactions on, 1989, 19(1): 17-30. DOI: http://dx.doi.org/10.1109/21.24528

[5] Wu Z, Palmer M. Verbs semantics and lexical selection. Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1994: 133-138. DOI: http://dx.doi.org/10.3115/981732.981751

[6] Leacock C, Chodorow M. Combining local context and WordNet similarity for word sense identification. WordNet: An electronic lexical database, 1998, 49(2): 265-283.

[7] Hirst G, St-Onge D. Lexical chains as representations of context for the detection and correction of malapropisms. WordNet: An electronic lexical database, 1998, 305: 305-332.

[8] Sánchez D, Batet M, Isern D, et al. Ontology-based semantic similarity: A new feature-based approach. Expert Systems with Applications, 2012, 39(9): 7718-7728. DOI: http://dx.doi.org/10.1016/j.eswa.2012.01.082

[9] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007, 1995.

[10] Lin D. An information-theoretic definition of similarity. ICML. 1998, 98: 296-304.

[11] Jiang J J, Conrath D W. Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008, 1997.

[12] Sánchez D, Batet M. A semantic similarity method based on information content exploiting multiple ontologies. Expert Systems with Applications, 2013, 40(4): 1393-1399. DOI: http://dx.doi.org/10.1016/j.eswa.2012.08.049

[13] Banerjee S, Pedersen T. An adapted Lesk algorithm for word sense disambiguation using WordNet[M]. Computational linguistics and intelligent text processing. Springer Berlin Heidelberg, 2002: 136-145. DOI: http://dx.doi.org/10.1007/3-540-45715-1_11

[14] Patwardhan S, Pedersen T. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. Proceedings of the EACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together. 2006, 1501: 1-8.

[15] Lesk M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. Proceedings of the 5th annual international conference on Systems documentation. ACM, 1986: 24-26. DOI: http://dx.doi.org/10.1145/318723.318728

[16] Tversky A. Features of similarity. Psychological review, 1977, 84(4): 327. DOI: http://dx.doi.org/10.1037//0033-295X.84.4.327

[17] Rodríguez M A, Egenhofer M J. Determining semantic similarity among entity classes from different ontologies. Knowledge and Data Engineering, IEEE Transactions on, 2003, 15(2): 442-456.

[18] Zhou Z, Wang Y, Gu J. A new model of information content for semantic similarity in WordNet. Future Generation Communication and Networking Symposia, 2008. FGCNS'08. Second International Conference on. IEEE, 2008, 3: 85-89. DOI: http://dx.doi.org/10.1109/FGCNS.2008.16

[19] Rubenstein H, Goodenough J B. Contextual correlates of synonymy. Communications of the ACM, 1965, 8(10): 627-633. DOI: http://dx.doi.org/10.1145/365628.365657

[20] Miller G A, Charles W G. Contextual correlates of semantic similarity. Language and cognitive processes, 1991, 6(1): 1-28. DOI: http://dx.doi.org/10.1080/01690969108406936

[21] Finkelstein L, Gabrilovich E, Matias Y, et al. Placing search in context: The concept revisited. Proceedings of the 10th international conference on World Wide Web. ACM, 2001: 406-414. DOI: http://dx.doi.org/10.1145/371920.372094

[22] Bruni E, Boleda G, Baroni M, et al. Distributional semantics in Technicolor. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012: 136-145.

[23] Luong M T, Socher R, Manning C D. Better word representations with recursive neural networks for morphology. CoNLL-2013, 2013, 104.

[24] Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

1st Complex Events and Information Modelling

COMPLEX Events and Information Modelling (CEIM). Event modelling is a method of intelligent analyzing streams of information (data, percepts) about things that happen (events), and deriving conclusions from them. The goal of CEIM is to identify meaningful events and respond to them appropriately and quickly. We define the complexity of the event as both the complexity of the modelled physical phenomenon (fire, weather, chemical reaction, biological process) as well as the heterogeneity of the data (digital images, percepts, sensory data, natural language, semi-structured and structured data). In addition, the emphasis should be placed on the intelligent aspect of these models. This means that systems should semi- autonomously perceive their environment and take action.

The workshop on Complex Events and Information Modelling provides an interdisciplinary forum for researchers and developers. The workshop is mostly intended for applications of Artificial Intelligence in Fire Safety domain, but we also encourage researchers from other fields: health care, smart buildings, ubiquitous computing, process mining and others. The workshop covers the whole range of theoretical and practical aspects, technologies and systems and aims at bringing together specialists for exchanging ideas and promote interdisciplinary discussions.

Topics

- Artificial Intelligence techniques in Fire Safety,
- Data Assimilation and Smart Buildings,
- Evacuation models,
- Cognition and Decision Making models during Emergency,
- Sensory data storage representation and processing,
- Ubiquitous computing,
- Uncertainty modelling,
- Automated reasoning,
- Risk Management,
- Knowledge Discovery, Data and Process Mining,
- Decision Support Systems,
- Knowledge Modelling.

Event Chairs

Krasuski, Adam, The Main School of Fire Sevice, Poland
Rein, Guillermo, Imperial College London

Program Committee

Bargiela, Andrzej, University of Nottingham, United Kingdom
Bazan, Jan, University of Rzeszów, Poland
Butler, Bret W., Missoula Fire Sciences Laboratory
Chaudhury, Santanu, Indian Institute of Technology Dehli, India
Gałaj, Jerzy, The Main School of Fire Service, Poland
Gelenbe, Erol, Imperial College London, United Kingdom
Hamins, Anthony, National Institute of Standard and Technology, United States
Hostikka, Simo, VTT Technical Research Centre of Finland, Finland
Jahn, Wolfram, Raindance Science International, Chile
Jankowski, Andrzej, Warsaw University of Technology, Poland
Jin, Peng, Leshan Normal U. China
Liu, Jiming, Hong Kong Baptist University, Hong Kong S.A.R., China
Meina, Michał, Nicolaus Copernicus University
Mendonca, David, Rensselaer Polytechnic Institute, United States
Merci, Bart, Ghent University, Belgium
Mirończuk, Marcin, Institute of Computer Science Polish Academy of Sciences
Muzy, Alexandre, National Center for Scientific Research
Nguyen, Hung Son, University of Warsaw, Poland
Ohsawa, Yukio, The University of Tokyo, Japan
Pilemalm, Sofie, Linköping University, Sweden
Robinson, Karen, NICTA, Australia
Ronchi, Enrico, Lund University, Sweden
Roszkowska, Ewa, University of Bialystok
Rykaczewski, Krzysztof, Nicolaus Copernicus University
Salamonowicz, Zdzisław, The Main School of Fire Service
Sikora, Beata, Silesian University of Technology
Simeoni, Albert, University of Edinburgh
Tarapata, Zbigniew, Military University of Technology, Poland
Tavares, Rodrigo Machado, RMT Fire & Crowd Safety, Brazil
Velasquez Silva, Juan D., Web Intelligence Research Chile Centre
Welch, Stephen, University of Edinburgh, United Kingdom

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 299–306
DOI: 10.15439/2014F353
ACSIS, Vol. 2

Heuristic to Build RCC8 for Event Locations

Majed Ayyad
University of Trento.
Via Sommarive 14 - 38123 Povo
Trento, Italy
Email: ayyad@disi.unitn.it

Abstract—**Events that are detected and reported by humans to actionable knowledge bases in multi-tier responding agencies have signi cant amount of spatial information. Humans have intuitive ability to triage repeated or duplicated events based on their spatio-temporal information. However, this cognitive process is not modeled easily and human ability is limited in situations where large number of events are reported simultaneously. The likelihood of two events to be the same is higher if they occur on the same place and time. In this work, we focus only on calculating location equivalence of events. For this purpose we use RCC8 theory to represent spatial relations between regional locations. The algorithm designed approximates the arbitrary shape of regions into circles and build region connection relations based on the size of the circle. The end result is a region of circular tiles with explicit RCC8 relations that could be used to reason on the relation between the locations of events. Additionally, we outline some experiments to evaluate the precision and recall of the results based on the used corpus. These results indicate that although the task is challenging, automated methods are capable of building spatial regional relations between events.**

I. Introduction

DESPITE advances in technology and ubiquitous computing, a large volume of events are still detected and reported by human beings. Statistics about emergency rooms or command and control rooms still report about receiving hundreds of calls per day from the public. In such situations, operators log each call and forward it to a dispatcher or commander. Typically , dispatchers and operators sit next to each other, so they can rapidly share information among one another in case the system is jammed with calls and there are emergencies to reply to. The efficiency of this system is measured by the number of abandoned calls and the response time of the first responder (the dispatched resource to the event site). To improve both elements, commanders could benefit from having a system that can work on the triage of events logged by operators and improve the decision making process before taking any action. In such situations, there is a need to triage repeated or duplicated events in real-time. We call this as the event matching problem which is the process of finding similarity between two events and is computed as 3-tuples $<e_1, e_2, R>$, where R specifies a similarity relation. Possible forms for R are : equivalence (=), sub-event ($\subset$), and mismatch ($\perp$).

Davidson[2] and Quine [3] argue that two events are identical if they occupy the same portion of space and time. Therefore, for two events to match or (when x= y ? ; if x an y are events) is to find the necessary and sufficient conditions for identical events. The matching criteria depends on a set of elements which we can summarize as : time, location, physical objects, cause and effect, existential conditions and properties. Therefore to match two events we need to calculate their time equivalence , location equivalence, causal equivalence and properties equivalence using other arguments such as participants and objects.

Comparing the location of two events is not always a straight forward, especially when events are reported using natural language. Different qualitative spatial relations are used to express the location of an event with other spatial entities. For the orientation aspect, events are described using qualitative terms such as “north of”, “in front of”, ”behind”, etc. Many approaches and calculi have been used to express the orientation of one object on reference to another. Most approaches use points as the basic spatial entities and use different versions of JEPD orientation relations. Distance qualitative relations are also used when describing the location of events. For the distance aspect, terms such as “near”, “far”, “close to” are commonly used. As mentioned by [7] combining the orientation and distance aspects is called positional information

In this work, we use the Region Connection Calculus (RCC8) theory to partially solve the event matching problem. A location is defined as an inherently grounded spatial entity, a location includes geospatial entities such as countries, mountains, cities, rivers, etc. It also includes classificatory and ontological spatial terms, such as edge, corner, intersection[4]. The location element covers both locations and places (where a place is considered a

978-83-60810-58-3/$25.00

functional category), and is assumed to be associated with a region whenever appropriate[10]

The main objective of this paper is to demonstrate how connectedness relations between geographical spaces could be calculated automatically. The following five topological relations between locations are built: (1) Equal (2) Externally Connected (3) Disconnected (4) Tangential Proper Part, and (5) Non-Tangential Proper Part. In this work, and by using a dataset of a country we build RCC8 relations between cities, towns, villages, suburbs and points of interests.

For this purpose , we use an approximation technique to represent a region as circular shape. Furthermore, we represent a country map from circular tiles. The radius of the circle is calculated based on the type of the region being a city or a hamlet as an example. Other parameters are also considered if available such as the area and population of a region. We show that our heuristic algorithm to build RCC8 relations between country regions and places is likely to achieve acceptable results

This paper is organized as follows. Section 2 includes background information on RCC8. Section 3 describes the estimation problems and their formulation. Experimental results are presented in Section 4 and the paper is concluded in Section 5.

II. The Region Connection Calculus

A. RCC8 relations

There are different aspects of space related to describing the event location on reference to another object. The location of an event could be expressed using a combination of orientation relations, distance relations and topological relations. While orientation and distance relations are important, in this paper we focus only on topological relations. Topology in mathematics concerned with the most basic properties of space, such as connectedness, continuity and boundary, while in qualitative spatial reasoning, the focus is on mereotoplogy [5].

In the Region Connection Calculus, regions are the basic spatial entities and relationships between spatial regions are defined in terms of the binary relation C(x; y),meaning spatial entity x connects with spatial entity y, which is true if and only if the closure of region x is connected to the closure of region y, i.e. if their closures share a common point[7]. Using the relation C, many versions of RCC could be found for instance RCC1, RCC2, RCC3, RCC5, RCC8, RCC15, and RCC23. The most common used and researched version is RCC8, which defines the following eight Jointly Exhaustive and Pairwise Disjoint (JEPD) relations : disconnected (DC), externally connected (EC), partially overlaps (PO), equal (EQ), tangential proper part (TPP), nontangential proper part (NTPP), tangential proper part inverse (TPPi) and nontangential proper part inverse (NTPPi) [8]. The intended meaning of these relations is illustrated in table (1).

B. Reasoning using RCC Relations

Since events are spatio-temporal entities, it is natural to use spatio-temporal reasoning to reason about the location of events. Studying how people report about the location of events, we notice that qualitative knowledge is used to express the event location as could be seen from the following example:

Event 1 : 8 Palestinians are arrested across the West Bank

Event 2: Thursday eight Palestinians arrested from Jerusalem, Jenin and Hebron, according to local and security sources.

In these two events , the event location is expressed using different qualitative representations which are used with different levels of granularity and expressiveness. When performing reasoning about the location of the two events, we may need to know if West Bank contains Jerusalem, Jenin and Hebron. Other aspects of event locations are usually described qualitatively, such as distance, orientation and topology.

Furthermore , There are many places that share the same or similar names (“AL-Tireh” :a neighborhood in Ramallah city;“AL-Tireh” : a Village in Ramallah region and “AL-Tireh”: a village north of Jenin city) . Also some places have multiple names(e.g. AL-Manarah square is also called Lions square). Some places are called after the most famous point of interest found near that place.

With RCC we can reason if two events have the same location by using the connection relations as explained in the

TABLE I.
DEFINING RCC8 RELATIONS[10][13]

Name	Symbol	Relation	Meaning	Definition
Equals	EQ	EQ(x,y)	X is identical with y	X = Y
Disconnected	DC	DC(x,y)	X is disconnected from	X ∩ Y = ∅
Externally Connected	EC	EC(x,y)	X is externally connected to y	i(X) ∩ i(Y) = ∅, X ∩ Y 6= ∅
Partially Overlap	PO	PO(x,y)	X partially overlaps y	i(X) ∩ i(Y) 6= ∅, X 6⊆ Y, Y 6⊆ X
Tangential Proper Part	TPP	TPP(x,y)	X is tangential proper part of y	X ⊂ Y, X 6⊆ i(Y)
Non-Tangential Proper Part	NTPP	NTPP(x,y)	X is non-tangential proper part of y	X ⊂ i(Y)

following rules :

Disconnected: Since one event cannot take place into two separate locations , and we have two events with disconnected locations, we can deduce that these are two different events .

$$\frac{(e_1\ in\ x) \wedge (e_2\ in\ y) \wedge (\boldsymbol{DC}(x,y))}{e_1\ \not\cong\ e_2}$$

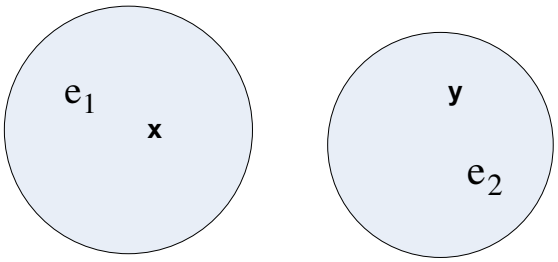


Fig. 1 Disconnected regions

Equal: If the two regions are equal , then at least one condition is met in the matching criteria, therefore it is possible that these two events are matched.

$$\frac{(e_1\ in\ x) \wedge (e_2\ in\ y) \wedge \left(\boldsymbol{EQ}(x,y)\right)}{e_1\ \cong\ e_2}$$

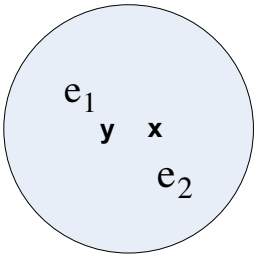


Fig. 2 Equal regions

As in the following two events, if we know that Radio street and Al-Ersal street are the same street from our knowledge base then this condition is met.

Event 3 : On 11 May 13, 11:14 hrs, reportedly, a car accident was reported in Radio street

Event 4 : On 11 May 13, 11:18 hrs, a car accident was reported in Al-Ersal street

Externally Connected: with externally connected regions, there is a possibility that the two events are taking place at the border of these two regions, therefore it possible that these two events have equal location and therefore a possible match.

$$\frac{(e_1\ in\ x) \wedge (e_2\ in\ y) \wedge \left(\boldsymbol{EC}(x,y)\right)}{e_1\ \cong\ e_2}$$

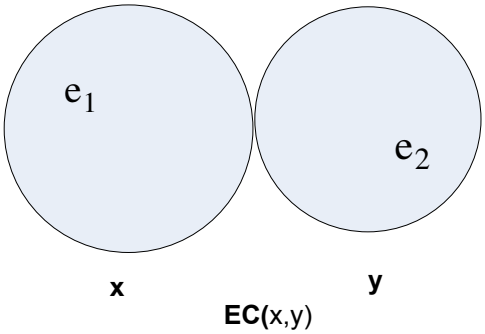


Fig. 3 Externally Connected regions

Event 1 : On 31 Mar 13, 0930 hrs, approximately 40 people demonstrated at DCO Beit-EL, NE Ramallah. It ended peacefully at1440 hrs.

Event 2 : On 31 Mar 13, between 0945-1200 hrs, families protested near City Inn Hotel, NE Ramallah against prisoners conditions.

Non tangential proper part: The semantic of the non tangential proper part is that region R1 is totally inside region R2 and that they are not equal and do not share any border.

$$\frac{(e_1\ in\ x) \wedge (e_2\ in\ y) \wedge \left(\boldsymbol{NTTP}(x,y)\right)}{e_1\ \cong\ e_2}$$

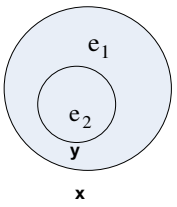


Fig. 4 Equal regions

Event 1 : A house in fire , in Jaffa street , second floor , near store AL-Manara close to AL-Families park

Event 2 : A smoke is seen ,near supermarket AL-Manara in Ain-Munjid area.

In these two events , Al-Families park is located in Ain-Munjid area.

Tangential proper part: in TTP relations, there might be more than two regions involved in the event. If x,y, and z are regions then y and x might be connected through a TPP , also y and z might be connected through a TPP**.**

$$\frac{(e_1\ in\ x) \wedge (e_2\ in\ y) \wedge \left(\boldsymbol{TTP}(x,y)\right)}{e_1\ \cong\ e_2}$$

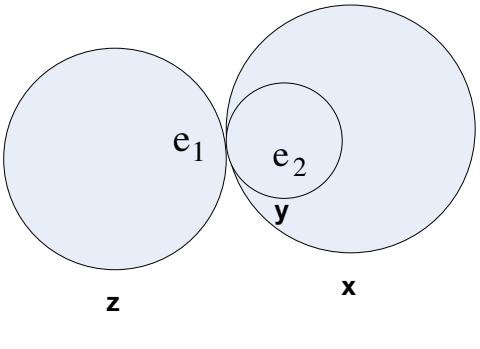


Fig. 5 Different TPP relations

Event 1 : A teen is injured in clashes near Jerusalem
Event 2 : A 17-year-old student was injured Thursday morning during clashes in the town of Abu Dis.

A main advantage for using RCC to reason about the location of events is that as examined by [9] , RCC is structurally similar to the way people reason about space and is a model of people's conceptual knowledge of spatial relationships).

III. Calculating RCC Relations

A. Region Ontology

In this paper all the examples are taken for events located in populated places. A populated place is an area of land inhabited by people. Therefore cities, villages , hamlets , towns, townships ,etc. are type of populated places. By definition, what mainly characterize an entity from another is its area. It is common to find the following definitions: a village is small human settlement, or a city is a large settlement and a hamlet is just a few dwellings[12]. Location and regions are more important for our work, however places are sometimes used to describe a region by its functional place like “city center”. A city center is a circle on a map to indicate the center of the city and it is only perceived by the human mind.

We have noticed that the three themes of geography (location, place and region) are used to describe where an event occurred or is happening. An observer uses `relative location` to describe the event when the observer is not familiar with the area. Also `absolute locations` are used when the observer knows the address of the event. Functional locations such as ‘city center’ or `formal` name such as ‘ name of the city’, or `vernacular` region such as ‘at the south area of the city’ are all used to describe an event location.

To model our regions, we use a region ontology where the country regions are classified into populated places and administratively declared places as shown in Figure

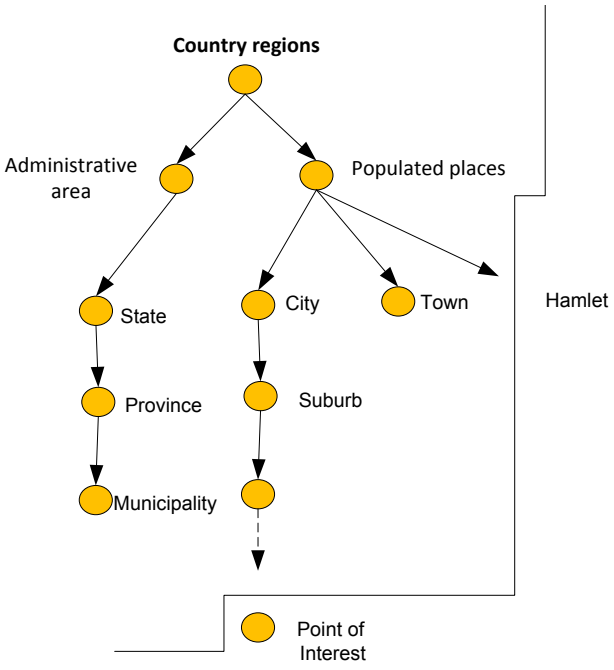


Fig. 6 Classification of a Country Region

Populated places are classified into `extended entities` such as city and `non-extended` entities such a point-of-interest. All populated places are disjoint classes and are continuous and have no holes. Suburbs and neighborhoods are part of a larger entity and is represented in one of the following forms :

NTPP: a suburb(S) has an NTTP relation with a town (T) if a suburb lies in a town and shares no border with it. The relation is denoted by S NTTP T

TPP: a suburb(S) has a TTP relation with a town (T) if a suburb lies in a town and shares borders with it. The relation is denoted by S TTP T

EC and **DC ,** this relation holds between suburb of a larger entity such as a city or town.

B. The Algorithm

The proposed methodology for calculating RCC between geographical regions is to approximate the exact region tiles by circular tiles as shown in Table (2). In the case of a country regions, the frame of reference is the partition of the country into cells which share boundaries but do not overlap. RCC relationships could then be calculated by using the longitude and latitude of the region as the center of the cell and then calculating the distance between cells.

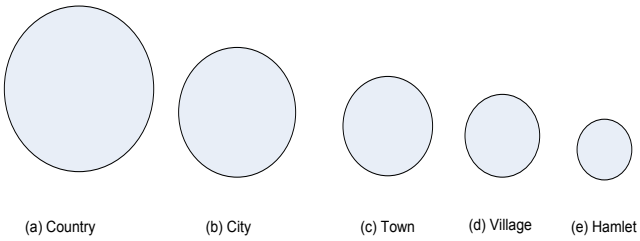


Fig. 7 Region Classification based on approximate area size

The difference between each type is identified by a set of features specially the size of the region. By comparing the distance between center of the cells and the reference distance, we can calculate the following relations:

Disconnected (DC): if two cells, R1 and R2, share no border then the relation between them is denoted by R1 DC R2. This is calculated using the following formula

DISTANCE(R1, R2) > (2* α + c) ; α denotes a constant that represents the maximum radius of a town and c denotes an error margin constant

Externally connected (EC): if two regions, R1 and R2, share borders then the relation between them is denoted by R1 EC R2.

DISTANCE(R1, R2) < (2* α + c)

Equals (EQ): the relation between each town, or any other location type, and itself is denoted by R1 EQ R2.

DISTANCE(R1, R2) < c

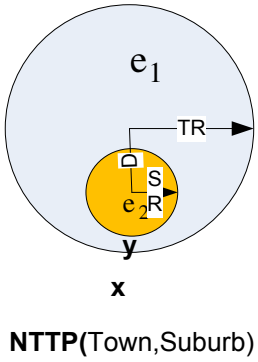
Both DC and EC relations are bidirectional. The algorithm is basically divided into three main parts : (1) calculates relations between town or cities (2) calculates relations between suburbs and towns (3) calculates relations between point of interests and suburbs or town with no suburbs. Following pseudo code illustrates how to calculate relations among towns in a country.

Pseducode for RCC8 Relations among towns/cities

```
Declare region Radius α // represents the maximum radius in meters
Declare c //  denotes an error margin constant defined in meters

Input region dataset containing longitude, latitude, place name
TOWNS_SET = FIND_ALL_LOCATIONS_BY_TYPE ("TOWN")
POIS_SET = FIND_ALL_LOCATIONS_BY_TYPE ("POIS")
SUBERBS_SET = FIND_ALL_LOCATIONS_BY_TYPE
("SUBERB")
BEGIN
   Build RCC8 Relations among  towns
   Build  RCC8 Relations among Suburbs and Towns
   Build RCC8 Relations among places of interests, towns and suburbs
   and towns
END
Output  set of Relations between all regions {EQ,DC ,EC} //
same type.
```

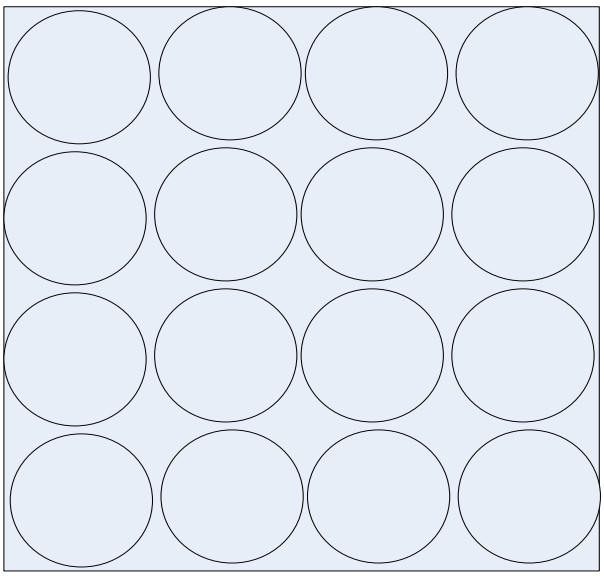
C. Building Town Suburb Relations



NTTP(Town,Suburb)

Fig. 8 NTTP(Town, Suburb)

TABLE II.
(A) APPROXIMATION USING CIRCULAR TILES (B) EXACT REGION TILES



(a) Approximation using circular tiles

(b) Exact region tiles

The second part of the algorithm is concerned with building the relations between towns and suburbs.

TownRaduis > Distance + SuburbRaduis + Constant

This part of the algorithm try to build the town or city suburbs only based on the input data which are are the lat,lon and suburb name.

D. Building Suburb- Suburb Relations

Building the suburb-suburb relations , follows the same approach for towns except we limit the comparison among a city or town suburbs.

Externally connected (EC): if two regions, S1 and S2, share borders then the relation between them is denoted by S1 EC S2.

Equals (EQ): the relation between each suburb, or any other location type, and itself is denoted by S1 EQ S2.

DISTANCE(S2, S2) < c

E. Building POI Relations

Specifying the relations among points of interests, villages, suburbs and towns: a point of interest can be located either in a village, suburb (town). The set of relations are all the relations listed above considering the semantics and the context of point of interests. At this stage we are mainly considering the NTPP relation between a point and a region (suburb,village and town).

NTPP: a POI(S) has an NTTP relation with a town (T) if a POI lies in a town and shares no border with it. The relation is denoted by S NTTP T

IV. Experimental Results and Validation

country region. All experiments were conducted on an Intel(R) Core(TM) i7 2.00 GHz running 32-bit Windows 7 Operating System with 4 GB of RAM.

A. An illustrative example

To build the data set for this experiment, we used Palestinian regions . We collected the shape

Figure 6 –

Fig. 9 Map of a region from a shape file

files from different municipalities like the one in Figure (6) and loaded the shape files into **PostGIS**/PostgreSQL database using the right coordination system for the selected region. The total spatial entities for this experiment is 5957 entity classified as shown in table ().

TABLE II.

SPATIAL ENTITIES PER TYPE

Type	Count
locality	144
hamlet	23
village	323
pois	5337
suburb	39
region	7
town	81
Border Crossing	1
city	10

The challenging question at this point is how to select the best radius for each region type. Obviously the algorithm will produce wrong results if the radius is chosen too small or too large. In order to select the best radius, we created a visual map that can help the user to select the best radius. As shown in figure (7) , choosing a radius of 800 meter will create more relations than 400 meters. Also we enhanced the algorithm by considering the area of the region. If the area of the region is found, then we can calculate the radius using the formula Area = sqrt((Area)/3.14) and thus we can get more reliable address

B. Validation of results

To develop our ground truth database for region relations, we had to build up the relations manually from existing maps. The ground truth data might include attribute data about the area size or population size of the region. However, not all towns or cities have these attributes filled. At this moment, we manually built the EC relationship between all towns, cities and villages. Also suburbs relations were built for two cities. Point of interests relations with their suburbs are built for nine suburbs.

The results are validated by computing precision , recall as shown in table

TABLE IIII.

EC PRECISION AND RECALL PER REGION

Region Id	**Total EC Relations (GT-Expert)**	**EC relations Built By System**	**EC relations Built (True) <true positive>**	**EC relations missed**	**EC relations Built (false) < false positive>**
94	8	6	5	3	1

91	6	2	2	4	0
236	8	9	7	1	1
196	9	8	8	1	0
177	6	8	6	1	2
…..					

For each region , we calculate the EC relations manually as shown in column (2). Column-3 represents the total relations built by the algorithm for each region ; column-4 shows how much of the calculated relations are true; column-5 shows how much relations are missed and column-6 represents how much relations are false.

- A – Number of relevant relations not retrieved
  B – Number of relevant relations retrieved
  C – Irrelevant relations retrieved

$$Precision = \frac{|B|}{|B|+|C|};$$

$$Recall = \frac{|B|}{|A|+|B|};$$

Precision = 0.82926829

Recall = 0.90265487

C. Discussion of results

Since the approach relies on approximating the area using a circle region. Selecting the radius (R) might produce wrong results as shown in the following cases. When the radius R is much smaller than region radius (RR) (R << RR), the algorithm creates no relations between the two regions. This is equivalent to region A is disconnected from region B . This could be improved by using the area of the region to calculate the radius and overriding the estimated one.

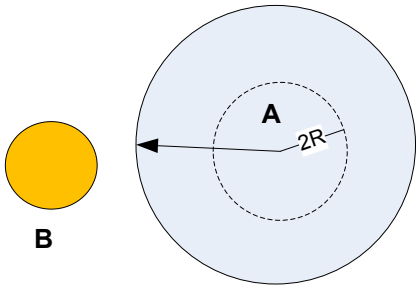


Fig. 10 City radius larger than double of selected radius

A second case occurs when the selected radius R is much larger than region radius (R >> RR)

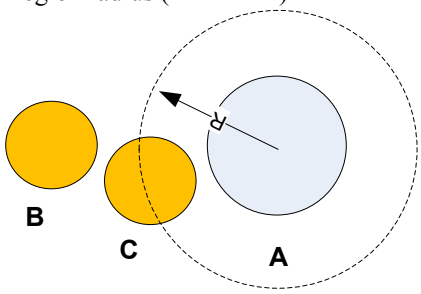
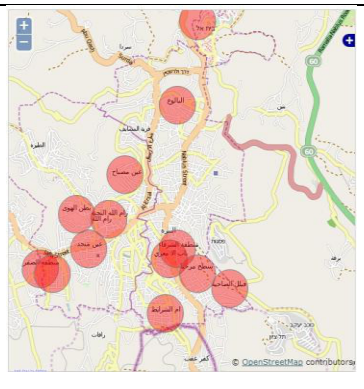


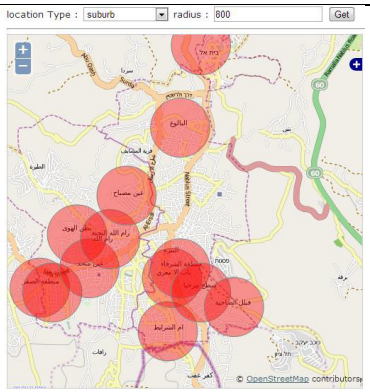
Fig. 11 City radius less than double of selected radius

When region radius is much lesser than selected radius , it is possible to make an EC relation with a region although there is another region in between. This is equivalent to having the following relations : a) A Externally connect to B b) A Externally connect to C c) C Externally connect to B

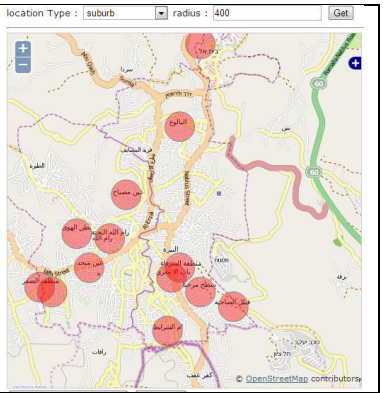
TABLE IIV.
DIFFERENT REGION RADIUS BASED ON VISUAL MAP



(a) Radius 500 meter



(b) Radius 800 meter



(c) Radius 400 meter

II. Conclusion

Using an automated method to build RCC relations between geographic regions is challenging especially if data has only attributes related to longitude and latitude. Although locating events is best done by its address, which is the more accurate among other methods like post address or boundary, the boundary approach in many rural areas is the only option available. However, from our experiments we found encouraging results. With such results it is now possible to use the new data set to find automatically the matching relationships between a pair of events such as in the two events presented earlier:

Event 1 : "8 Palestinians are arrested across the West Bank "
Event 2 : "Thursday eight Palestinians arrested from Jerusalem, Jenin and Hebron, according to local and security sources". Since Jerusalem, Jenin and Hebron has NTPP relationships with West Bank, we can infer that the location of these two events is the same.

However, we did not consider the issue of integrating topological relation with other binary relations such as distance constraints and directional constraints. In future research, the similarity measure between location A and location B is computed using the three types of constraints.

References

[1] J. M.. Zacks, B. Tversky, "Event structure in perception and conception" . Psychological Bulletin , 2001, 3-21

[2] D. Davidson, "The individuation of events", in Davidson ,1985 , p 179 doi: http://dx.doi.org/10.1037//0033-2909.127.1.3

[3] W. V. Quine,. "Events and reification", 1985

[4] J.Pustejovsk, "ISO-Space: The Annotation of Spatial Information in Language" , Proceedings of the Sixth Joint ISO - ACL SIGSEMWorkshop on Interoperable Semantic Annotation isa-6 ,2011

[5] A. G. Cohn and J. Renz, "Qualitative Spatial Representation and Reasoning. Handbook of Knowledge Representation", pages 551{596, 2008.

[6] B. L. Clarke. "A calculus of individuals based on \connection". Notre Dame J. Formal Logic, (22), 1981. doi:10.1305/ndjfl/1093883455

[7] J. Renz and B. Nebel. "Qualitative spatial reasoning using constraint calculi". In Handbook of Spatial Logics, pages 161{215. Springer, 2007. ISBN 978-1-4020-5586-7. doi:10.1007/978-1-4020-5587-4_4

[8] D. A. Randell, Z.Cui, A. G. Cohn," A Spatial Logic based on Regions and Connections". In B. Nebel, C. Rich, and W. Swartout (Eds.), Principles of Knowledge Representation and Reasoning. Morgan Kaufmann, San Mateo, CA, 1992, 165–176 DOI: 10.1016/j.artint.2008.10.009

[9] M. Knau, R. Rauh, J.Renz, "A cognitive assessment of topological spatial relations: Results from an empirical investigation". In Proceedings of the 3rd International Conference on Spatial Information Theory (COSIT'97),volume 1329 of Lecture Notes in Computer Science, pages 193{206, 1997 doi: 10.1007/3-540-63623-4_51

[10] Z. C. David, A. Randell, A. G. Cohn, "A spatial logic based on regions and connection" in 3rd International Conference on knowledge representation and reasonning, vol. 1, 1992, pp. 165–176.

[11] T. Bittner, J.G. Stell, "Approximate qualitative spatial reasoning" 1Department of Computer Science, Northwestern Universit doi: 10.1023/A:1015598320584

[12] http://vocab.org/places/schema.html

[13] J. Renz, "A canonical model of the Region Connection Calculus", in: Proc. 6th International Conference on Principles of Knowledge Representation and Reasoning (KR-98), Trento, Italy, 1998. DOI: 10.3166/jancl.12.469-494

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 307–312
DOI: 10.15439/2014F238
ACSIS, Vol. 2

An approach to discover false alarms in monitoring system of the copper mine

Bartłomiej Karaban, Jerzy Korczak
Wroclaw University of Economics.
ul Komandorska 118/120,
53-345 Wrocław, Poland
{bartlomiej.karaban,jerzy.korczak}@ue.wroc.pl

***Abstract*–The key task of telecommunication systems in deep mining is to ensure safety and continuity of production. These systems, despite modern and innovative infrastructure monitoring solutions, are not free from drawbacks. The occurrence of false alarms of damage to infrastructure is the practical problem which causes many negative effects, such as increasing the cost of the current operation of the system, information overload of operators, and service errors. In this paper a method for detecting false alarms in the communication system of the copper mine is proposed, presenting some rules that provide useful knowledge extracted from the database. A variety of experiments were carried out on real data from the telecommunications system operating in the copper mine KGHM Polska Miedź S.A.**

I. Introduction

SAFETY at work in the mine, good organization, and continuity of production requires an efficient and effective system of monitoring the state of the telecommunications installation, machinery, equipment and employees [1]. One of the important functions of a telecommunication network monitoring system is to collect and provide operators with information about the status of communication with network devices, the presence of voltage, time, location of failure, and values of parameters of the equipment, as well as the dangers of the failure [2]. For the system operator, information about the failure means loss of connectivity in excavations in which these devices are installed.

This article presents research focused on a well-known and difficult problem occurring in automated monitoring systems - the problem of identifying false alarms about the lack of communication with the device network infrastructure [3]. Let it be noted that the monitoring systems in the mine, and this fact is also confirmed by our study, always generate an alarm about the lack of communication in cases of actual loss [4].

The diagnosis of false alarms and minimizing their occurrence is one of the current problems in design and operation of automatic monitoring systems [5]. One of the eligibility criterion alarms is its duration. In practice, in cases of short duration of the alarm, operators treat it as a false alarm. This approach results in "anticipation" at the end of an alarm condition (restoration of communication), which causes delays in reporting failures to the service, making their duration longer [4]. A large number of alarms also cause adverse effects such as information overload for operators, the resulting operating errors and increase in the cost of the system operation.

The article discusses how the diagnostic system operates in one of the mining companies belonging to KGHM Polska Miedź SA. This system informs the operator about damage which has the status of a fault or breakdown. A fault is an event that causes a break in the functioning of communications. Any damage where the effect is different than the loss of communication we call a breakdown [4].

The aim of this publication is to present a method of automatic recognition and classification of alarms, allowing for the extraction of new, useful information and rules for the identification of false alarms generated by the monitoring system.

The following section describes the telecommunication system of the mine . Section 3 presents the method of data exploration. The results of experiments are discussed in the section 4. The experiments carried out on real data extracted from the system database are also presented.

II. The telecommunication network in the mine

In telecommunication systems operating in the mines of KGHM Polska Miedź SA, technical and organizational solutions were applied, ensuring a very high level of reliability of their work. The majority of the systems have very modern and innovative solutions. The most important of them are: a global mine communication system, a system of emergency propagation, and a radio communication system which where the database was used to discover false alarms [6].

The radio system used consists of the physical components and software packages that within a data communication system allow one to carry out the telecommunication functions, administrative services and diagnostics. The main element of the system is the Mine Station, which serves as a node that sets the connection, the monitoring application server and administrative point of connection of the terminals. The Mine Sta-

978-83-60810-58-3/$25.00

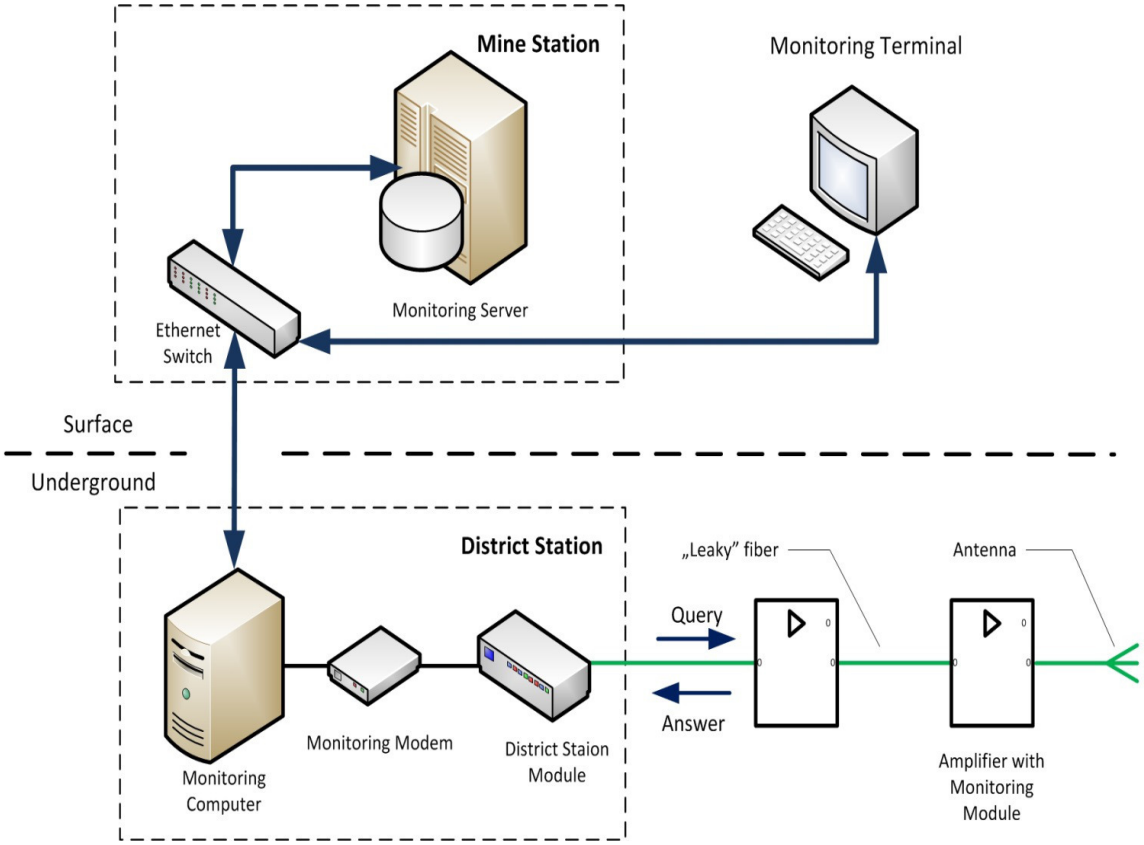


Fig 1. Diagram of the monitoring system

tion is connected with fiber optic links in District Stations installed at the bottom of the mine. District Stations are equipped with modules and modems monitorings that "interrogate" the network devices to determine their communication status and to control operating parameters. An "enquiry" about the "presence" of the device is sent approximately every 11 minutes; if there is no response, then an alarm is activated which indicates lack of communication with the device [4].

The described elements of the monitoring system are illustrated in Figure 1

The system described is a modern solution in terms of hardware and application, with a high degree of reliability dedicated to deep mines. A particular advantage is the graphical interface Monitoring Portal, allowing a quick and intuitive way to locate the location of the failure in the network. One of the interesting features of the system is the ability to "poll" the ad hoc device selected by the system operator, which greatly speeds up the execution of maintenance work. The Monitoring Portal map window is shown in Figure 2. As you can see, the color red is signaled by the failure of infrastructure elements for the lack of communication, information that can be read in the dialog caused by the operator by clicking on the device. Despite the advantages indicated in the monitoring system, it is not devoid of drawbacks. The most important one is the generation of false alarms about the lack of communication in the excavations.

III. THE METHOD OF EXPLORATION DATABASES DIAGNOSTIC

G. Piatetsky-Shapiro defines the process of data mining and knowledge discovery as "the process of nontrivial extraction of potentially useful and previously unknown information, or general patterns existing in databases" [7].

Referring to the specific problem of recognition and classification of alarms, it is possible to use two approaches well described in the literature: supervised or unsupervised classification. The main difference between them is *a priori* knowledge about the target class. Due to the possession of information whether a given historical alarm was true or false, it was desirable to apply a supervised classification.

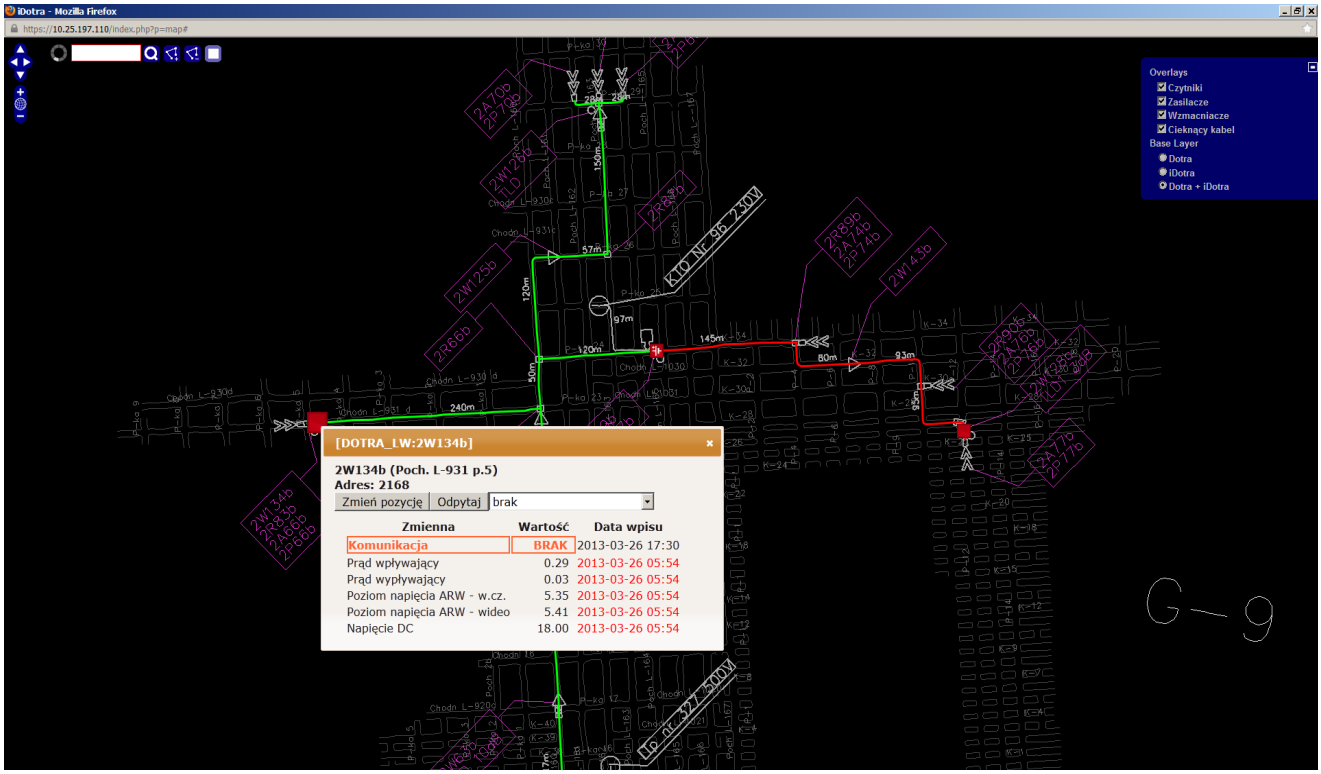


Fig 2. Map of the antenna system of the mining branch, with the window of the amplifier parameters (scale 1:7500)

There are many supervised classification methods, which include: logistic regression, discriminant analysis, neural networks, genetic algorithms, SVM, Naive Bayes, CN2, and inductive decision trees (IDT). It is not possible to use many of these methods, they are dedicated for the quantitative variables, while the alarm is mostly described using symbolic attributes. In contrast, methods such as SVM, Naive Bayes, and inductive decision trees do not have this restriction, therefore these methods were considered to be applied [8].

Many experiments and tests of several methods for classification were carried out in the project, amongst which we identified inductive decision trees. For choosing this method, several conditions appealed. The first condition resulted from a fundamental principle of inductive inference leading to the generalization of observations and facts in the form of rules and statements. An analyst who has domain knowledge should verify the authenticity of the generated rules and model the tree, as long as these rules are pragmatic enough to apply them to solve the problem. Another important advantage was the simplicity of interpretation derived rules in both graphical and decision rules. The last advantage, which prompted us to use tree induction, is the ability to control the complexity and generality of generated rules. The weakness of the IDT is the possibility of generating too large and too deep a tree, which might overfit and generate erroneous classifications [9].

There are many measures of evaluation of classifiers, such as sensitivity, precision, specificity or accuracy [10]. The studies demonstrate that the most important objective is to assure that all true alarms will be identified, then the number of false alarm would be minimized. In our approach the discovery will be focused on the minimization of the error of the first kind (False Positive rate), specifying the number of false alarms which were classified as true. The second important quality measure in this project was the precision, which takes into account the number of real alarms misclassified as false. In general, when choosing a classifier, we strive to achieve a compromise between readability and usability rules and maximizing the value of these measures. The first criterion is considered more important in the case of a small difference in the assessment of classification.

The process of data mining has been carried out according to the methodology CRISP-DM [2], using the data mining platform Orange. The data mining process is shown in Figure 3. The first step was the data preprocessing was performed (highlighted by orange color), part of which was done using MS EXCEL and MS ACCESS. The next step concerned the data analysis (box of brown contour), after which the model was built and applied to explore the data using the chosen methods of classification (box with a green outline). The last step was to evaluate the selected models (box with blue outline).

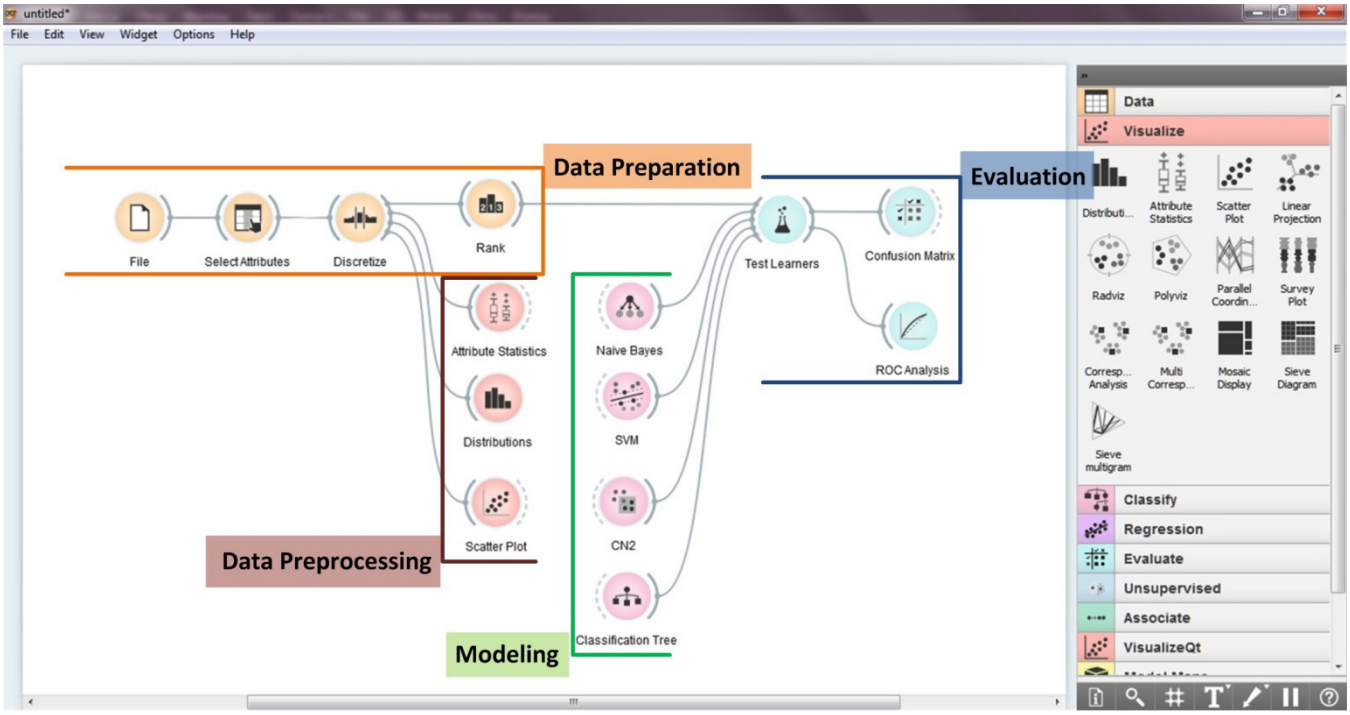


Fig 3. Schema of data mining process using Orange platform

The next section presents the process of building a model of a decision tree induction, and evaluation in comparison with other classifiers (SVM, Naive Bayes and CN2).

IV. Experimental Results

The analyzed data was real data extracted from: system monitoring, technical documentation, and system entries in the service log and saved in the Monitoring Database. The data set was randomly selected and contains 1316 observations from 01.05.2012 to 18.05.2012. In a first stage, there were necessary transformations of attribute values carried out in order to use the data in the selected classification algorithms. For example, one of the attributes ("duration") was discretized (by equal density intervals). The transformed data were rang using the Gain Ratio criterion, then the attributes selected that were used for modeling.

In order to recognize and classify false alarms a two-step approach was carried out. In the first step, the rules that classify the true alarms with the greatest possible "purity" were discovered. In the fig.4 this set of rules is named RulesTP. The algorithm is the following:

After inducing rules that cover all true alarms, in the second step the rules to discover false alarms were generated.

In fig.5 two sets of false are alarms are shown; the blue one illustrates the false alarms indicated by the current system, but the rose one shows the false alarms generated by the RulesTP.

As indicated and justified in the third section, the decision tree induction algorithm based on entropy (IDT) was chosen [11]. During the experiments about 100 variants of decision trees were generated by changing the value of various parameters such as the tree pruning ratio and the number of class attributes describing the duration of the alarm.

In the first series of experiments, the number of class attributes was specified: "duration" to 10 (the maximum for the reduction in the package Orange). The controlled parameter was the minimum number of observations in the leaf, which value after 10 experiments was set to 5.

A tree was generated in which the 6 leaf nodes of

```
let the RulesTPSet be empty
  repeat
    Find_BestRule(TrueAlarm)
    if the BestRule is not nil
      then
        let the TrueAlarmsDiscovered be the observations
              covered by the BestRule
        remove from the TrueAlarms the observations in the
              TrueAlarmsDiscovered
        append to the RulesTPSet the rule
  until the TrueAlarms is empty or the BestRule is nil
return the RulesTPSet
```

Fig 4. Text of rules named RulesTP

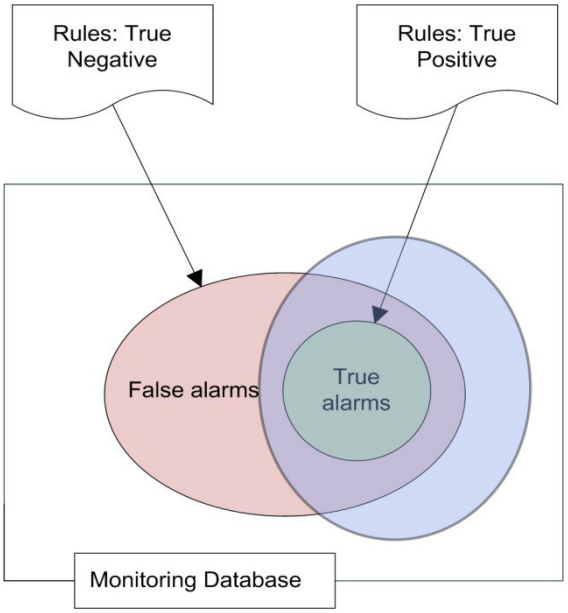


Fig 5. Schema of alarms identification by the rules

attribute "duration" contained only observations of false alarms were aggregated, which resulted in the reduction in their number from 10 to 5. The indicated change negatively affected the quality of all four classifiers. The obtained specificity ratio (TN) worsen the value than in previous experiments.

In the last series of experiments, the parameter to halt the tree construction has been changed. Just as in the first series of experiments, during the test the number of observations in the leaf was monitored, at which point the construction of the tree was stopped. The values of the parameter ranged from 1 to 10. Satisfactory results in terms of measures of quality and usability of the generated rules were obtained for the trees with minimum two observations per leaf.

In about 100 experiments, hundreds of decision rules were generated. The article presents only four of the ten rules that provide useful, previously unknown knowledge about the false and true alarms generated by the monitoring system. One of these rule was the following:

IF
alarm_ occurrence(t) = ‘stoppage’
∧ alarm_retreat_miners_work(t+1)= ‘miners_overlay’ v ‘miners_relay’
THEN true_alarm

The indicated rule covers 39 cases of "purity" of 100%. After analyzing the cases, it turned out that the rule identifies alarms which arose as a result of switching off the electrical switchboard during the weekend. Application of this rule increased besides of efficiency of alarm recognition indicated the switchgears which due to the communication requirement that should be maintained in a continuous operation. This information was forwarded to the electrical services at the mine.

The next rule that identified the true alarm was more complex, namely:

IF
alarm_occurence(t) = ‘mining’
∧ alarm_duration> 5 hours and 30 minutes
∧ alarm_occurence_miners_work(t) = ‘blasting_work’
THEN true_alarm

The rule allows to identify the cases where the loss of communication occurs most likely as a result of conducting blasting work. This rule covers only three cases (the "purity" of leaf 100%), however, taking into account the information it provides, we can identify the situation in which there is damage to the telecommunication line because of conducting blasting work. Such information provided to service may reduce the incidence of such situations; it helps to reduce system failures and the costs associated with damage to the network. Duration of alarm in the last o rule may seem too long. Such situations are due to the specific nature of the work environment in the mine (climatic and geological conditions), work organization (logistics, time to arrive to the place of accident), and the difficulties encountered in removing failure (signs of non-access to dangerous places).

Having analyze the generated rules, it was noted that in each of the created trees after the aggregation of classes of "Alarm Duration" (third series of experiments), there is the following "strong" rule of "purity" of 96.5% covering 800 alarms:

IF
alarm_occurence(t) = ‘mining’
∧ alarm_duration ≤ 45 minutes
THEN false_alarm

This rule is a useful new piece of knowledge about the alarm time threshold below which an alarm can be considered with high probability to be false. The value of this threshold is approximately 45 minutes..

The next rule classifies the alarms of long duration:

IF
alarm_occurence(t) =’mining’
∧ alarm_duration > 5 hours and 30 minutes
∧ alarm_occurence_miners_work(t) = ‘miners_relay’
∧ alarm_retreat(t+1) = ‘stoppage’
THEN false_alarm

It should be noted that in practice it is particularly difficult to determine the authenticity of such alarms.

The rules provided provide useful knowledge about a false alarm of long duration. As indicated in the introduction, operators often use heuristics when trying to determine the veracity of a given alarm. So far it has been assumed that the false alarm takes no more than two hours. Based on the first rule, it can be induced

that in 45 cases the alarms of about 5.5 hours duration were the false alarms ("purity" of the leaf = 100%), which exemplifies the existing heuristic assumptions.

V. CONCLUSIONS

Telecommunication systems in the mining industry play a key role in terms of safety, good organization and continuity of production. That is why it is necessary to monitor the state of the telecommunication infrastructure. Despite modern and innovative solutions discussed, the monitoring system is not free from inconveniences.

This paper proposes a method for the detection of false alarms about the lack of communication in the mine and the acquisition of new, useful knowledge from data. In about 100 experiments controlling the initial conditions of detention tree construction and the number of classes of the attribute "Alarm Duration" using the inductive decision trees, information about the duration threshold value has been found below which the alarm can be considered objectively false. Another very important and previously unknown piece of information is to identify individual devices that are deprived of the power supply in the days when there is no mining work, resulting in a loss of communication. Another potentially useful rule applies to identification of devices that can be damaged by blasting work.

The results obtained give encouragement to undertake further research on the diagnosis and extension of alarm recognition, signaled by the monitoring system in the mine.

REFERENCES

[1]. L. Tong, Y. Dou, "Simulation study of coal mine safety investment based on system dynamics", *International Journal of Mining Science and Technology*, vol. 24, March 2014, pp. 201-205.

[2]. S. Ding, *Model-Based Fault Diagnosis Techniques. Design Schemes, Algorithms and Tools*. Springer: London, 2013, pp. 4.

[3]. J. Korbicz, J. Kościelny, Z. Cholewa, *Diagnostyka procesów. Modele. Metody sztucznej inteligencji. Zastosowania.* WNT: Warsaw, 2002, pp. 11.

[4]. B. Karaban, *Indukcyjne drzewa decyzyjne w analizie alarmów systemu telekomunikacyjnego –Master Thesis*. Wroclaw University of Economics: Wroclaw, 2013.

[5]. Ch. Li, Xin Zhang, Xin Liu, "Mine safety information technology in the framework of Digital Mine", *Safety Science*, vol. 50, April 2012, pp. 846-850.

[6]. J. Korczak, B. Karaban, „ Metoda wykrywania fałszywych alarmów w systemie monitorującym sieć telekomunikacyjną kopalni", *Przegląd Górniczy*, vol. 70, no. 5, May 2014, pp. 108-112.

[7]. G Piatetsky-Shapiro, W. Frawley, *Knowledge discovery in databases*, The AAAI Press: Menlo Park, 1991.

[8]. T. Rivas, M. Paz, J. E. Martín, J. M. Matías, J.F. García, J. Taboada, "Explaining and predicting workplace accidents using data-mining techniques", *Reliability Engineering & System Safety*, vol. 96, July 2011, pp. 739-747.

[9]. M. Bramer, *Undergraduate Topics in Computer Science. Principles of Data Mining*, Springer: London, 2013,pp. 121-122.

[10]. T. Morzy, *Eksploracja danych. Metody i algorytmy*, PWN: Warsaw, 2013, pp. 326-327.

[11]. F. Gorunescu, *Data Mining. Concepts, Models and Techniques*, Springer: Berlin, 2011, pp. 165-166.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 313–321
DOI: 10.15439/2014F94
ACSIS, Vol. 2

Virtual Reality for Fire Evacuation Research

Max Kinateder
Department of Cognitive, Linguistic, and Psychological Sciences
Brown University, Box 1821
Providence, RI 02912, USA
Email: max_kinateder@brown.edu

Enrico Ronchi & Daniel Nilsson
Department of Fire Safety Engineering and Systems Safety
Lund University
P.O. Box 118, 22100 Lund, Sweden
Email: enrico.ronchi@brand.lth.se, daniel.nilsson@brand.lth.se

Margrethe Kobes
Institute for Safety
PO Box 7010
6801 AH Arnhem
The Netherlands
Email: Margrethe.Kobes@ifv.nl

Mathias Müller & Paul Pauli
Department of Psychology I
University of Würzburg
Marcusstr. 9-11, D-97072 Würzburg, Germany
Email: m.mueller@psychologie.uni-wuerzburg.de, pauli@psychologie.uni-wuerzburg.de

Andreas Mühlberger
Department of Psychology, Clinical Psychology, and Psychotherapy
University of Regensburg
Universitätsstr. 31, D-93053 Regensburg, Germany
Email: Andreas.Muehlberger@psychologie.uni-regensburg.de

***Abstract*—Virtual reality (VR) has become a popular approach to study human behavior in fire. The present position paper analyses Strengths, Weaknesses, Opportunities, and Threats (SWOT) of VR as a research tool for human behavior in fire. Virtual environments provide a maximum of experimental control, are easy to replicate, have relatively high ecological validity, and allow safe study of occupant behavior in scenarios that otherwise would be too dangerous. Lower ecological validity compared to field studies, ergonomic aspects, and technical limitations are the main weaknesses of the method. Increasingly realistic simulations and other technological advances provide new opportunities for this relatively young method. In this position paper, we argue that VR is a promising complementary laboratory tool in the quest to understand human behavior in fire and to improve fire safety.**

I. Introduction

STUDIES on fire evacuation seek to understand how occupants react when they are confronted with fire emergencies. Various disciplines, such as safety engineering, computer modeling, human factors, and psychology contribute to this field of research, all aiming to better understand human behavior in fire (HBiF) and ultimately to improve safety. One of the biggest challenges in this field is the access to ecologically valid and at the same time experimentally controlled empirical data (see for example references [1, 2]). Researchers in HBiF need safe, objective, reliable, and valid methods of data collection. The scope of this position paper is to discuss how virtual reality (VR) studies complement other well established research methods, such as case studies, unannounced drills, field studies, laboratory studies, and hypothetical studies [3-6].

The present article seeks to analyze **S**trengths, **W**eaknesses, **O**pportunities, and **T**hreats (SWOT) of VR research on HBiF. SWOT analysis originates in the management literature and has been applied to VR in the context of rehabilitation research [7].

I. VR in Fire Evacuation

VR has been defined as a "real or simulated environment in which the perceiver experiences telepresence" (the feeling of being present in a virtual environment) [8]. Note that this very wide definition implies that VR is not limited to computer generated environments or any specific technology. In a way, a real world laboratories can also be seen as virtual environments. However, for the scope of this article, VR refers only to computer generated simulations. The experience of telepresence comprises the illusion of being in the place displayed by the VR technology, and the illusion that events happening in the virtual environment (VE) are plausible and real [9]. Note that this definition does not imply the use of any specific technology. However, VR systems generally use computer generated visual and auditory simulations to immerse participants into a VE. Although less immersive systems – such as simulations on desktop computers [10] – can be used to study HBiF, the present paper mainly addresses highly immersive VR systems using CAVE (Cave Automatic Virtual Environment) systems, Powerwalls, or head mounted displays (HMD). These systems allow the presentation of highly realistic interactive visual and auditory stimuli to participants. Enhanced multimodal systems extend VR with olfactory [11, 12] and proprioceptive stimuli like wind, heat, or motion [13].

Paul Pauli and Andreas Mühlberger are shareholders of a commercial company that develops virtual environment research systems for empirical studies in the field of psychology, psychiatry, and psychotherapy. Mathias Müller is executive officer of the same company. No further potential conflicting interests exist.

The presentation of this paper was supported by the German Academic Exchange Service (DAAD).

VR has become a well-established method in other research fields such as traffic behavior [14], and clinical, social, or experimental psychology, e.g., for the psychotherapy of phobias [15], post-traumatic stress disorder [16], and for rehabilitation [7]. However, the usefulness of VR for HBiF is still under discussion and the method needs to be validated. *Ecological* validity can be assumed, if participants show similar behavioral, emotional, cognitive, and psychophysiological reactions in VR and in the real world [17]. The extent of emotional responses to a real world or virtual laboratory scenario is not necessarily the same as one might expect in a real fire emergency. Ecological validity does not mandate that participants have to believe that a simulated fire scenario is real. In fact, perceptual input (e.g., a visual simulation) can elicit emotional reactions, such as fear reactions, even if participants know that what they see is a simulation [18]. These reactions are probably of a lower intensity, however, future research is necessary to shed light on this question in the context of HBiF. More importantly, VEs have to be designed in a way that the observations from participants' behavior allow valid conclusion for real world scenarios. One study found promising results by comparing participants' behavior in VR evacuation scenario with real world case studies [19]. Other studies found HBiF comparable in conventional laboratory and VR simulated tunnel emergency scenarios [20, 21]. Note that similarity of two forms of artificial experimental methods (VR and classical laboratory studies) does not warrant ecological validity. There are still not enough studies systematically comparing virtual and real HBiF. In related research fields, however, validation studies have repeatedly demonstrated the ecological validity of VR. For example, several driving simulator studies documented ecological validity of VR simulations in terms of driving behavior, as well as the ability to elicit adequate emotional responses to VR [22-24]. In addition, several validation studies of virtual driving simulators demonstrated similar behavior in the real and the virtual world [25-28].

So far, VR has been used in several studies on diverse aspects of human behavior in fire, such as evacuation from buildings [10, 19, 29-32], occupant behavior in road tunnel fires [33-35], fire training [36-40], and other areas of safety and security research [41-43].

If proven sufficiently valid, VR will be a promising route to gain objective and reliable insights in HBiF. Results from VR studies can be used to test theories of HBiF, verify and validate evacuation models [44], and be integrated into VR training measures [29, 45].

II. VR in Comparison to other Methods

Table I compares six different empirical research methods, which gather data on HBiF, on several important aspects such as the degree of experimental control, experimental setting, and the type of data that can be collected with each method.

In *hypothetical studies* have been used in evacuation research [46]. Participants are usually either shown videos or are instructed to imagine a certain scenario and then asked how they would react in that situation. Another example would be data acquisition from experts who evaluate the outcome of a given hypothetical scenario. These scenarios can be in the form of an interview or questionnaires. Data

TABLE I.
COMPARISON OF RESEARCH METHODS

	Hypothetical study	**"Classical" lab experiment**	**VR experiment**	**Field studies**	**Drills**	**Case Studies**
Setting	laboratory	laboratory	laboratory	real-world	real-world	real-world
Experimental control	yes	Yes (less than in VR)	yes	limited	no	no
Setting	laboratory	laboratory	laboratory	real-world	real-world	real-world
Type of data	subjective (statements from participants or experts)	subjective, objective (behavior & psycho-physiology)	subjective, objective (behavior & psycho-physiology)	subjective, objective (behavior)	subjective, objective (behavior)	subjective, objective (behavior)
Possibility of use of stressors	no (only hypothetical)	limited	limited	limited	limited	yes
Ecological validity	low	medium	medium	medium	high, if unannounced; limited, if announced	high
Possibility of adjusting experimental setting	yes	yes	yes	limited	no	no
Possibility of exact replication	yes	yes	yes	limited	no	no
Time and cost intensity for data collection	very low	low	low	high	medium	-
Automatic data collection possible	yes	yes	yes	limited	limited	no

from hypothetical studies is always subjective as they reflect the participants' personal opinion, knowledge, or experience. Subjective data, although being prone to bias, can be highly useful to gain insights into how occupants experienced an event or to reconstruct chains of events.

In *"classical" laboratory studies* real world scenarios are transferred into the controlled environment of a laboratory. Here, causal effects can be investigated with experimental methods by manipulating independent variables and measuring dependent variables (e.g., behavior, subjective data, and physiological data). Participants have to be assigned randomly into at least two experimental conditions for a true experiment which vary only in one condition (the independent variable). It is crucial that laboratory studies are ethical acceptable. That is, the experimenter may, for example, only use stimuli/stressors that are not actually harming the participant.

In *VR experiments,* participants can be confronted with simulated fire emergencies. Simulations of fire emergencies can be presented to participants in an extremely controlled way. VR experiments allow the convenient recording of behavioral and physiological data with a very high resolution as well as the collection of subjective data. In comparison to other methods, the presentation of stressors (e.g., flames) is ethically less critical compared to classical laboratory and field studies.

In *field studies,* emergency scenarios can be reenacted in a naturalistic setting outside of a laboratory. Unlike in laboratory settings, field studies are usually in less controlled environments (although certain infrastructures like road tunnels are highly controllable). Similar to classical laboratory studies and VR experiments, field studies use subjective and objective data (recorded behavior).

Drills are either announced or unannounced practice scenarios in real world settings. Although very similar to field studies, the focus of drills is usually on practicing emergency procedures. They allow the observation of occupant behavior under naturalistic conditions in a specific location. Similar to field studies, observational data and self-report data can be acquired.

Case studies refer to the descriptive, exploratory or explanatory analysis of a real fire emergency. Subjective self-report data from occupants and analysis of closed-circuit television footage can be used to reconstruct the events of a real emergency.

In addition, mixed methods may help to overcome limitations of individual methods. For example it is possible to modify participants experience in real world settings using augmented reality, or increase the immersiveness of a VR system by adding real elements (for example objects that participants can touch) to a VR study.

When planning studies on HBiF researchers have to consider certain factors and restrictions (See Table I). These include the necessary degree of experimental control, the choice of setting (laboratory or real-world), or the type of data required (subjective vs objective) and whether or not it is important to be able to adjust or replicate the experimental scenario during data acquisition. There are also factors related to the efforts necessary for the realization of a study. Efforts can be financial (e.g., costs for hard and software, personnel, participant recruitment, or lab space in VR experiments) but also whether or not data can be collected and processed automatically (e.g., with tracking devices) or has to be extracted from video footage.

These comparisons do not necessarily reflect strengths or weaknesses, rather factors that should be considered when deciding on which research method is most suitable for a certain research question. The methods discussed here do not provide the best solution for every research issue. There are arguments for and against the use of each method to address specific concerns. In this section, the VR studies are analyzed with respect to key aspects of a research question.

II. SWOT Analysis

SWOT analysis refers to the analysis of **S**trengths, **W**eaknesses, **O**pportunities and **T**hreats of a given method or product [7]. The present SWOT analysis (Table II) aims to uncover internal strengths and weaknesses of VR as a research tool in HBiF and to identify its surrounding conditions (opportunities and threats). A detailed description of SWOT analysis can be found in reference [7].

- Strengths refer to inherent resources and capacities of VR helping to gather objective, reliable, and valid empirical data on HBiF.
- W*eaknesses* describe inherent shortcomings, limitations, and problems of VR to achieve its goal.
- *Opportunities* comprise surrounding conditions or trends from which VR research will potentially benefit and which is promising to overcome weaknesses.
- *Threats* are surrounding conditions which are detrimental to the use of VR as a research tool in HBiF and which need to be overcome.

I. Strengths

Internal validity is possibly the most important strength of VR studies. Entire VEs can be easily controlled. Stimulus control and experimental stimulus manipulation is a key feature in investigating cause and effect relations [47]. It is extremely difficult to impossible to control the environment in field studies, drills, and even classical lab studies. For instance, in VR smoke can be numerically calculated and then repeatedly presented in exactly the same way to several participants. "Real" smoke, even in the controlled environment of a classical laboratory study, will always vary, and consequently visibility conditions may change across observations. Lack of experimental control limits the reliability and consequently the internal validity of these methods.

Replication. VR studies can be replicated to the last detail, given the usage of the same or comparable equipment. One major criterion for empirical studies is that they can be/should be reproducible. Replication refers to the

repetition of a study using the same methods but different participants and experimenters. Studies need to be replicated in order to test their reliability and validity and to test their generalizability and the role of confounding variables. Real world studies, especially field and case studies, provide data for only one specific event and are extremely difficult to replicate.

Ecological validity refers to the degree with which the methods of a study represent the real world scenario that is being examined. VR offers a similar degree of ecological validity as classical laboratory studies, but depending on the research question one method or the other may be more suitable. For example, certain features of a fire emergency, such as the visual simulation of flames, may be simulated with higher control in VR but other features (e.g. touch) may be more difficult but not impossible to simulate in VR (e.g. using a mix of virtual and real elements). However, simulation of heat or olfactory cues is possible but still limited as it is both technically challenging to present olfactory stimuli in an experimentally controlled manner. Ecological validity of VR studies is higher than in hypothetical studies since the latter require the ability of participants to correctly imagine a scenario. High – but not absolute – ecological validity of VR studies can be assumed if the visualization, observed behavior, and task difficulty of a simulated fire emergency is realistic, i.e., based on valid models and representative of real world events. VR simulations can have the same degree of visual realism as simulations in classical laboratory studies.

All laboratory experiments including VR studies, however, are abstractions of reality and therefore some loss of ecological validity is inherent to the method in comparison to real events [47]. Even the most sophisticated field experiment and the most advanced VR simulation on human behavior in dangerous situations cannot (and should not) claim absolute ecological validity. Participants will always know that they are taking part in an artificial situation. However, this is true for all methods compared in the present article with the exception of unannounced drills and case studies. Knowing that one takes part in a study and/or that there is no real danger, may lead to systematic biases in participants' responses.

External validity refers to the question whether the results of a study can be generalized from the experimental setting to other situations or populations [48]. VR and other laboratory studies allow controlling confounding factors and thus studying general underlying effects in HBiF is possible. Results from uncontrolled studies (e.g., drills, case studies, and to some extent also in field studies) cannot not be generalized because confounding variables are not controlled.

Safety for participants. VR allows the controlled simulation of perilous scenarios, such as extreme tunnel fires, without putting the participants at risk of a physical harm. That is, VR studies are ethically less problematic than field studies since it is possible to simulate catastrophic and life threatening situations without risking to physical harm participants. However, there are also limitations for VR studies (see *Threats*).

Real-time feedback. Precise tracking of various parameters as well as the highly controlled visual input technologies allow to give participants and researchers immediate feedback of behavior, performance, and even psychophysiological processes. For example, task-performance or physiological parameters such as heart rate can be displayed online during trials. This allows the experimenter to have real time access to data. Real time feedback for participants can be used to test training measures (e.g., fire evacuation training).

Multi-modal simulations. In theory, simulation of any modality is possible. To date, combined simulation of visual and auditory stimuli are very well developed. Olfactory, nociceptive, or thermoceptive simulations are also possible, however, still less technologically advanced.

Precise measurement. Precise tracking technology allows accurate analysis of various aspects of participants' behavior (e.g., full body tracking, head movement, eye tracking) with extremely high sampling rates.

Psychophysiological monitoring. In addition to behaviors, psychophysiological parameters such as heart rate or skin conductance can be measured easily in a VR laboratory. Measuring physiological correlates of behavior while being immersed into a VE allows researchers to analyze emotional reactions (e.g., fear reactions) to simulated emergencies.

Low costs. Once a VR system is set up, it can be used, in theory, infinitely. Virtual scenarios can be re-used and easily modified. With the decrease in prices for hardware and software (some VR simulation software are even free to use), VR experiments have become more and more affordable. Although costs for individual studies vary significantly, VR studies are generally cheaper than field studies. However, setting up a complete complex VR laboratory such as CAVE systems is cost intensive and requires space but is relatively affordable to run.

Repeated measurements. Participants can easily be immersed repeatedly into VEs and repeated measurements in identical scenarios are possible. Recreating identical conditions is complex or even impossible with other methods (See also *Replication*). Repeated measurements can be used to test, for example, training measures aimed to improve HBiF.

Flexibility. Experimental settings in VR can be adjusted easily, allowing to run pilot studies and to quickly develop minor alterations of the experimental set-up.

Control of confounding variables. There are many variables that potentially confound the effect of a given independent variable but are not of primary interest (e.g., minor changes in starting positions, left/right turning preferences). These can be easily controlled in VR and laboratory studies but is difficult to impossible in other designs.

Independent of imagination abilities/willingness of participants. Producing highly immersive VEs reduces the variance in participants' response caused by individual

differences in the ability and willingness to imagine a given scenario. Hypothetical studies rely on the ability of participants to imagine a scenario. Here, researchers have no control of the ability and willingness of participants to imagine, for example, a fire evacuation from a high-rise building.

Participant recruitment. Although not an important strength of VR studies, it is worth mentioning that recruiting a sample for a VR study has less restrictions than recruiting for a drill or field study. In field studies, the experimental set-up is often only available on limited occasions and may be time and cost intensive to install. For example, certain infrastructures such as underground public transportation systems or road tunnels may only be accessible to researchers at a very limited time or during certain hours of the day making data collection difficult. Once a VR scenario has been set-up it can be repeated, in theory, at any given time which allows longer and more flexible time-windows for data collection.

II. Weaknesses

Need for confirmation/validation. To date, there are still not enough validation studies. These are necessary to test the assumptions that behavior in a simulated scenario can predict or be transferred to "real" HBiF.

Non-intuitive interaction methods. Although VEs are interactive, participants often need devices like gamepads to navigate in and interact with the virtual environment. This always reminds the participants that they are in an artificial scenario, and if the scenario is not well designed, may bias behavior. For example, evacuation times may vary depending on how well participants can handle their navigation device. However, developments in input devices may partially address this weakness (See also *Opportunities*).

Inter-individual differences in ease of interaction with VR. Depending on various factors, such as age or experience with VR, participants may have difficulties when using VR. For example, participants who have a lot of experience using 3D video games may find it easier to navigate in a VE. Participants, who have less experience with computers (e.g., elderly participants) may need longer practice sessions before they can navigate without limitations in a VE.

Technical limitations. Visual input as well as well as interaction methods are still limited. Although visual simulation of virtual environments has improved tremendously in recent years, the current simulations will always be recognized as such by participants. Such imperfections (e.g. in model rendering, spatial resolution, field of view (for HMDs), graphic update rate, lags between head tracking and visualization) of VEs may lead to artifacts [49]. Especially the simulation of behaviorally realistic virtual humans is still challenging. Another technological limitation is the need for interaction tools, such as game pads or HMDs, to immerse into and interact in the VE. For example, navigation, even in a highly immersive CAVE system is either limited to a few square meters or participants have to use interaction devices. These technical challenges may limit the immersiveness of a VR system and lead to a lower experience of presence.

Technology-induced side effects. Prolonged exposure to VR may cause symptoms of nausea and vertigo (simulator sickness; for a review on simulator sickness, see references [49, 50]). The incidence of these side effects depends upon characteristics of the VR system (e.g., display field of view, lag between tracking device and update of the visualization) and participants [51]. Such side-effects need to be considered when planning and evaluating the ethical innocuousness (e.g., participants need to be able to terminate the experiment whenever they want to).

TABLE II.
SUMMARY OF A SWOT ANALYSIS FOR VR IN FIRE EVACUATION RESEARCH.

Strengths	Weaknesses	Opportunities	Threats
• Internal validity • Replication • Ecological validity • External validity • Safety for participants • Real-time feedback • Multi-modal simulations • Precise measurement • Psychophysiological monitoring • Low costs • Repeated measurements • Flexibility • Control of confounding variables • Independent of imagination abilities/willingness of participants • Participant recruitment	• Need for confirmation/validation • Non-intuitive interaction methods • Inter-individual differences in ease of interaction with VR • Technical limitations • Technology-induced side effects • Efforts	• Intuitive and natural navigation • Graphical developments • Multi-modal simulation and feedback • Usability for researchers • Exchange of 3D-scenes or experiments	• Failure to show ecological validity • Ethical challenges • Side-effects due to interaction with other medical conditions • Misleading expectations • Technical faults

Efforts. Setting up a highly immersive VR laboratory is time and cost intensive. [49]. Creating plausible VEs is also complex (the differences between less immersive and simple environments and complex highly immersive VEs is extreme) and requires expertise with special hard- and software systems. Given the rapid developments in this type of technology, constant investment may be required to stay up-to-date.

III. Opportunities

Intuitive and natural navigation. Although, highly immersive VR systems, such as CAVE or HMD systems, allow participants to move freely with their whole body within the VE [52], even the most advanced systems still have movement restrictions and participants have to use navigation and input devices. Advances in tracking technology, innovative interaction devices, and VR systems that allow natural navigation (e.g., bigger CAVE systems or wireless HMDs and walking platforms) may reduce the weakness of limited space and non-intuitive interaction devices, e.g. [52]. These advances promise even better immersion into VEs and consequently improved ecological validity.

Graphical developments. The dramatic improvement of graphical simulations allows more and more photorealistic simulation of fire emergencies. In addition, numerical calculated fire and smoke have been successfully implemented into VR simulations [37]. Similar to the advances in navigation devices, improved realism of simulations will lead to increased experience of presence for participants and better ecological validity.

Multi-modal simulation and feedback. The integration of multi-modal simulations for visual and auditory simulation extended by kinesthetic, olfactory, haptic, thermoceptive simulation allows the simulation of more complete scenarios. For examples, see references [53, 54].

Usability for researchers. The widespread use of VR technology depends highly on its usability for researchers. Recent developments in easy to use VR tool kits make VR technology more accessible. The improved cross platform compatibility helps the use of VR over different platforms and operation systems.

Exchange of 3D-scenes or experiments. Researchers can easily exchange 3D models or even entire experiments with each other. This may foster cooperation between laboratories and also lead to the development of standard scenarios which could be used as references and thus increase comparability of different VR studies.

IV. Threats

Failure to show ecological validity. This is the biggest threat to VR as a research tool to study HBiF. Systematic validation of VR for HBiF has still not demonstrated its range of applicability. Future studies are clearly necessary to test the ecological validity of VR to study HBiF.

Ethical challenges. Scientific studies on HBiF have to comply with ethical standards such as the Declaration of Helsinki which define ethical standards for studies with human subjects [55]. Even though most participants are aware that a virtual fire provides no threat to them, some participants may still experience extreme fear. Just as with any other method, VR research needs to ensure that the experienced fear cannot lead to longer lasting difficulties for participants such as traumatization, especially if one has in mind that VEs are getting closer and closer in means of realism to real scenarios. In addition, a VR system that causes extreme side effects (e.g., seizures or strong nausea) would be ethically unacceptable.

If participants cannot differentiate between a simulated and a real scenario, which may be the case, for example, with small children, the same ethical restrictions as with other methods apply.

Side-effects due to interaction with other medical conditions. Some scenario for HBiF may be particularly risky in causing side-effects in interaction with pre-existing medical conditions. For example, studies using flashing lights may cause seizures in at-risk populations; patients with specific phobias (e.g., of tunnels or heights) may experience extreme fear; Simulation of fire emergencies may induce flashbacks in participants who previously have experienced a traumatizing event. Other methods, however, bear similar risks.

Misleading expectations. The expectation that VR experiments can completely replace real world tests and holistically covers all aspects of human behavior in fire is misleading. Similar to classical laboratory experiment, VR allows investigating general underlying processes of HBiF and testing specific aspects (e.g., the effect of safety installations on evacuation behavior). The conclusions from these studies may even lead to changes in the design of real world safety installations. However, HBiF is highly complex; one can never exclude that individual decision-making, behavior, and experience in a specific real scenario may differ significantly from trends found in VR studies.

Certain *technical faults* in the implementation of a VR system (e.g., jitter errors, discrepancies in simulation or tracking latency) even can reduce the immersiveness and even may increase side-effects like simulator sickness.

III. General Discussion

The present position paper provides a SWOT analysis for VR as a research tool to study HBiF. We provided an overview of various methods used in HBiF and systematically compared VR to these methods.

The biggest strength of VR is surely its ability to create highly immersive, externally valid, highly controlled, and safe experimental set-ups. The biggest weakness is the reduced ecological validity in comparison with field and case studies, as well as the lack of validation studies specifically for HBiF. These studies should compare VR experiments with the results of other laboratory experiments and field studies.

The diverse methods used to study HBiF always have to trade-off between ecological validity and experimental control. For instance, case and field studies in real world

settings provide almost perfect ecological validity. However, strict experimental control is impossible to achieve here, and financial and logistic efforts as well as ethical limitations need to be considered. Hypothetical studies need to consider less strict ethical limitations and are easier to realize, but rely heavily on the ability of the participant's imagination and are prone to response biases.

Field studies are often characterized by the combination of setting and participants, e.g., real world settings with participants that naturally are in these settings. In the evacuation area, this allows doing unannounced experiments. This can never be achieved in VR or other laboratory experiments, as participants need to be recruited and enter the VR-lab (or ask them to put on some equipment). At most, participants may be "deceived" by telling them that they will take part in one study and then exposing them to something else. However, this is easily feasible in classical laboratory studies but requires more efforts in VR studies. It can be argued that affects the external validity of a study.

The differentiation between ecological and external validity is important. Ecological validity refers to how good a research method represents reality. External validity describes how well study results can be transferred to other situations and generalized over populations. Whereas ecological validity is not, external validity is a prerequisite for the overall validity of a study. A study, can be ecologically valid (e.g., the results from an unannounced drill) but not generalizable to other settings, populations, cultures etc., if it lacks experimental control and, therefore, internal validity).

I. What can we study in VR?

VR can be used to design complex laboratory experiments on HBiF. It allows studying how occupants react to fire cues, such as flames or smoke; it allows collecting precise behavioral and psychophysiological data during controlled simulated events. Virtual scenarios can be designed with an extremely high level of detail. That way, we can use VR to study underlying processes of HBiF (e.g. phenomena like risk perception of occupants, social influence, architectural influences, way-finding abilities in smoke, etc.). That way, VR studies can contribute to a better understanding of HBiF.

In addition, evacuation concepts for large complex buildings can be tested in VR making it possible to identify potentially problematic evacuation routes *before* a new building is constructed. This is particularly useful since evacuation models implemented in simulation software tools still oversimplify HBiF (e.g., some models assume that occupants always take the shortest route to an emergency exit [33]).

It is important to note that VR cannot replace any of the other methods mentioned above but is complementary. VR studies can be used in experimental pilot studies in order to test a number of possible factors that may theoretically be influencing HBiF (e.g. various design aspects of safety equipment). Then, those factors deemed as the most important ones in VR can then be tested in field experiments or used to predict behavior in case studies or drills.

II. What can we not study in VR?

Virtual reality is not reality. Participants will always know that they take part in an artificial situation. It is impossible to generate situations in which participants' would risk actual physical harm. Extremely perilous situations may induce effects (e.g., extreme fear) which are not attainable with artificial scenarios, which in turn may affect behavior. Only observations from real events and to some degree unannounced drills may have this effect. It is impossible to investigate these parts of HBiF using VR laboratory studies.

III. Conclusion and positioning statement

We argue that VR is a powerful approach to study HBiF. VR allows shedding light on aspects of occupant behavior that were previously impossible to investigate under controlled conditions. Although we identified several weaknesses and limitations of the method, the most important one being the need for validation studies, it seems possible that these can be overcome, either by technical progress or by combining several different research approaches (triangulation approach). None of the state of the art research methods (including VR) are able to validly grasp all aspects of HBiF, and VR does not aim to replace any of the other presently established research methods. We see it as a promising complementary laboratory tool in the quest to understand HBiF and to improve fire safety.

IV. REFERENCES

[1] K. Fridolf, D. Nilsson, and H. Frantzich, "Fire evacuation in underground transportation systems: a review of accidents and empirical research," *Fire Technology,* vol. 49, pp. 451-475, 2013.

[2] M. Kobes, I. Helsloot, B. de Vries, and J. G. Post, "Building safety and human behaviour in fire: A literature review," *Fire Safety Journal,* vol. 45, pp. 1-11, 1// 2010.

[3] R. F. Fahy and G. Proulx, "A comparison of the 1993 and 2001 evacuations of the World Trade Center," in *Proceedings of the 2002 Fire Risk and Hazard Assessment Symposium*, 2002, pp. 111-117.

[4] D. Nilsson, H. Frantzich, and W. Saunders, "Coloured Flashing Lights to Mark Emergency Exits - Experiences from Evacuation Experiments," presented at the Fire Safety Science - Proceedings of the Eighth International Symposium, Beijing, China, 2005.

[5] T. Shields and K. Boyce, "A study of evacuation from large retail stores," *Fire Safety Journal,* vol. 35, pp. 25-49, 2000.

[6] P. Burns, G. Stevens, K. Sandy, A. Dix, B. Raphael, and B. Allen, "Human behaviour during an evacuation scenario in the Sydney Harbour Tunnel," *Australian Journal of Emergency Management, The,* vol. 28, p. 20, 2013.

[7] A. S. Rizzo and G. J. Kim, "A SWOT Analysis of the Field of Virtual Reality Rehabilitation and Therapy," *Presence: Teleoperators and Virtual Environments,* vol. 14, pp. 119-146, 2005/04/01 2005.

[8] J. Steuer, "Defining virtual reality: Dimensions determining telepresence," *Journal of communication,* vol. 42, pp. 73-93, 1992.

[9] M. Slater, B. Spanlang, and D. Corominas, "Simulating virtual environments within virtual environments as the basis for a psychophysics of presence," *ACM Transactions on Graphics (TOG),* vol. 29, p. 92, 2010.

[10] N. W. Bode, A. U. K. Wagoum, and E. A. Codling, "Human responses to multiple sources of directional information in

virtual crowd evacuations," *Journal of The Royal Society Interface,* vol. 11, p. 20130904, 2014.

[11] W. Barfield and E. Danas, "Comments on the use of olfactory displays for virtual environments," *Presence: Teleoperators and Virtual Environments,* vol. 5, pp. 109-121, 1996.

[12] E. Richard, A. Tijou, P. Richard, and J.-L. Ferrier, "Multi-modal virtual environments for education with haptic and olfactory feedback," *Virtual Reality,* vol. 10, pp. 207-225, 2006.

[13] F. Hülsmann, N. Mattar, J. Fröhlich, and I. Wachsmuth, "Wind and Warmth in Virtual Reality–Requirements and Chances," in *Proceedings of the Workshop Virtuelle & Erweiterte Realität 2013*, 2013.

[14] L. N. Boyle and J. D. Lee, "Using driving simulators to assess driving safety," *Accident Analysis and Prevention,* vol. 42, pp. 785-787, May 2010.

[15] K. Meyerbröker and P. M. Emmelkamp, "Virtual reality exposure therapy in anxiety disorders: a systematic review of process-and-outcome studies," *Depression and anxiety,* vol. 27, pp. 933-944, 2010.

[16] B. K. Wiederhold and M. D. Wiederhold, "Virtual Reality Treatment of Posttraumatic Stress Disorder Due to Motor Vehicle Accident," *Cyberpsychology Behavior and Social Networking,* vol. 13, pp. 21-27, Feb 2010.

[17] C. A. Anderson and B. J. Bushman, "External validity of "trivial" experiments: The case of laboratory aggression," *Review of General Psychology,* vol. 1, pp. 19-41, 1997.

[18] H. M. Peperkorn, G. W. Alpers, and A. Mühlberger, "Triggers of Fear: Perceptual Cues Versus Conceptual Information in Spider Phobia," *Journal of clinical psychology,* 2013.

[19] M. Kobes, I. Helsloot, B. de Vries, and J. Post, "Exit choice, (pre-)movement time and (pre-)evacuation behaviour in hotel fire evacuation — Behavioural analysis and validation of the use of serious gaming in experimental research," *Procedia Engineering,* vol. 3, pp. 37-51, 2010/01// 2010.

[20] F. Malthe and Vukancic, "Virtual Reality och människors beteende vid brand [Virtual Reality and human behavior in fire]," Lund University LUCATORG: 011033007, 2012.

[21] J. Johansson and L. Petersson, "Utrymning och vägval i Virtual Reality."

[22] A. Mühlberger, H. H. Bülthoff, G. Wiedemann, and P. Pauli, "Virtual reality for the psychophysiological assessment of phobic fear: responses during virtual tunnel driving," *Psychological Assessment,* vol. 19, pp. 340-346, Sep 2007.

[23] A. Calvi and M. R. De Blasiis, "How Long is Really a Road Tunnel? Application of Driving Simulator for the Evaluation of the Effects of Highway Tunnel on Driving Performance," in *6th International Conference Traffic and Safety in Road Tunnels*, Hamburg, Germany, 2011.

[24] A. Calvi, "Analysis of Driver's Behaviour in Road Tunnels: a Driving Simulation Study," in *2010 International Symposium on Safety Science and Technology*, Zhejiang, China, 2010.

[25] J. Törnros, "Driving behaviour in a real and a simulated road tunnel - A validation study," *Accident Analysis and Prevention,* vol. 30, pp. 497-503, Jul 1998.

[26] T. Hirata, T. Yai, and T. Tagakawa, "Development of the driving simulation system MOVIC-T4 and its validation using field driving data," *Tsinghua Science & Technology* vol. 12, pp. 141-150, 2007.

[27] O. Shechtman, S. Classen, K. Awadzi, and W. Mann, "Comparison of Driving Errors Between On-the-Road and Simulated Driving Assessment: A Validation Study," *Traffic Injury Prevention,* vol. 10, pp. 379-385, 2009.

[28] S. Heliovaara, J.-M. Kuusinen, T. Rinne, T. Korhonen, and H. Ehtamo, "Pedestrian behavior and exit selection in evacuation of a corridor - An experimental study," *Safety Science,* vol. 50, pp. 221-227, Feb 2012.

[29] U. Rüppel and K. Schatz, "Designing a BIM-based serious game for fire safety evacuation simulations," *Advanced Engineering Informatics,* vol. 25, pp. 600-611, 2011.

[30] E. Duarte, F. Rebelo, J. Teles, and M. S. Wogalter, "Behavioral compliance for dynamic versus static signs in an immersive virtual environment," *Applied Ergonomics,* in press.

[31] G. Lawson, S. Sharples, D. Clarke, and S. Cobb, "Validating a low cost approach for predicting human responses to emergency situations," *Applied Ergonomics,* vol. 44, pp. 27-34, 1// 2013.

[32] L. Gamberini, P. Cottone, A. Spagnolli, D. Varotto, and G. Mantovani, "Responding to a fire emergency in a virtual environment: different patterns of action for different situations," *Ergonomics,* vol. 46, pp. 842-858, Jun 20 2003.

[33] E. Ronchi, M. Kinateder, M. Müller, M. Jost, M. Nehfischer, P. Pauli*, et al.*, "Evacuation travel paths in virtual reality experiments for tunnel safety analysis," submitted.

[34] M. Kinateder, E. Ronchi, M. Müller, M. Jost, M. Nehfischer, P. Pauli*, et al.*, "Social influence on route choice in a virtual reality tunnel fire," submitted.

[35] M. Kinateder, M. Müller, A. Mühlberger, and P. Pauli, "Social Influence in a Virtual Tunnel Fire - Influence of Passive Virtual Bystanders," in *Human Behaviour in Fire 2012*, Cambridge, 2012, pp. 506-516.

[36] M. Kinateder, P. Pauli, M. Müller, J. Krieger, F. Heimbecher, I. Rönnau*, et al.*, "Human behaviour in severe tunnel accidents: Effects of information and behavioural training," *Transportation Research Part F: Traffic Psychology and Behaviour,* vol. 17, pp. 20-32, 2013.

[37] Z. Xu, X. Z. Lu, H. Guan, C. Chen, and A. Z. Ren, "A virtual reality based fire training simulator with smoke hazard assessment capacity," *Advances in Engineering Software,* vol. 68, pp. 1-8, 2// 2014.

[38] K. M. O'Connell, M. J. De Jong, K. M. Dufour, T. L. Millwater, S. F. Dukes, and C. L. Winik, "An Integrated Review of Simulation Use in Aeromedical Evacuation Training," *Clinical Simulation in Nursing,* vol. 10, pp. e11-e18, 1// 2014.

[39] S. L. Farra, E. T. Miller, and E. Hodgson, "Virtual reality disaster training: Translation to practice," *Nurse Education in Practice*.

[40] M. Cha, S. Han, J. Lee, and B. Choi, "A virtual reality based fire training simulator integrated with fire dynamics data," *Fire Safety Journal,* vol. 50, pp. 12-24, 5// 2012.

[41] A. I. Ginnis, K. V. Kostas, C. G. Politis, and P. D. Kaklis, "VELOS: A VR platform for ship-evacuation analysis," *Computer-Aided Design,* vol. 42, pp. 1045-1058, 11// 2010.

[42] K. Andree, M. Kinateder, and D. Nilsson, "Immersive Virtual Environment as a Method to experimentally study human behaviour in fire," in *13th International Conference and Exhibition on Fire Science and Engineering*, Royal Holloway College, University of London, UK, 2013, pp. 565-570.

[43] J. Drury, C. Cocking, S. Reicher, A. Burton, D. Schofield, A. Hardwick*, et al.*, "Cooperation versus competition in a mass emergency evacuation: A new laboratory simulation and a new theoretical model," *Behavior research methods,* vol. 41, pp. 957-970, 2009.

[44] T. Kretz, S. Hengst, A. P. Arias, S. Friedberger, and U. D. Hanebeck, "Using a Telepresence System to Investigate Route Choice Behavior," *arXiv preprint arXiv:1111.1103,* 2011.

[45] J. Ribeiro, J. E. Almeida, R. J. Rossetti, A. Coelho, and A. L. Coelho, "Using Serious Games to Train Evacuation Behaviour."

[46] W. Saunders, "Decision making model of behaviour in office building fire evacuations," PhD PhD thesis, Department of Psychology, Victoria University of Technology, 2001.

[47] J. M. Loomis, J. J. Blascovich, and A. C. Beall, "Immersive virtual environment technology as a basic research tool in psychology," *Behavior Research Methods, Instruments, & Computers,* vol. 31, pp. 557-564, 1999.

[48] W. R. Shadish, T. D. Cook, and D. T. Campbell, "Experimental and quasi-experimental designs for generalized causal inference," 2002.

[49] J. Loomis, J. Blascovich, and A. Beall, "Immersive virtual environment technology as a basic research tool in psychology," *Behavior Research Methods,* vol. 31, pp. 557-564, 1999.

[50] R. Patterson, M. D. Winterbottom, and B. J. Pierce, "Perceptual issues in the use of head-mounted visual displays," *Human Factors,* vol. 48, pp. 555-573, Fal 2006.

[51] K. M. Stanney and R. S. Kennedy, "Simulation Sickness," in *Human factors in simulation and training*, P. A. Hancock, D. A. Vincenzi, J. A. Wise, and M. Mouloua, Eds., ed Boca Raton, Florida: CRC Press, 2010, pp. 117-124.

[52] A. Nybakke, R. Ramakrishnan, and V. Interrante, "From virtual to actual mobility: Assessing the benefits of active locomotion through an immersive virtual environment using a motorized wheelchair," in *3D User Interfaces (3DUI), 2012 IEEE Symposium on*, 2012, pp. 27-30.

[53] B. Weber, M. Sagardia, T. Hulin, and C. Preusche, "Visual, Vibrotactile, and Force Feedback of Collisions in Virtual Environments: Effects on Performance, Mental Workload and Spatial Orientation," *in Virtual Augmented and Mixed Reality. Designing and Developing Augmented and Virtual Environments.* vol. 8021, R. Shumaker, Ed., ed: Springer Berlin Heidelberg, 2013, pp. 241-250.

[54] Heidelberg, 2013, pp. 241-250. J. Hummel, J. Dodiya, R. Wolff, A. Gerndt, and T. Kuhlen, "An evaluation of two simple methods for representing heaviness in immersive virtual environments," in *3D User Interfaces (3DUI), 2013 IEEE Symposium on,* 2013, pp. 87-94.

[55] W. M. Association, "World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects," Bulletin of the World Health Organization, vol. 79, p. 373, 2001.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 323–330
DOI: 10.15439/2014F371
ACSIS, Vol. 2

A Framework for Dynamic Analytical Risk Management at the Emergency Scene

From Tribal to Top Down in the Risk Management Maturity Model

Adam Krasuski*
*Section of Computer Science,
The Main School of Fire Service
ul. Słowackiego 52/54, 01-629 Warsaw, Poland
krasuski@inf.sgsp.edu.pl

***Abstract*—We present a framework designed for the risk management at the emergency scene. The system that implements the framework is focused on supporting an Incident Commander during the fire and rescue actions. The system is able to assess and manage the risks with the use of sensory data, ontology modelling and reasoning techniques from AI domain. Within the framework we propose the novel approaches for perceiving and modelling the emergency scene, for reasoning, for assessing the state and the relations among the objects at the scene, for assessing the risk mitigation and for communicating the risks to the Incident Commander.**

***Keywords*-Fire Service, Decision Support, Risk Management, Sensory Data, Domain Ontology**

I. Introduction

EMERGENCY scene is considered one of the most challenging decision making environments [1]. The safety and the success of the fire & rescue (F&R) action depends strongly on the evaluation of the risks at the emergency scene. The *Incident Risk Management* is the principal consideration of an Incident Commander (IC) in order to ensure the safety of the rescuers. Therefore, prior to deciding upon the tactics, risks must be assessed. The IC must identify the threats and the vulnerabilities (subjects to threats) as well as assess the risks and implement all reasonable control measures. The risks must be recognized and controlled before committing rescuers into the danger zone.

In the State Fire Service of Poland there are no regulations that impose an obligation of risk assessment. There are no procedures or habits that introduce the methods of risk assessment or management. The management of F&R actions is regulated according to the general procedures. The procedures in the scope of the evaluation of the emergency scene distinguish reconnaissances: initial, complete and continuous. However, even the experienced ICs are not able to distinguish how these reconnaissances differ from each other, and what exactly should be done within the instance of each of these reconnaissances.

Having the incident – in the scope of the risks – poorly evaluated there is also a problem with proper controlling (by leading) the processes emerging during the F&R action. Therefore, the safety of rescuers and success of the F&R action depends strongly on the experience, knowledge and skills of individuals. The risk management maturity model [2] defines such a process management as *tribal and hectic*. The management of the emergency scene is ad-hoc and chaotic. The success depends primarily on individuals heroics, capabilities and verbal wisdom. The emerging processes are unpredictable, poorly controlled and reactive.

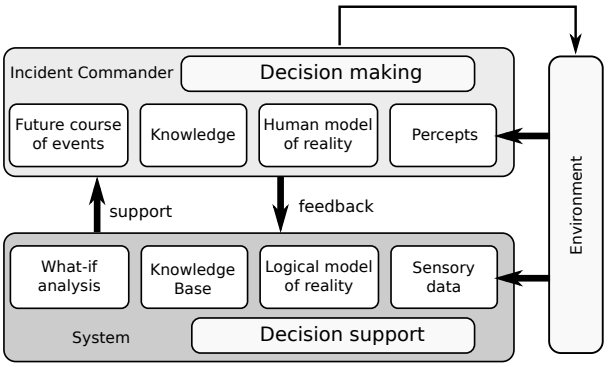


Fig. 1. Cooperation between the IC and the system.

So far there are no standalone computer systems that are able to support the IC at the emergency scene in the risk management activities. This is mainly caused by a) specificity of the decision making environment, b) problems of communicating the risk assessment to the IC. The issue of a) is caused by significant uncertainty and dynamically changing conditions of the objects and phenomena at the emergency scene. There is a problem with obtaining the information which satisfies the IC's *information triangle* rule [3]. It means that the information reported to the IC should be *relevant, accurate* and *timely*. The b) issue originates from the problem that the IC operates under time and mental pressure and has no time for longer analyses and more complex reports. The IC is very sensitive to information overload, simply not important from the intervention objectives point of view [4]. Moreover,

978-83-60810-58-3/$25.00

during the F&R action the IC reasons using the very abstract and vague concepts, such us safety, danger, threats, potential losses and others. Therefore, the system that supports and cooperates with the IC during the F&R actions should use the same concept's namespace as the IC. The system should gather information through the sensory layer and translate the concepts to be "compatible" with the model residing in the IC's mind. The accuracy of such an approximation is crucial in order to follow the IC's strategy and to provide help whenever any new risks arise. Figure 1 illustrates the correspondence of the IC and the software with respect to different aspects [5].

Creating the system which satisfies these constrains is a real challenge. There were a few attempts [6], [7] to build such systems. However, they depended strongly on a dense sensors networks which are not currently operating in the real world. Also, there was an issue with translating all these sensory data into whatever the IC could comprehend.

There are practical implementations that introduce the risk assessment in other countries' Fire Services. However, they are either complex and demand comprehending of large amount of information [3] by the IC or the are based on the IC experience [8]. Therefore, the safety of rescuers and the success of F&R actions depend again on the individuals. Implementing such approaches in the State Fire Service of Poland can only result in the advancement to the *specialist silos* [2] level in the risk maturity model (see Figure 2).

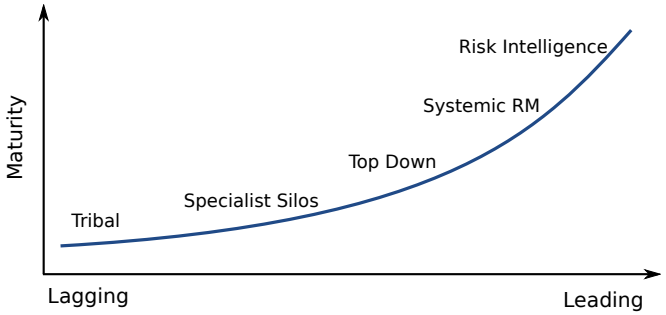


Fig. 2. Risk management maturity model chart.

In this article we present an approach which is able to transform the current *tribal* risk management model into *top down* (see Table I). The top down model is characterized by a) common framework, program statement and policy, b) routine risk assessment c) proper communication of strategic risks to the IC, d) knowledge sharing across risk functions e) awareness activities.

The rest of the article is structured as follows: in section II we present the context for risk management at the emergency scene, giving examples of applied risk assessment methodologies. Section III contains our proposition for the risk assessment. Section IV describes our methods for risk management at the emergency scene. The article is concluded with the evaluation of the approach and a discussion on the perspectives for the future work.

II. The Context

There are two leading approaches implemented for the risk assessment at the emergency scene. One of them is used by German Fire Service and is called *Threat Matrix* (in German – Gefahrenmatrix) [10], [8]. After arriving at the emergency scene German commanders have to recognise and evaluate the appearing risks. In order to do this systematically and not to miss any of the threats they have to fill the Threat Matrix. The Threat Matrix helps to identify both the threats emerging at the scene and the threatened objects (vulnerabilities). Having this information, the commanders can recognize the primary danger to deal with. The approach structures the problem of risk assessment, defining and limiting the set of threats and vulnerabilities to be recognized. However, the method strongly depends on individual experience and intuition. The definition of the consecutive threats are vague and there is no method of risks evaluation – the risks either exist or not. There is no evaluation of the likelihood of risk materialization and the consequences.

The more advanced approach — much more complicated as well — is one used by British Fire Brigade. The approach is composed from three risk assessment methods *Generic Risk Assessment (GRA) Dynamic Risk Assessment (DRA)* and *Analytical Risk Assessment (ARA)* [3].

GRA is a general framework for risk assessment in the Fire Service, regardless of the scope and nature of an incident. The approach takes under account the risks at every stage of duties – from the activities in the fire station via travel to the emergency scene up to the incident commanding. The approach links the risk with the conditions at the emergency scene and the tasks performed by the rescuers. "Generic" means that the values of the risks come from statistics based on the similar actions from the past. The results of the approach consist of a set of rules matching the given situation [11]. During the F&R action GRA allows the IC to operate under the standard procedures. GRA forms the foundations for DRA, operating procedures and training schemes. It also assists in the completion for ARA at incidents.

The second one, DRA, is used to describe the continuing assessment of the risks that is carried out in a rapidly changing environment at the emergency scene. DRA is defined in the initial phase and then reviewed continuously and updated. The outcome of DRA is a declaration of a tactical scheme for the IC, i.e. *offensive* or *defensive*. DRA is a continuous process and takes into account the continually and sometimes rapidly evolving nature of an incident. During DRA phase the IC refines the general rules defined by GRA [11] and fits them according to the state of the phenomenon, objects involved, equipment available and others. DRA must be reviewed continuously and updated as required. Having carried out the DRA and the tactical scheme established, the IC is aware of the immediate threats, vulnerabilities at risk and the control measures necessary to protect those vulnerabilities. This initial assessment of DRA further forms the basis of a more detailed risk assessment – ARA. ARA is introduced to analyse situation

TABLE I
THE RISK MANAGEMENT MATURITY MODEL [9]

Tribal and Hectic	**Specialist Silos**	**Top Down**	**Systemic**	**Risk Intelligence**
Ad-hoc/chaotic.	Independent risk management activities.	Common framework, program statement, policy.	Coordinated risk management activities.	Embedded in strategic planning, capital allocation, product development etc. across silos.
Depends primarily on individual heroics, capabilities, and verbal wisdom.	Limited focus on the linkage between risks.	Routine risk assessments.	Risk appetite is fully defined.	Early warning risk indicators.
	Limited alignment of risk to strategies.	Communication of too strategic risks to the Board.	Enterprise-wide risk monitoring, measuring and reporting.	Linkage to performance measurement/incentives.
	Disproportionate monitoring and reporting functions.	Executive /Steering committee.	Technology implementation.	Risk modelling/scenarios.
		Knowledge sharing across risk functions. Awareness activities.	Consistency plans and escalation procedures. Risk Management training.	Industry benchmarking.

in more detail on the basis of information obtained from the reconnaissance and from the rescuers. The special forms are defined and provided to the IC in order to help calculating and recording ARA [12]. The outcome of the review of ARA either confirms that the DRA and chosen tactical scheme was correct, or it results in a change of the scheme. This also provides the basis for the current and future DRA.

The discussed approaches enhance the risk assessment at the emergency scene and improve the safety of the rescuers. However, they have a major shortcoming: it is not easy to implement the risk assessment as a stand-alone, unsupervised computer process since they require a) a rich sensory infrastructure and b) a clever AI processing.

The issue with a) is continuously improving: the technology can deliver more precise, more modern and cheaper sensors each year which can produce lots of streams of data about various phenomena. The b) issue improves as well: there is a significant improvement in the fire and evacuation modelling [13], [14], the ontology modelling methodologies are being invented and evaluated, AI-based algorithms can support big data analytics and so on. We can therefore support the claim that the computer-driven Dynamic Analytical Risk Assessment, independent from IC is becoming more and more feasible.

III. DYNAMIC ANALYTICAL RISK ASSESSMENT

We propose an approach which allows for supporting the IC at the emergency scene in the managing of the risks. We called our approach Dynamic Analytical Risk Assessment due to the fact that the method reacts dynamically to the changing at the emergency scene and is detailed enough to be considered an analytical risk assessment. Our approach derives the foundations from the risk approaches presented in section II and uses the methods from AI to implement the ideas.

A. Scene Modelling

We start our process of creating the scene model from the review of the domain. We used a Use Case diagram for this purpose. The elaboration of the diagrams results also in a better mutual understanding between architect of the system, analytics and domain experts. The Use Case diagrams allow to extract the main objects, concepts and relation within the domain. Then, we used a set of documents called *incident analysis* in order to obtain the more detailed description of the domain. The documents study in detail the selected incidents and contain a comprehensive description of them, including the context, previous trainings at the objects involved, their recognition, the course of action minute after minute, decisions made and their background. It allows us for extracting (with support of domain experts) the complex objects, their hierarchy and spatio-temporal relation among them. The description was too complex to model it using Use Case diagrams. Therefore, we use the taxonomic hierarchies of classes defined by [15] in order to better represent the hierarchy and relationships between complex, plain objects at the emergency scene and their attributes.

In our approach we consider an emergency incident as a set of frames [16] from time t_s when the incident begins to the time t_e when the last crew come back to the fire station. A single frame F_n from the set represents the emergency scene at time t_n. The frame can be considered as a complex object which is composed by other complex objects such as buildings, equipments, rescuers, occupants and others. Figure 3 depicts the idea of the perception of the scene.

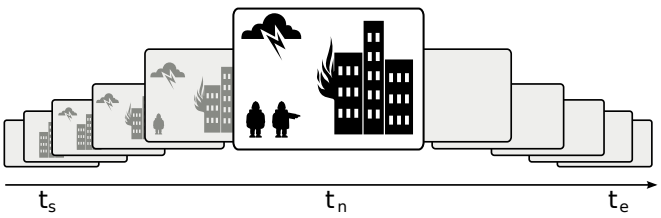


Fig. 3. The idea of frame-based scene modelling.

The complex objects within the frame may be composed from other complex objects or just by plain objects. Plain

objects are represented only using vectors of attributes values. We can create for each of the complex objects at the scene the hierarchy of sub-objects, attributes or both.

The attributes describing the objects can be static i.e. *nominal pressure* of the firefighting nozzle or dynamic when they reflect the current state of the object i.e. *a firefighter is exhausted*. The static attributes can be quite easily defined when the object is created. The values of dynamic attributes change continuously and are much more difficult to define. They depend on the situation at the scene and need communication with the sensory layer. For example, the level of the fatigue of the firefighter can be defined at the basis of breathing ratio and exhausted carbon dioxide concentration.

Modelling such aspects is challenging as it require to apply sophisticated methods. In order to face this problem we extend our ontology by the spatio-temporal perceptual concepts modelling defined by [17], [18]. The approach needs a domain ontology which is a core for reasoning processes. The creation of the domain ontology needs tight cooperation with domain experts. A cooperation with domain experts towards definitions of ontology is poorly studied. There are also no measures evaluating the correctness and completeness of the created ontology. Therefore a high attention should be paid in order to perform this correctly. First we created a draft of the ontology on the basis of domain literature [8], [19], [3]. Then we extended this draft with support of domain experts and contextual visualization [20] of the situation. In this process we involved not only the experts but also the software architects and psychologists. Figure 4 depicts the simplified snapshot of the created ontology.

B. Risk Assessing

The created ontology is only a carrier in our reasoning processes. It allows for structuring the problem and applying the method of *divide and conquer*. The evaluation of the value of the selected risk needs applying the hierarchical classification. This means that at the basis of lower layer concepts from our ontology we *approximate* a higher level concepts. We consider this process as an approximation because the concepts, objects or attributes from lower levels are not in such relation with higher level, which allows for its crisp definition.

The approximation of the higher level concepts by the lower level is a problem which in our case can be reduced to the classification problem. The standard classification uses the information system [21] for training the classifiers. The classifiers have to extract the features and their impact on decision class. However, in the hierarchical classification, where the decisions classes of lower level classifiers become the attributes for higher classes, the approach is insufficient due to the computing complexity. For example, in our case we have a sub-system for recognition of activities performed by rescuers [22]. The sub-system consists of a set of accelerometers and magnetometers placed in different body parts of the rescuers. In the purpose of the recognition of the activity of single rescuer the standard classification approach is good enough. However, if the recognition of some activity needs observing the group of rescuers (i.e. a tactic used to access the room on fire) the standard approach fails. This is caused by the necessity of consideration of a Cartesian product of each of the attributes values from the sensors.

In our approach, domain experts assist not only in the creation of ontology and categorization of objects/situations but also in learning the phase of classifiers. This recalls the human learning process when the tutor filters the information indicating features which plays the key role in the classification problem. In our case it is important that the higher level concepts (objects) are created as relational structures in which the points are represented by the vectors of attributes values from the lower hierarchical level and relations between them represent constrains. Over such objects the new attributes are defined with domain experts support. On the basis of this idea, the methods for ontology approximation were developed [23], [24].

We use the classifiers in order to induce the rules [25], [26]. Rules learnt from data can be used to support approximate reasoning about the concepts. Approximations can be considered both with respect to degrees of satisfaction of particular patterns in the observed data and the degrees of correspondence of previously unseen situations to already established ontology areas. It allows us for building the dynamic and spatio-temporal model of the emergency scene.

The presence, the state and the relations among complex objects at the scene define the concepts used by IC during the reasoning process. As was mentioned in the Introduction the concepts originate from the risk assessment field. Those are such concepts as: threats, vulnerabilities, risk, safety, danger and others. In order to approximate these concepts the across-hierarchy reasoning about objects within the frame is used, as well as spatio-temporal across-frames relations reasoning. For example, in order to evaluate, whether in a given moment the risk of explosion for rescuers exists, we have to consider the chances of *backdraft*[1] occurrence and recognition whether rescuers are currently entering the compartment on fire. Figure 4 depicts a simplified snapshot of the ontology created for recognition of the risk of an explosion for rescuers.

We present the methodology of risk assessment performed by the system, on the following example. The sub-systems for recognition of the activity of the rescuers and their position [22], [27] generate the stream of data. The set of classifiers uses these data to approximate the lower level concepts from our ontology (see Figure 4). These concepts are related to the navigation in the building, fire location as well as usage of rescue equipment. The decision classes from those classifiers constitute the attributes for higher level classifiers which recognize for example the usage of forcible entry tools. The usage of forcible entry tools simultaneous with the kneeling position of other rescuer approximate the concept of starting position of rescuers to enter the compartment on fire. The starting position of rescuers preceded by "gaining access to the fire" indicates that rescuers are already entering

[1] http://en.wikipedia.org/wiki/Backdraft

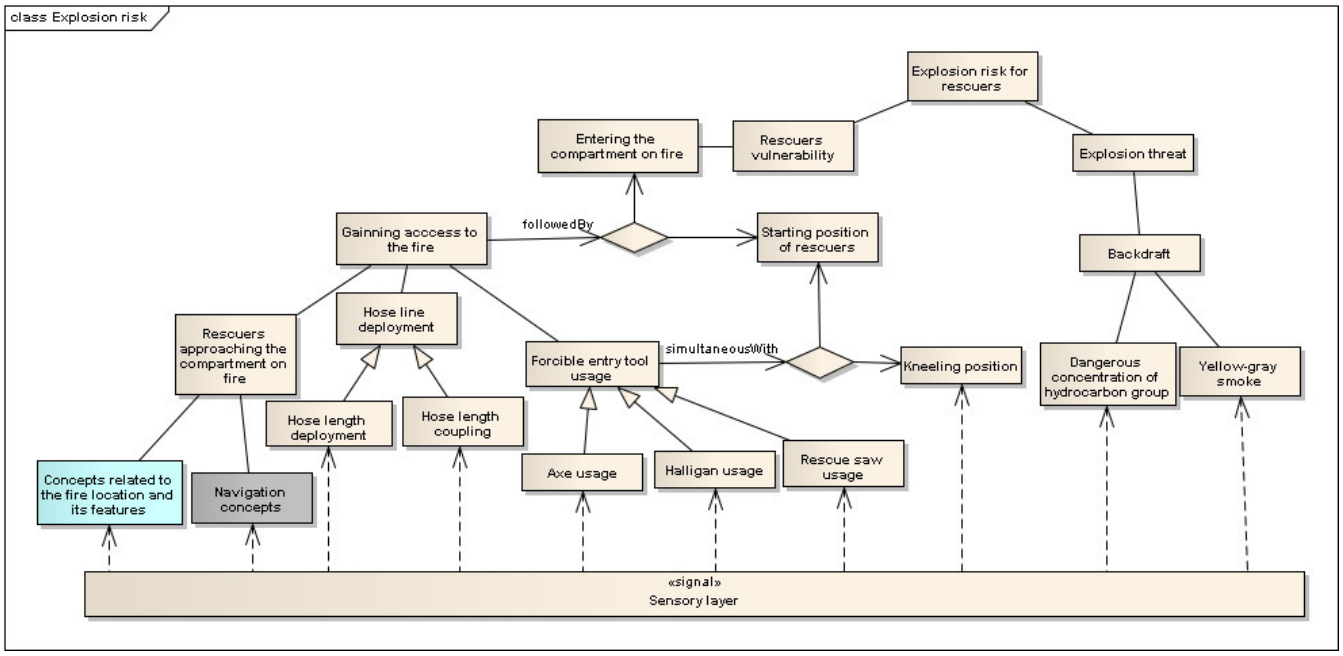


Fig. 4. The ontology for hierarchical spatio-temporal reasoning.

the compartment on fire creating vulnerability on the explosion threat caused by the backdraft phenomenon. The sensors from fire detector aspiration system and smoke observation provide the data for the classifiers which recognize the likelihood of explosion threat. All the concepts defined are introduced to evaluate the root concepts of *explosion risk for rescuers* (see Figure 4). The presented methods of feature extraction and filtering from lower level to upper level is supported by domain experts.

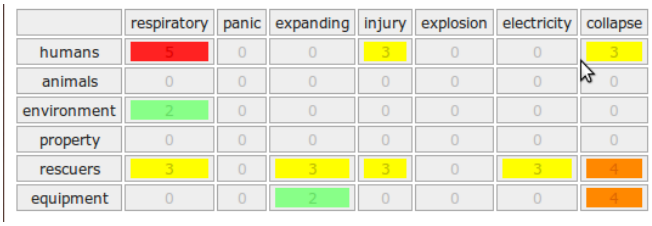


Fig. 5. An example Threat Matrix

C. Risk Communicating

As we mentioned in the Introduction, the IC is a very demanding subject to efficient risk reporting. Therefore, we introduce a hierarchical level of risk communication. We use the general indicator about the actual risk level at the top level of the hierarchy. Due to the fact that we use augmented reality glasses (characterized by low resolution) to communicate with intervention level commander, the highest level risk indicators are just two squares with colors: green, yellow or red indicating risk for human and rescuers. If the IC needs a more detailed information about the actual risks, the second level of risk communication is Threat Matrix. The Threat Matrix presents the threats at the emergency scene and the vulnerabilities which can be subjects to the threats. The example Threat Matrix is presented in Figure 5.

In most cases the matrix contains enough information for the IC to evaluate the emergency scene [8]. However, the less experienced ICs may need a more detailed explanation concerning the origins of the risks presented in the Threat Matrix. Therefore, we created the next level of information provided to the IC. We present the rules which were launched and served to calculate the given risks. Table II depicts the presentation of the rules.

TABLE II
THE PRESENTATION OF THE RULES THAT WERE USED TO CALCULATE RISKS IN THE THREAT MATRIX

Id	Situation description	L	S	Control
A1.1	farm site fires: presence of dust	2	1	use Personal Protection Equipment (PPE)
A1.3	hazardous atmospheres	1	4	Use PPE, Resuscitation equipment immediately available / monitoring of atmosphere
A1.7	asbestos on the roof (the hazard comes from breathing)	1	2	Use PPE, decontamination after intervention

The rules presented in the table contain also the control measures aimed at decreasing the severity if the risks do materialise. This is additional guide for less experienced ICs to help them in risk mitigation.

IV. Risk Management

The presented approach allows for the complete risk assessment at the emergency scene and for its presentation to the IC. However the risk assessment is only part of the process of risk management. The other important parts of this process are methods of risk mitigation and measurements of the effectiveness of the controls applied.

A. Risk Mitigating

The presented approach for risk assessment allows for reasoning under uncertainty about vague concepts at the emergency scene. This allows the IC for better assessing of the situation and keeping the operation safe. However, the better situational awareness is only part of the success of the incident risk management. Equally important is planning and operating. The good incident action plan (IAP) plays a key role in the successful incident management [19]. According to the [28] the IAP should contain the strategy of risk mitigation. A good mitigation plan predicts future course of events and proposes the adequate controls to mitigate the risks. Therefore, the risk management could be perceived as a game between the nature and the IC. In order to win the game IC has to recognize the "strategy of the nature". The recognition of the strategy and creating own strategy is a challenging task. We are not able to address the problem yet. Therefore we are going to face the problem in our future work aimed at transition of the system into *Risk Intelligence* (see Figure 2). At the current state of the research we are only able to hint the general recommendation and propose the control measures matching the rules from risk assessment (see example in Table II).

The system determines, on the basis of the risk assessment expressed by the Threat Matrix, whether the potential benefits (saved live or property) outweighs the undertaken risks. If this is the case, the system proposes the general recommendation – the tactics scheme – (*offensive* or *defensive*). The proposition of the scheme is based on the rules, taking mostly into account the chances that people are present inside the building and the building construction type. If this scheme is accepted by the IC, the system is trying to endeavour to reduce the risks to an acceptable level.

The second level in our *hierarchy of control measures* are the general strategies of applying the control measures. At every moment in the F&R action the system is trying to recognize whether any of the following strategies should be applied. *Eliminate* the risk or substitute it with something less dangerous. For example changing the scheme to defensive thus preventing rescuers access the danger zone. *Reduce* the risk by preventing or reducing the number of vulnerabilities that come into contact with the risk or reducing the time of the exposure to the risk. The strategy is calculated according the evaluation of the parameters of the fire [29] and the amount of resources needed to extinguish the fire or to rescue people. Ensuring that *discipline* is maintained throughout the exposure to the risk. This is performed by monitoring and visualization of activities performed by the rescuers [22].

The third level in hierarchy of control measures constitute the rules used for risk assessment. As it was mentioned in section III-B there were rules induced from ontology which approximate the concepts related to the risks. We asked the domain experts to define the controls which should be used if the given risks materialize. The number of rules, even limited to active at the moment, is significant. Moreover, the controls proposed are very detailed and need some attention while reading. Therefore, leaving the navigation across the rules to the IC would result in information overload. We tried to partially address the problem by introducing a tool called what-if analysis. The approach allows for keyword search, faced search or fast navigation across the rules. The IC or her/his assistant at the control room can quite quickly find, using the keywords, the rules matching the actual situation. This allows for fast review and implementation of proper control measures. The IC has access to the appropriate risk related information to assist with the identification of suitable control measures. This, in conjunction with other specific facts regarding the premises, for example information gained on risk visits, will assist the IC to formulate an effective plan.

B. Performance Indicators

The approach presented so far is designated to deal with the risk defined as a likelihood of threatening the vulnerabilities and the potential consequences [28]. However, during the rescue action there is also a risk related to the definition provided by ISO 31000 defined as an effect of uncertainty on intervention objectives [30]. This type of the risk is related to the tactics applied by the IC. Each of the activities of the rescuers committed by IC are characterized by uncertainty about the obtained outcome. This type of the risk should be measured by the defined performance indicator of applied strategy.

The performance should be measured against agreed standards to reveal when and where improvement is needed. Active self-monitoring of the system reveals how effectively the management system is functioning, looking at the equipment, processes and individual behaviour/performance.

Every incident has an objective that reflects the mission's objective – protect life, property and the environment from harm. An IC develops a strategy for accomplishing this mission, depending on the conditions that exist at the time. The rescuers execute the full mechanics of the tasks to complete each phase of the operation at the emergency scene. These tasks are based on fundamentals – ventilation; nozzle operation; water flow rates; secondary egress and emergency bailout by ladder.

Although each of the incidents is different there are common tactical phases of the incident management. There can be distinguished: arriving, reconnaissance, resource deployment, gaining access to the fire, search and rescue activities and extinguishing. Each of the phases has a set of activities and the outcome. The activities should be performed with accordance to the tactic and training processes, executing the rule "play as you train". Therefore, we can define in each tactical phase

the checklist which should be completed if the IC obeys to the procedures. There also should be observable, measurable effects for each of the tasks that is performed, which can be observed by the system, and the effects of completion of each step as the outcome.

If the set of activities at each phase is performed according to the checklist and the outcome is consistent with the expectations (trained) then we can state that the risk of the intervention objectives is low. If the IC is not well-trained, insubordinated and does not obey the defined checklist, then the risk related to the uncertainty on the intervention objectives rises. The analogous situation is when the conditions of the incident are changing in an unexpected way – then the risk also rises.

For example, the rescuers are in the phase of entering the compartment on fire. The checklist consists of: rescuers in full gear kneeling before the door, breathing apparatus in use, hose line wet, forcible entry tools ready to use, second squad ready for assistance, etc. The outcome of the phase are the jets cooling the ceiling and opened windows. Every task and tactical procedure completed is also reported as a "benchmark".

C. Call for Action

Apart from the organization of the scene where the risks are assessed, the main idea of this process is the *call for action*. Having the risks assessed we have to communicate them in such a way that forces the stockholders to the action of the risk mitigation. In our system the call for action is implemented by risk exposure and control activity level matrix. Figure 6 depicts the idea of the matrix.

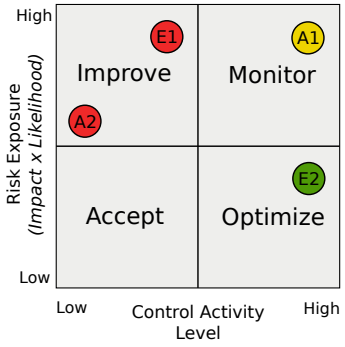


Fig. 6. Risk maps for presenting risk/control relationships.

There are four areas distinguished in the matrix. *Improve* is the area of high risk exposure with a low level of control. This area must be key priority for improvements in management and control activities. *Monitor* is the area of high risk exposure where controls are deemed adequate. This area should be monitored to provide the ongoing assurance of control effectiveness. *Accept* is the area of low risk exposure that also have a lower level of control. This area may be consciously accepted by the IC. *Optimize* is the area of low risk exposure with a high level of control. This area may generate opportunities for the IC to optimize the management and control activities.

Such a visualization of the risks and the controls illustrates clearly the areas where the current controls are not sufficient to operate in safe condition. This forces the IC to mitigate the risk in such areas or to lower the risks by it avoidance. The location of different risks within the given areas of risk maps is calculated according to the rules presented in Table II. The rules beside the calculation of the risks have also defined the control measures which help to mitigate the risks. If the selected rule is active and there is no adequate control applied then balance between the risk and the control is biased. The location of a given risk within the risk maps depends also on the risk level. Figure 6 illustrates that the risk caused by threats A2 and E1 should be handled by implementing the additional control measures; threat A1 creates a high risk, however it is properly controlled and some controls from E2 may be relaxed.

V. Conclusions

We present an approach allowing for the management of the risks at the emergency scene. The approach defines the framework for scene modelling, introduces the reasoning algorithms, risks mitigation methods, performance measuring and risk reporting and communicating. We implemented all the presented ideas into a standalone computer system. However, so far the system is not deployed in the State Fire Service of Poland. Our system is currently at the 4 level of Technology Readiness Level[2] It means that the main technological components of the system are integrated to establish that they will work together. This is relatively “low fidelity” compared with the eventual system. The system was tested in the laboratory with controlled parameter of the fire and many assumed simplifications. However, on the basis of the performed experiments we can support our claim that Dynamic Analytical Risk Assessment, independent from IC is becoming feasible.

We argue that it is possible to improve the quality of interventions and minimize the corresponding risks by providing IC with the support in the following areas: a) grouping and interpreting incoming information by means of higher level concepts and linking them with intervention objectives; b) filtering and ranking information; c) indicating which information is missing in order to make reliable decision; d) indicating where and how to acquire important information; e) monitoring situation and decisions made so far. We claim that such functionalities can be achieved through a combination of modern methods from the domain of Information System Security Risk Management, data organization with compliance to ontological approaches and interactive algorithms processing recommendations for IC. We pointed out that there is still a gap between analytical models and human abilities to benefit from them.

[2]http://en.wikipedia.org/wiki/Technology_readiness_level

ACKNOWLEDGMENT

Supported by Polish National Centre for Research and Development (NCBiR) – Grant No. O ROB/0010/03/001 in the frame of Defence and Security Programmes and Projects: "Modern engineering tools for decision support for commanders of the State Fire Service of Poland during Fire&Rescue operations in the buildings".

REFERENCES

[1] B. Brehmer, "Strategies in Real-Time, Dynamic Decision Making," *Insights in decision making*, pp. 262–279, 1990.

[2] P. X. Zou, Y. Chen, and T.-Y. Chan, "Understanding and improving your risk management capability: Assessment model for construction organizations," *Journal of Construction Engineering and Management*, vol. 136, no. 8, pp. 854–863, 2009. [Online]. Available: http://dx.doi.org/10.1061/(ASCE)CO.1943-7862.0000175

[3] Department of Communities and Local Goverment, *Fire Service Operations, Incident Command*, 3rd ed., ser. Fire Service Manual. London TSO, 2008.

[4] A. Cowlard, W. Jahn, C. Abecassis-Empis, G. Rein, and J. L. Torero, "Sensor Assisted Fire Fighting," *Fire Technology*, vol. 46, no. 3, pp. 719–741, 2010. [Online]. Available: http://dx.doi.org/10.1007/s10694-008-0069-1

[5] A. Krasuski, A. Jankowski, A. Skowron, and D. Slezak, "From sensory data to decision making: A perspective on supporting a fire commander," in *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*, vol. 3. IEEE, 2013, pp. 229–236. [Online]. Available: http://dx.doi.org/10.1109/WI-IAT.2013.188

[6] H. Liangxiu *et al.*, "FireGrid: An e-infrastructure for next-generation emergency response support ," *Journal of Parallel and Distributed Computing*, vol. 70, no. 11, pp. 1128 – 1141, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.jpdc.2010.06.005

[7] N. Ashish, J. Lickfett, S. Mehrotra, and N. Venkatasubramanian, "The software ebox: Integrated information for situational awareness," in *Intelligence and Security Informatics, 2009. ISI'09. IEEE International Conference on*. IEEE, 2009, pp. 77–82. [Online]. Available: http://dx.doi.org/10.1109/ISI.2009.5137275

[8] A. Graeger, U. Cimolino, H. de Vries, and J. Sümersen, *Einsatz- und Abschnittsleitung: Das Einsatz-Führungs-System (EFS)*. Ecomed Sicherheit, 2009.

[9] B. Endicott-Popovsky, "End-to-End Risk Assessment Approach," in *Building an Information Risk Management Toolkit*. Coursera.org, 2014, p. 23.

[10] Bundesamt für Bevölkerungsschutz und Katastrophenhilfe, "Feuerwehr-Dienstvorschrift 100 Führung und Leitung im Einsatz : Führungssystem," FwDV 100 Stand: 10. März 1999.

[11] Department of Communities and Local Goverment, *Generic Risk Assessments, GRA 3.1 Fighting fires in buildings*, ser. Fire and Rescue Authorities Operational Guidance. London TSO, 2011.

[12] Department of Cummunities and Local Government, "Fire and Rescue Service Operational guidance. Operational Risk Information," 2012.

[13] W. Jahn, G. Rein, and J. Torero, "Forecasting fire growth using an inverse zone modelling approach," *Fire Safety Journal*, vol. 46, no. 3, pp. 81–88, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.firesaf.2010.10.001

[14] ——, "Forecasting fire dynamics using inverse computational fluid dynamics and tangent linearisation," *Advances in Engineering Software*, vol. 47, no. 1, pp. 114–126, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.advengsoft.2011.12.005

[15] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge acquisition*, vol. 5, no. 2, pp. 199–220, 1993. [Online]. Available: http://dx.doi.org/10.1006/knac.1993.1008

[16] I. Düntsch and E. Orlowska, "A discrete duality between apartness algebras and apartness frames," *Journal of Applied Non-classical Logics*, vol. 18, no. 2-3, pp. 213–227, 2008. [Online]. Available: http://dx.doi.org/10.3166/JANCL.18.213-227

[17] A. Mallik, H. Ghosh, S. Chaudhury, and G. Harit, "Mowl: An ontology representation language for web-based multimedia applications," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 10, no. 1, p. 8, 2013. [Online]. Available: http://dx.doi.org/10.1145/2069276.2069280

[18] A. Mallik, S. Chaudhury, and H. Ghosh, "Nrityakosha: Preserving the intangible heritage of indian classical dance," *Journal on Computing and Cultural Heritage (JOCCH)*, vol. 4, no. 3, p. 11, 2011. [Online]. Available: http://dx.doi.org/10.1145/2542205.2542210

[19] Emergency Management Institute, "Introduction to Incident Command System, ICS-100," http://training.fema.gov/EMIWeb/IS/courseOverview.aspx?code=IS-100.b, 2013, access: 22.02.201.

[20] P. Teicholz, R. Sacks, and K. Liston, *BIM handbook: a guide to building information modeling for owners, managers, designers, engineers, and contractors*. Wiley, 2011.

[21] Z. Pawlak, "Information systems theoretical foundations," *Information systems*, vol. 6, no. 3, pp. 205–218, 1981. [Online]. Available: http://dx.doi.org/10.1016/0306-4379(81)90023-5

[22] M. Meina, K. Rykaczeweski, and B. Celmer, "Towards robust framework for on-line human activity reporting using accelerometer readings," *Lecture Notes in Computer Science*, vol. 8610, pp. 350–361, 2014.

[23] J. Bazan, "Hierarchical classifiers for complex spatio-temporal concepts," in *Transactions on Rough Sets IX: Journal Subline*, ser. Lecture Notes in Computer Science, J. F. Peters, A. Skowron, and H. Rybiński, Eds. Heidelberg: Springer, 2008, vol. 5390, pp. 474–750.

[24] J. G. Bazan and A. Skowron, "Classifiers based on approximate reasoning schemes," in *Monitoring, Security, and Rescue Tasks in Multiagent Systems (MSRAS'2004)*, ser. Advances in Soft Computing, B. Dunin-Kęplicz, A. Jankowski, A. Skowron, and M. Szczuka, Eds. Heidelberg: Springer, 2005, pp. 191–202. [Online]. Available: http://dx.doi.org/10.1007/3-540-32370-8_13

[25] S. H. Nguyen, J. Bazan, A. Skowron, and H. S. Nguyen, "Layered learning for concept synthesis," in *Transactions on Rough Sets I: Journal Subline*, ser. Lecture Notes in Computer Science, J. F. Peters and A. Skowron, Eds. Heidelberg: Springer, 2004, vol. 3100, pp. 187–208. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-27794-1_9

[26] S. H. Nguyen, T. T. Nguyen, M. Szczuka, and H. S. Nguyen, "An Approach to Pattern Recognition based on Hierarchical Granular Computing," *Fundamenta Informaticae*, vol. 127, no. 1-4, pp. 369–384, 2013.

[27] M. Meina, K. Rykaczeweski, and B. Celmer, "Certain Aspects of Foot-Mounted Inertial-based Indoor Navigation Systems," in *Type II Proceedings of WIC 2014 Conference.Warsaw, August 11-14*, 2014.

[28] G. Stoneburner, A. Goguen, and A. Feringa, "Risk management guide for information technology systems," *Nist special publication*, vol. 800, no. 30, pp. 800–30, 2002.

[29] M. Fliszkiewicz, A. Krasuski, and K. Kreński, " Evaluation of a Heat Release Rate based on Massively Generated Simulations and Machine Learning Approach," in *Proceeding of FedCSIS 2014 Conference, Warsaw, September 9-11*, 2014.

[30] "ISO 31000 - Risk management," 2009.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 331–335
DOI: 10.15439/2014F406
ACSIS, Vol. 2

Data Cleansing of the Fire & Rescue Text Corpus.
The Case Study of Correction of the Misspellings and Segmentation into Sentences

Karol Kreński*, Mateusz Fliszkiewicz†
*Section of Computer Science, The Main School of Fire Service
ul. Słowackiego 52/54, 01-629 Warsaw, Poland
krenski@inf.sgsp.edu.pl,
†Section of Computer Science, The Main School of Fire Service
ul. Słowackiego 52/54, 01-629 Warsaw, Poland
fliszkiewicz@inf.sgsp.edu.pl

***Abstract*—The article presents a case study of applying data cleansing methods and segmentation procedures in order to correct and enhance the structure of the domain corpus of fire service. During the study we present our approach and the results in the task of correcting the misspellings, as well as the method of segmenting the corpus into sentences.**

***Index Terms*—Fire Service, Data Cleansing, Text Corpus, Misspellings, Segmentation**

I. Introduction

OVER the years, the National Fire Service of Poland collected a large corpus of texts from about 6 million incidents. Unfortunately, there was little validation of the input which resulted not only in the (too much) free form of the texts which is difficult to automatically process, but also in lots of misspellings and lack of structure (sentences boundaries) which further impede computer analyses.

Our research was focused on finding the characteristics of the above problems and by using mostly regular expressions, n-gram analysis, spell checkers and databases of some entities (e.g. geographic locations) as well as reference domain texts (fire & rescue journal) we tried to cleanse the corpus.

The paper is structured as follows: In section II we introduce the fire & rescue text corpus named EWID. In section III we present our motivation and the context of our research. In sections IV and V we provide the details of how we corrected the corpus.

II. Characteristics of the EWID corpus

The national fire and rescue services (just like, e.g., police) are typically equipped with the incident data reporting systems (IDRS), which gather the information about conducted actions. Each of approximately 500 Fire and Rescue Units (JRG) of the State Fire Service of Poland (PSP) conducts around 3 fire & rescue actions per day. After every action a report is created in EWID – the internal computerized reporting system of PSP [1]. As of 2014, the total number of the reports in EWID is around 6 million, of which about 0.3 million records were available for the purposes of this research. Each record contains 560+ attributes (only a few dozens are usually set per record). Most of these attributes provide yes/no information about action parameters (binary), but there are also timestamps, quantities and short text entries. There is one attribute which we consider distinct: the natural language *description* of the action.

The collection of the 0.3 million EWID *descriptions* contains about 60 MB of texts, which is about 8 mln of words, written in semi-natural, technical language. Following is a sample passage from the corpus (awkward vocabulary and misspelled words are intentional):

"Affer arriving at the fire scene the undergrowth fire was observed. Two firefighting jets ware applied and suction line from the nearby lake was created. After putting out the fire, appliance crew came back to fire station".

The concern is that over the years this large corpus which contains valuable information has been collected with limited validation of the input. This situation is considered quite common in the real world data collections [2], [3]. The corpus in scope requires data cleansing followed by further processing in order to improve the semantics. This case study is focused on the data cleansing only. It may be beneficial for other text corpora, which are affected by typographic errors. We know of projects where data cleansing step was explicitly skipped, as the expected solution was no other than a laborious human work [4], [5]. In this work we propose a mostly automatic, iterative process supervised by domain experts. It is important to mention that we are more interested in having the entities in the text unified (disambiguated) rather than grammatically correct. We assume, that for the purpose of further operations on EWID corpus, such as clustering, statistical analysis, and so on, this unification may be beneficial.

III. Motivation

The motivation for the corpus reparation is to prepare the ground for further processing. In our particular case there was a need to have the data cleansed when working on a concept of a decision support system named CLEWID [6]. CLEWID is a proposition of a platform for fire & rescue data analyses. It is built of 4 layers, namely: 1) the raw data layer (EWID

978-83-60810-58-3/$25.00

and other sources), 2) the quality data layer, 3) the granular (semantic) layer, 4) the models layer. The scope of this article is to transform the raw data layer into the quality data layer to have the higher level layers operate on the more precise and strict data. The semantics are fixed on the granular layer, since granulation is about organising the data based on various aspects of their similarity.

IV. Details of corrections of the misspellings

The process was divided into stages. Each stage describes another approach and provides the information how much gain was achieved.

1) Removal of redundant characters: The lack of the validation for the *description section* in EWID database results in various characters being incorrectly inserted. This may result in creation of alternate forms of the same entities (e.g. GBA3 vs GBA-3). The selection of the characters which should be removed requires the input from the expert – this step can be done by searching for the words containing non-alphanumeric characters and deciding which of the characters should be dropped. In the case of EWID corpus most of non-alphanumeric characters were replaced by space. Additionally, digits at the beginning of a word boundary and hyphens within word boundaries after a letter and before a digit were removed.

2) Frequent words not recognized by the dictionary: At this stage there were 8,044,535 words including 309,036 words which were not recognized by the popular *aspell* spell checker [1] (3.8% error ratio). 500 most frequent (the reasonable number for a human to manually process) of the 309,036 not recognized words were extracted from the corpus. 200 of these entries proved to be valid words from domain vocabulary – it was reasonable that aspell didn't have them in its database. Automatic corrections by aspell were proposed for the remaining 300 and they were later manually adjusted by the expert. The knowledge from the domain expert was instrumental in achieving reasonable outcome, as some cases were not quite obvious. For example, aspell proposed 'dzielenie' (division) as the correction for the misspelled word 'dzialenie', which was overruled by expert's 'działanie' (action). The corrections were applied to the corpus and the spell checker was rerun. The error ratio dropped to 2.9%. This particular fix was an example of a huge gain with little effort.

3) The additional dictionaries: In order to extend the spell checker, we searched for collections of the domain vocabulary. There exist a number of texts collections which could serve as a reference in composing the domain vocabulary (domain knowledge). Ultimately, the expert decided that the domain journal "Przeglad Pozarniczy" (*PP*)[2] would contain the texts that are most relevant for the operational content of EWID corpus. PP publishes fire & rescue related articles, and by the fact that it is a journal it (hopefully) contains very small amount of misspellings. We spell checked the acquired PP corpus and all the misspellings reported by aspell (words not found in standard dictionary) were treated as candidates for domain vocabulary, thus domain dictionary was created. The extended spell checker reported error rate of 2.15%.

The idea to use a good quality domain corpus as an extension of a spell checker came after we already fixed the 500 most frequent words manually (as described in the previous section). The spell checker extended by journal-based domain vocabulary would likely have recognized most of the frequent words from EWID since they come from the same domain of fire & rescue. The lesson learned is that the domain vocabulary should be used as early as possible (if it is available).

By knowing the content of EWID the expert added more elements to dictionary. Geographical entities – streets, cities and districts were obtained from the external public sources (Polish governmental/administration organisations) and became another extension to the dictionary. The spell checker extended with the domain vocabulary and geographical entities was rerun and the error ratio dropped to 1.75%.

Another key step was the inclusion of surnames. Surnames in EWID are frequently reported as misspellings by aspell. Fortunately most first names are recognized by aspell. The public database of Polish surnames and first names was acquired and roughly checked for the completeness against our students surnames database (around 200 entries). 97% of the surnames were recognized, so the completeness of the surnames database was reasonable. However, this mechanism proved to be too greedy – too many actual misspellings were being forgiven as possible surnames. We needed to drop the surnames database. The spell checker extended with domain vocabulary, geographical entities and names reported 1.56% of errors.

4) The n-gram approach: The knowledge-based extensions of spell checker's dictionary exhausted the inventory of easy fixes. The remainder of the misspellings in the corpus required more extensive approach. The method that we have applied replaces (corrects) a misspelled words using their nearest correct neighbor. The neighbor(s) of a given word needs to be identified in a meaningful way. For this purpose, the list of all 3-grams (unique triplets of words in the corpus) was created. This list was spell checked with the use of the extended spell checker introduced above and, as a result, split in two. The `3-grams.correct` and `3-grams.errors` contain 3-grams recognized as correct and misspelled, respectively. Then the `3-grams.errors` list was iterated to find the nearest entry on the `3-grams.correct` list. The measure we use is the Levenshtein (editorial) distance [7]. The correction was applied if the distance between the misspelled trigram and correct trigram was less or equal 2. The threshold of 2 was set by the domain expert after his inspection of a sample of such corrections.

At this stage we faced a computational problem. The corpus is a collection of about 8 mln words. The building of the 3-grams.errors (about 0.3 mln entries) and 3-grams.correct (about 2 mln entries) databases proved to be unexpectedly quick. However, the performance of finding the closest match for each entry from 3-grams.errors in 3-grams.correct database

[1] GNU aspell, http://www.aspell.net/

[2] ISSN 0137-8910, http://www.ppoz.pl/

was very poor when choosing a simple approach: for each trigram in 3-grams.errors iterate over 3.grams.correct and calculate the Levenshtein distance between the two elements in each step. Database indexing of 3-grams.errors was not an options, since we didn't operate on exact matches, but needed to always calculate the difference.

Therefore we searched for a better method than scanning this large corpus and calculating the distance. The imaginary example below illustrates our solution:

Let us consider an example corpus of words

a-correct b-correct c-error d-correct e-error f-error.

For this corpus, there are 4 possible trigrams. Let us search for contexts contain the c-error:

(a) *a-correct b-correct c-error*

(b) *b-correct c-error d-correct*

(c) *c-error d-correct e-error*

A number of conclusions can be drawn from this observation: 1) the best context to fix the c-error is the (b)-trigram as it provides most likely the best (left and right) context for the misspelling, 2) the trigrams are redundant – it is enough to consider just one from the three above to have the c-error placed in the context, 3) the other two words in each trigram can be either correct or misspelled.

The third conclusion can be inspected further. At this stage the corpus contained around 2% of words with errors. The chances for two misspelled words occurring in one trigram seem low; assuming the misspellings are normally distributed across the corpus – there should be very few such trigrams. However, the assumption of normality proved to have a flaw, since in the population of humans, there are ones that tend to produce misspellings and others who do not. The result is that there occur trigrams with 2 misspelled words and less (but still) with all 3 words misspelled. Luckily, the prevailing majority of the trigrams were composed of one misspelling in the context of two correct words.

Considering the above, our approach proceeded as follows: the trigrams containing misspellings were split into two groups: i) a large group of trigrams with only one misspelled word, and ii) a small group of trigrams with two or three misspelled words.

Concerning the ii) group the plan was simple: the accepted Levenshtein distance was increased from 2 to $2 \cdot n$, where n is the number of misspelled words in the trigram. Then these trigrams were a subject to a linear scanning through the `3-grams.correct` database and because the small number of misspelled trigrams it proved not to be a computational issue.

In the group i) we started with our conclusion that trigrams are redundant. There are 3 setups for a misspelled word to be placed in the trigram, of which we choose the scenario (b) *b-correct c-error d-correct* (misspelling in the middle). The other two setups can be safely dropped since the goal of fixing the misspelled word can be achieved based on just a single context. The trigram was then reorganized into an associative array with the context as the key and the misspelled word as the value, i.e. *key="b-correct d-correct"* and *value="c-error"*. This structure later evolved: since between *b-correct* and *d-correct* more misspelled words may appear in the corpus, the value of the array should be a placeholder for more objects than just *c-error*. Therefore the final data structure has the form: *key="b-correct d-correct"*, *value="array('x1-error', 'x2-error', 'xN-error')"*. The same data structure was applied to the `3-grams.correct` database. The keys were hashed. The task has now become the searching for a hashed key of the misspelled trigram in the hashed keys of `3-grams.correct` database. Once the matching key is found, the Levenshtein distance between the given *xN-error* word and all the correct words (a small array) for the corresponding key in `3-grams.correct` database is calculated. This method proved to be very effective computationally and resolved the issue. The overall error rate dropped to under 1% after incorporating the ngrams method.

V. The Segmentation into Sentences and the Abbreviations

Another step in enhancing the nature of the date was the segmentation of corpus into sentences. It is important to note that the standard procedures of segmentation into sentences assume that the corpus is rather free from misspellings, that upper/lower case and other language rules are strictly obeyed – for such pure corpora the approaches like [8] could be more easily applied.

There is a couple of aspects related to sentences: 1) it is not proper to treat a dot as a terminator of a sentence since dots also appear in abbreviations 2) words which end with a dot may be not recognized by aspell either because they are misspelled or because they are correct domain abbreviations not known to aspell 3) for the n-grams analysis: trigrams should not cross the sentence boundaries 4) having the corpus segmented into sentences allows for enhanced further processing in more abstract layers. In many applications the sentences can be the smallest building blocks, e.g. in the Computer Aided Translation systems such as OmegaT [3] the sentences are atoms.

For the sake of simplifying our further considerations, let us introduce the terms *a sentence terminator* meaning *the last word of the sentence* and *a dotted word* meaning *word ending with a dot*.

We tried to automatically extract the abbreviations from the corpus. First we found all dotted words and sorted them by the number of the occurrences in the corpus. Table I is the header of the resulting list.

As this list extends there are less and less abbreviations, but we can not make any assumptions that after a certain position of this list there won't be any abbreviations. This is particularly true if we realize that the distribution of words in a text corpora is a Zipf distribution [9]:

"In human languages, word frequencies have a very heavy-tailed distribution, and can therefore be modeled reasonably well by a Zipf distribution (...)" [10].

[3] OmegaT, The free (GPL) translation memory tool, http://www.omegat.org

TABLE I
OCCURRENCES OF DOTTED WORDS

word	en translation	occurrences	abbreviation?
st.	fireman	54716	Y
ul.	street	40162	Y
C.	Celsius	39772	N
sprawny.	operating	26733	N
ok.	around	22184	Y
temp.	temperature	18295	Y
p.	floor	17087	Y
ůmieci.	garbage	13662	N
zdarzenia.	incident	12464	N
wody.	water	9706	N
budynku.	building	8097	N
lasu.	forest	7159	N
zach.	west	6368	Y
...			

The result from being a heavy-tailed distribution is that most words in the corpus (say 80%) appear relatively seldom (say 3 times). Taken the large number of the words, that means that we should be aware of the weakness of any manual action against the corpus, as we will only process a small portion of all the entities. Therefore, we look for a more automatic way of distinguishing the abbreviations from the sentence terminators.

We assumed that any sentence terminator may also appear in other position than at the end of the sentence, thus not end with a dot. Then we inspected the fraction: $f = w_d/(w_d+w)$, where w_d is *the frequency of occurrence a dotted word* and w is a *the frequency of occurrence of the same word without a dot*. f should return higher values for abbreviations. By inspecting the table II we can expect that the good threshold should be somewhere around 0.50 – higher values would be the abbreviations, lower values would be sentence terminators.

TABLE II
OCCURRENCES OF DOTTED WORDS AS A FRACTION OF $dotted/(dotted + notdotted)$. HIGH VALUES SHOULD INDICATE ABBREVIATIONS

word	en translation	fraction	abbreviation?
st.	fireman	0.99	Y
ul.	street	0.95	Y
C.	Celsius	0.10	N
sprawny.	operating	0.28	N
ok.	around	0.84	Y
temp.	temperature	0.89	Y
p.	floor	0.77	Y
ůmieci.	garbage	0.16	N
zdarzenia.	incident	0.09	N
wody.	water	0.14	N
budynku.	building	0.20	N
lasu.	forest	0.34	N
zach.	west	0.52	Y
...			

This method allows for pretty good results and abbreviations can be easily separated. However, a quick inspection of the full list reveals that false positives (not abbreviations) happen for values of above even 0.80. Therefore a few more features are added:

1. The number of characters in the word. Abbreviations should be short, that is an implicit part of their definition. The dotted words were getting benefit/penalty points for being short/long.

2. Position at the end of a paragraph indicates towards a sentence terminator. We added benefit points for each dotted word ending any paragraph.

3. Similarly, position directly before the beginning of a sentence indicates towards a sentence terminator. How to define the beginning of a sentence? The first idea was to treat any word beginning with an upper case as a likely beginning of a sentence. However, "kpt. John Snow" phrase quickly proves it is not entirely true. Instead, we built a list of bigrams starting with an upper case. Then we selected just the bigrams that occur often, more often than bigrams containing Names and Surnames – "John Snow" is not a frequent phrase in the corpus mentioning probably thousands of humans. Such frequent bigrams should very likely be the beginnings of the sentences. We added benefit points for the dotted word if it occurred before any beginning of a sentence (one occurrence is sufficient as it proves that such a word is a proper sentence terminator).

Finally we constructed the classifier based on the above features 1) the ratio of occurrence the dotted word with/without the trailing dot 2) the number of characters in the word 3) position at the end of a paragraph 4) position before the beginning of a sentence. The classifier was simply the sum of the 4 indicators, each of them normalized to <0,1> range. The list of sentence terminators was obtained and the corpus was segmented at each point were the sentence terminator with a trailing dot occurred. Manual browsing proved that this method was correct in about 97% cases, which seems a good score.

A. An example of the segmentation of a block of text into sentences

Let us illustrate our approach with the segmentation of an imaginary block of text: *Today temp. was 10 C. Strong wind from east. The fire was successfully put out.*

There are following dotted words to consider: temp, C, east, out. According to all of our considerations, the following would happen:

a) temp seldom appears without a dot in the EWID corpus. Some humans tend to write it without a dot (which is a mistake), but most write it properly and the statistics suggest it is an abbreviation. Our classifier correctly identified it as an abbreviation.

b) C is very short which suggest an abbreviation. Some firemen do follow C by a dot: "10 C." while it should be "10 °C". However, C is often spotted without the trailing dot, also spotted before the beginning of a sentence or even at the end of paragraphs which is a strong premise for a sentence terminator. Our classifier identified it as a sentence terminator.

c) east appears at the end of paragraphs and often without a dot. Our classifier identified it as a sentence terminator.

d) out appears at the end of paragraphs and often without a dot. Our classifier identified it as a sentence terminator.

The above block of text was therefore split into 3 sentences, which is correct.

B. Discussion

The knowledge-based data (text) correction method that we propose makes it possible to reduce error (typo) ratio from 4% down to below 1% (four-fold) in the EWID corpus. The cleansing/correction methods described above may also be tweaked for clearing the corpus from sensitive and private data. For example, there is an issue with sharing EWID corpus because it contains personal data (names, addresses, etc.) These sensitive data are not always easy to pinpoint, and the presented methods may help in this task, making anonymization of the text corpus feasible.

We also managed to quite successfully segment the corpus into sentences – the algorithm correctly proposed the endings of the sentences with about 97% accuracy. EWID system was hopefully carefully designed, but life often proves to find shortcomings in many designs, once these designs start to operate in the real world. The issue is that the designers seem to have lost their control over the content – there are difficulties in finding certain, fuzzy information (e.g. finding the information about all the accidents with the buses). Over the years EWID became the collection of lots of information in form of unstructured texts and it became a playground for researches like this one. One of our future ideas is to semantically inspect the content of EWID and sentences seem to be the proper building blocks for such an analysis. Once we have the sentences correctly defined we can cluster the whole corpus based on the sentences and then inspect the meaning (semantics) of each cluster. We believe that the system could improve the validation of the input by checking the input against it's knowledge base and then tag/correct/propose or otherwise interact with the human introducing the data.

What was learned from the experiment is a confirmation of [11]: "Usually the process of data cleansing cannot be performed without the involvement of a domain expert, because the detection and correction of anomalies requires detailed domain knowledge. Data cleansing is therefore described as semi-automatic but it should be as automatic as possible because of the large amount of data that usually is be processed and because of the time required for an expert to cleanse it manually. The ability for comprehensive and successful data cleansing is limited by the available knowledge and information necessary to detect and correct anomalies in data.".

The process of data cleansing has an iterative nature. Different aspects appear after the nature of data is better know, new thresholds must be checked, then parameters tweaked and then the whole process must be rerun. There is a difficulty with the order of the undertakings. On one hand we would like to start segmenting into sentences very early in the whole process of data cleansing. But at this time we would like to have the misspellings fixed already. In order to fix misspellings on the other hand, we use n-grams analysis which should not cross the sentences boundaries, but the sentences boundaries are not yet defined. We therefore need to run the analyses simultaneously and iteratively, as stated before. There is also the question whether bothering with data cleansing is worthwhile – the alternative is to accept that there is noise in the data (google and other search engines accept such noise after all). Answering the question of how much gain we achieve by cleansing the data would be possible after performing specific researches in higher level layers, e.g. the proposed CLEWID platform, where models operate on these lower level data. However, we didn't conduct such experiments.

ACKNOWLEDGMENT

Supported by Polish National Centre for Research and Development (NCBiR) – Grant No. O ROB/0010/03/001 in the frame of Defence and Security Programmes and Projects: "Modern engineering tools for decision support for commanders of the State Fire Service of Poland during Fire&Rescue operations in the buildings".

REFERENCES

[1] C. work, "Ewidencja zdarzeń - EWID99," Abacus, http://www.ewid.pl/, Tech. Rep., [Access: 23.04.2014].

[2] M. A. Hernández and S. J. Stolfo, "Real-world data is dirty: Data cleansing and the merge/purge problem," *Data mining and knowledge discovery*, vol. 2, no. 1, pp. 9–37, 1998.

[3] M. L. Lee, H. Lu, T. W. Ling, and Y. T. Ko, "Cleansing data for mining and warehousing," in *Database and Expert Systems Applications*. Springer, 1999, pp. 751–760.

[4] P. Elzinga, J. Poelmans, S. Viaene, G. Dedene, and S. Morsing, "Terrorist threat assessment with formal concept analysis," in *Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on*. IEEE, 2010, pp. 77–82.

[5] J. Poelmans, P. Elzinga, G. Dedene, S. Viaene, and S. Kuznetsov, "A concept discovery approach for fighting human trafficking and forced prostitution," *Conceptual Structures for Discovering Knowledge*, pp. 201–214, 2011.

[6] A. Krasuski, K. Kreński, P. Wasilewski, and S. Łazowy, "Granular approach in knowledge discovery," in *Rough Sets and Knowledge Technology*. Springer, 2012, pp. 416–421.

[7] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet physics doklady*, vol. 10, pp. 707–710, 1966.

[8] M. Rudolf and M. Świdziński, "Automatic utterance boundaries recognition in large polish text corpora," in *Intelligent Information Processing and Web Mining*. Springer, 2004, pp. 247–256.

[9] G. K. Zipf, "Selected studies of the principle of relative frequency in language." 1932.

[10] Wikipedia, "Zipf's law," http://en.wikipedia.org/wiki/Zipf's_law, [Access: 23.04.2014].

[11] H. Müller and J.-C. Freytag, *Problems, methods, and challenges in comprehensive data cleansing*. Professoren des Inst. Für Informatik, 2005.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 337–342
DOI: 10.15439/2014F375
ACSIS, Vol. 2

A MAT-based Granular Evacuation Modeling Framework.

Wojciech Świeboda
Institute of Mathematics
The University of Warsaw
Banacha 2,
02-097, Warsaw Poland

Andrzej Krauze
Institute of Computer Science
The Main School of Fire Service
Słowackiego 52/54,
01-629 Warsaw Poland

Hung Son Nguyen
Institute of Mathematics
The University of Warsaw
Banacha 2,
02-097, Warsaw Poland

***Abstract*—In this paper we describe an evacuation modeling framework based on a graph representation of the scene which is derived from its geometric description. Typically such graphs (geometric networks) are constructed using Medial Axis Transform (MAT) or Straight Medial Axis Transform (S-MAT). In our work we use Voronoi tessellation of a set of points approximating the scene (a single floor plan) along with the dual graph – Delaunay triangulation. Using these two graphs we extract not only the information about paths in the building, but also information about path widths and areas assigned to vertices. Typically only path lengths from MAT or S-MAT based geometric networks are used in evacuation modeling. Our approach enables us to include flow analysis and e.g. locate bottlenecks. We discuss a typical density-based evacuation model coupled with a partial behavioral evacuation model within proposed framework.**

I. Introduction

ICRA project (http://icra-project.org/) aims to provide modern engineering tools to support fire commanders during firefighting and rescue operations.

As a module in this project, we aim to build a framework which would provide the basis for implementation of density-based (flow-based) evacuation models. While density-based models only represent one approach to evacuation modeling [1], we nevertheless stress that they are very similar to hand calculations carried out using methods described in [2], [3] and thus are the easiest for experts to understand and assess. Advantages and limitations of different methodological approaches have been discussed in [4].

While the assumptions of specific models, their degree of complexity as well as their specifics may differ, all of these models will operate on the same underlying representation of the scene. On one hand, the scene is represented by a building model which describes geometric properties, on the other hand, it is represented by a graph that describes the topology of connections in the building. Such a graph is called a *geometric network* [5].

Vertices in a geometric network correspond to so-called transition points, whereas edges correspond to paths. Geometric networks are widely used in geographic information systems [5], and indoor navigation systems [6]. A computational geometry tool that has been often used in construction of geometric networks from planar plans is *Medial Axis Transform* (MAT) or *Topological Skeleton*, introduced by Blum [7].

Indoor navigation models for the purpose of emergency movement were previously studied in [8], [9], [10]. Building representations that supplement each other seem to be a common trend in the literature, see e.g. [11], [12], [13].

In this paper we describe an approximate process of constructing a geometric network that is further enriched: Vertices are assigned to areas in the building and edges are enriched by information about path lengths, widths, and stairs parameters that affect maximum flows. We also discuss an implementation of a typical density-based model [2] coupled with a partial behavioral model based on PD 7974-6 norm [3] within the proposed framework.

The outline of our paper is as follows: First we briefly describe the process of evacuation and the role of evacuation modeling in ICRA project. Afterwards we introduce two evacuation models: one includes a behavioral component, the other one is strictly a (density based) movement model. In the following sections we briefly discuss Building Information Modeling (BIM) and the process of geometric network calculation in our model.

Granularity announced in the title will become apparent when we discuss the duality of vertices in the graph and areas in the geometric description of a building.

II. Evacuation process

Evacuation models are typically implemented in the context of building safety analysis. In this usage scenario, which needs to address the worst-case usage pattern of a building, one is usually interested in comparison of available safe egress time (ASET) and required safe egress time (RSET). ASET is the period of time which permits safe escape from a building. RSET, on the other hand, is the time between ignition of fire and the completion of evacuation. Methods of ASET estimation are outside of the scope of actual evacuation modeling, although some models (e.g. buildingEXODUS [14] and FDS+EVAC [15]) enable joint fire and evacuation calculations or inclusion of fire simulation results in evacuation

This work was partially supported by the Polish National Centre for Research and Development (NCBiR) – grant O ROB/0010/03/001 under Defence and Security Programmes and Projects: "Modern engineering tools for decision support for commanders of the State Fire Service of Poland during Fire&Rescue operations in buildings", and by the Polish National Science Centre grants DEC-2011/01/B/ST6/03867 and DEC-2012/05/B/ST6/03215.

calculations [1]. In what follows, we remind the typical [3], [2] model of RSET calculation. Following [2] and [3] (from which we borrow the notation) we define:

$$RSET = \Delta t_{\text{det}} + \Delta t_a + \Delta t_{\text{evac}}$$

where Δt_{det} denotes the time between fire ignition and detection, Δt_a is the time of alarming, and t_{evac} is the actual evacuation time. Δt_{evac} is further decomposed as:

$$\Delta t_{\text{evac}} = \Delta t_{\text{pre}} + \Delta t_{\text{trav}}$$

where Δt_{pre} is pre-movement time and Δt_{trav} is travel time. Pre-movement time consists of recognition time (the time it takes an alarm aware occupant to take action) and response time (additional time it takes the occupant before he starts walking towards exit).

III. Evacuation model as a decision support system

There are important differences between typical usage of evacuation models and the usage scenarios considered in ICRA project: In building safety analysis, one typically aims to analyse potential worst-case scenarios, whereas during a Fire Rescue Action the placement of occupants may be known, and the model is used to assess this specific situation.

In ICRA project, the end user is the fire commander. Two primary use cases that we consider are direct assistance in Search and Rescue operations and providing a rough estimate of egress time along with bottleneck analysis. Usage scenarios differ mainly in assumptions of occupant localization. We cover these from the least to the most specific:

- In most situations, rough occupant density assumptions can be made based on domain knowledge. For example, if a fire alarm is triggered at a school at 9am, we can expect the highest overall density, with most classrooms utilized. Further assumptions about typical class size (e.g. 20 pupils) can lead to accurate evacuation time estimates. A uniform density in all rooms may be assumed without further clues.

The following two points are still an area of research, but we nevertheless stress them now as the current roadmap of our research:

- If technology permits, we may infer vague hints as to density placement of people in different parts of a building based e.g. on cellular traffic or information from other sources. Thus, we may rule out certain bottlenecks that would not be apparent if we assumed an overall uniform density of people in the building.
- We also consider a very specific scenario where only few occupants are left in the building, and their locations are (approximately) known. The module could provide hints for navigation of fire fighters and directly support Search and Rescue operation rather than provide egress time estimation.

In this paper we only consider the first usage scenario: we assume that occupants are uniformly placed in the building. We wish to estimate total egress time and determine possible bottlenecks in building structure. The model should also provide a plausible forecast for a given timestamp.

IV. Implementation of PD 7974-6 norm

PD 7974-6 norm by British Standards Institute [3] describes an algorithm of RSET calculation that encompasses two scenarios: a sparsely populated and a densely populated building. If the building is densely populated, first occupants will usually have shorter pre-movement times than in the other scenario [3], but a queue may form quickly afterwards, limiting the outgoing flow. If the building is sparsely populated, the movement is unobstructed, but pre-movement times of last occupants may be longer (e.g. there may be nobody around to notify them). Thus, the overall evacuation time for the first scenario may be calculated as:

$$\Delta t_{\text{trav}}^{\text{sparse}} = \Delta t_{\text{pre}(99)} + \Delta t_{\text{trav(walking)}}$$

and for the second scenario as:

$$\Delta t_{\text{trav}}^{\text{dense}} = \Delta t_{\text{pre}(1)} + \Delta t_{\text{trav(walking)}} + \Delta t_{\text{trav(flow)}}$$

where:

- $\Delta t_{\text{pre}(99)}$ is the pre-movement time of the 99th percentile of occupants in a sparsely populated building,
- $\Delta t_{\text{pre}(1)}$ is the pre-movement time of the first percentile of occupants in a densely populated building,
- $\Delta t_{\text{trav(walking)}}$ is the unimpeded evacuation time of an occupant with the longest path to the exit.
- $\Delta t_{\text{trav(flow)}}$ is the queuing time of occupants at the exits.

Usually $\Delta t_{\text{trav}}^{\text{sparse}} < \Delta t_{\text{trav}}^{\text{dense}}$, although in some situations the opposite may hold true, hence both scenarios are usually considered.

PD 7974-6 also describes the calculation procedure of pre-movement times for several behavioral scenarios and building types (characterized by their complexity, management level and alarm system). For the rest of this paper we assume that parameters required for calculation of pre-movement times (according to the discussed model) can be provided by an operator of Fire Command Center.

It is worth stressing that in this setting, alarm times and pre-movement times for certain scenarios are intervals, thus the output of the discussed evacuation model is not a single number, but a 2×4 matrix that describes the sparsely-populated and densely-populated scenarios, e.g.:

scenario	Δt_a	Δt_{pre}	$t_{\text{trav(walk)}}$	$t_{\text{trav(flow)}}$
sparse	$2-5$	>20	10	0
dense	$2-5$	>10	10	15

Fig. 1. The output of the model consists of time (in minutes) of alarming-time, pre-movement time and evacuation time (walking and queuing) for different scenarios. Pre-movement time in the sparse scenario corresponds to the last occupants (99th percentile), whereas in the dense scenario it corresponds to the first occupants (1st percentile).

In this setting, Δt_{pre} corresponds to inter-percentile range of pre-movement times of all occupants in the building. Pre-movement time of a single occupant in the building typically

follows approximately log-normal or normal distribution [3] (log-normal), [16] (log-normal or normal). It is worth stressing that other approaches to pre-movement time modeling are possible, see e.g. [16] for a discussion of a sampling-based approach, or overview of various other approaches in [1].

When hand calculations are used, $\Delta t_{trav(walk)}$ is usually approximated by the longest route to exit multiplied by a conservative estimate of human speed, whereas $\Delta t_{trav(flow)}$ is usually estimated by finding the dominating bottleneck in evacuation plan and performing calculations for this bottleneck alone.

V. Flow Model

The simplest model of $\Delta t_{\text{trav(flow)}}$ calculation hinted in the previous section requires determining the dominating bottleneck, which in turn requires analysis of the overall flow of people in the building. Since we are not hindered by time constraints that enforce simplicity (and approximations) of hand calculations, we describe a somewhat more general scheme. In what follows, we remind a flow model described in [2]. We begin by reminding an informal definition from the paper:

Consider a set of evacuation paths in a building. A *transition point* is any point where (i) a path becomes narrower or wider; or (ii) paths split, converge or join stairs.

From now on we consider a graph (V, E) such that vertices V represent transition points in a building and edges E represent evacuation paths. We will think of *flow* of people through edges $e \in E$, and thus for clarity we will assume that graph $G = (V, E)$ is directed. The unit of flow in this model is $\left[\frac{\text{persons}}{s \cdot m}\right]$ (flow denotes the number of people that pass through a corridor or a door of a given width in a time interval).

For each transition point $v \in V$:

$$\sum_{e=\langle v,w\rangle \in E} F_e W_e = 0$$

where F_e denotes flow departing (or arriving, if negative) from (or at) v through e and W_e is the minimum effective width of path $e \in E$. Furthermore, [2] defines maximum flow through e for horizontal travel and for stairs with different parameters (Riser and Tread) as a function of occupant density.

We remind that t_a, t_{pre} and $t_{\text{trav(walk)}}$ are calculated separately and the purpose of flow analysis is to determine $t_{\text{trav(flow)}}$ component only.

We may assume that queuing occurs and that a queue is already formed. For this reason, flow in this network can be approximated by optimization algorithms for flow networks ([17], [18]), by assuming fixed values of maximum flows through edges and iterating further calculations over vertices corresponding to consecutively depleted sources.

We stress that in the simple formulation above we make several gross simplifications, for example:

- we do not take into account the effects of fire or smoke,
- we assume that occupants act semi-rationally so as to reach (locally, at each moment) the maximum flow out of the building (in particular, no congestion points are formed at dominating bottlenecks),
- effective widths of corridors are estimated (using the floor plan) by actual corridor widths (i.e., we are ignoring potential obstacles not directly described on the floor plan),

However, all of these simplifications (and various other points) can be addressed within the same framework by extending the basic model.

VI. BIM and indoor navigation models

Afyouni et al. [11] presents a taxonomy of indoor spatial models proposed in the literature. The taxonomy is briefly summarized on Fig. 2. Authors of the paper further stress that hybrid spatial models that combine geometric and symbolic approaches may complement each other in various applications.

Building Information Modeling (BIM) is a general framework of representing, archiving and processing information about buildings in a structured digital format. One specific format which is widely used in practice is Industry Foundation Classes, and particularly IFC2x3 (IFC4 is published as ISO 16739, but IFC2x3 is still prevalent in practice today). Indoor navigation models for buildings represented in IFC were previously studied in e.g. [19], [20], [21]. IFC2x3 models combine the boundary-based geometric representation and an object-oriented (symbolic) model: typical objects in IFC2x3 files are storeys, walls or doors, along with information about their placement.

While IFC2x3 files may contain topological/graph-based representation of a building (described in terms of entities *IfcPath*, *IfcEdge* and other entities of supertype *IfcTopologicalRepresentationItem*), such graph-based representations are usually missing in IFC files exported by CAD tools.

Thus, in order to define the graph required for Flow Model calculations mentioned in the previous section, we transform the input BIM file to a geometric network, a structure that encompasses the geometric part defined in directly BIM and the topology which we infer from geometric representation. The topology is a graph whose vertices correspond to transition points.

From the perspective of evacuation modeling, a classification of the underlying grid or structure of the floor plan is discussed in [1] for various models. Authors discuss a fine network, a coarse network and a continuous network geometry. From the perspective of this ontology, the geometric network on which our model operates is a coarse network.

VII. Geometric Network Calculation

Usually Medial Axis Transform (MAT) or Straight-Medial Axis Transform (S-MAT) [5] is used to define a geometric network. See [6] for a discussion of other algorithms and [9] for an example of an alternative approach of indoor navigation that does not require topology construction. A Medial Axis of a set $F \subseteq V$ is the set of points $M \subseteq V$ that have at least two

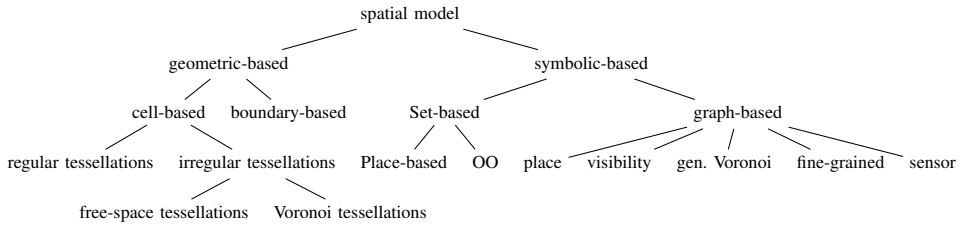


Fig. 2. Taxonomy of indoor spatial models presented by Afyouni et al. [11]

closest neighbours in F (see Fig. 3). If F is finite, Voronoi diagram [22] of F is the Medial Axis of F.

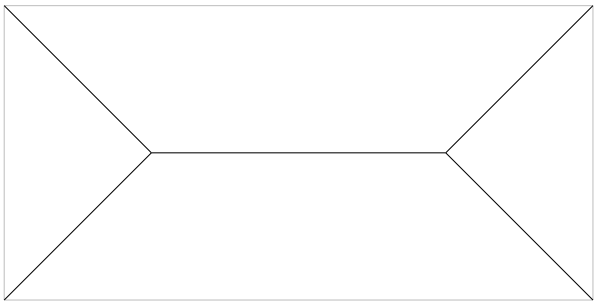


Fig. 3. Medial Axis Transform of the gray rectangle consists of five line segments shown on the picture.

Consider a plan of a floor in a building $F \subset R^2$ (Fig. 4). Instead of calculating MAT of F directly, we approximate F by a finite set of points S (Fig. 5) and calculate the Voronoi diagram of S, which consists of line segments. Denote by (V', E') the graph that consists of subset of line segments from Voronoi triangulation that do not intersect F (see Fig. 7). Edges in graph (V', E') describe permissible paths in our model.

Delaunay triangulation of S is the dual graph of the Voronoi diagram of S (Fig. 6). We utilize this triangulation in two ways: We use it to approximate the minimum width of a path ($e \in E$) by the shortest line segment in Delaunay triangulation that intersects e. Secondly, triangles in Delaunay triangulation are assigned to vertices $v \in V$ and provide a partitioning of the geometric view of the scene.

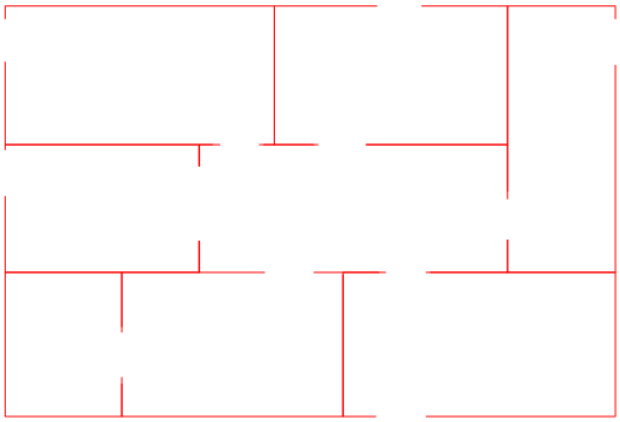


Fig. 4. A 2-dimensional slice of the building that represents a single floor (with doors removed).

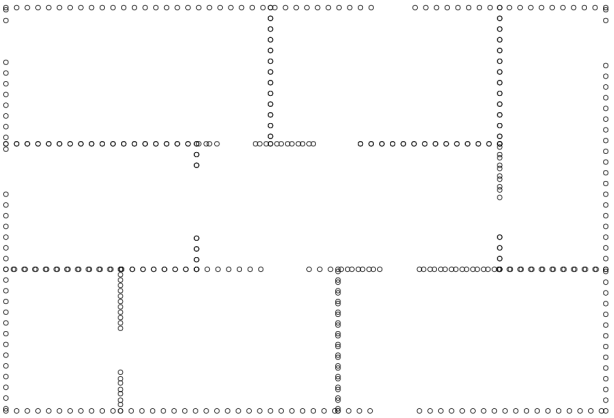


Fig. 5. Approximation of line segments by a set of points.

VIII. Conclusions and Future Work

In this paper we have discussed a framework for evacuation modeling based on building topology extraction from the building. We supplemented the typical graph representation of a floor plan derived from MAT by information obtained from Delaunay triangulation of the point set that approximates a single floor: path widths and areas assigned to vertices.

We have mentioned various areas of future research throughout our paper:

- Design of evacuation models based on different assumptions of occupant localization. Our current research focuses on localization of people within buildings.
- More refined models, e.g. taking into account effects of fire or smoke.
- Staircases are often bottlenecks during evacuations, which suggests a more detailed modeling of the effect of different types of staircases and their parameters on movement speeds of crowds.

Other possible areas of future research are:

- Specification of evacuation scenarios in a dialogue with the user (during the ride to fire scene). The dialogue necessarily needs to be very limited, but it could aid

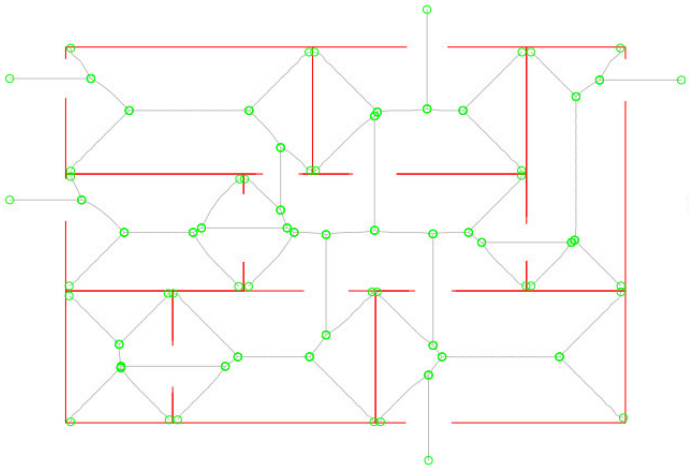


Fig. 8. A graph (V, E) resulting from contraction of edges of degree 2 in graph (V', E').

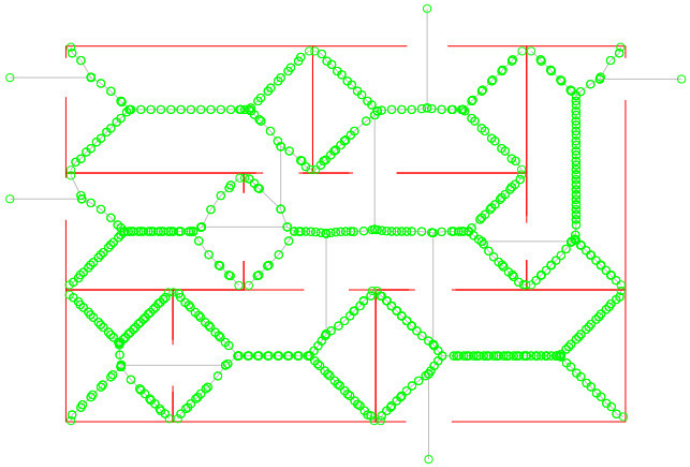


Fig. 7. A subset of line segments from Voronoi tessellation determines permissible paths. This is the initial graph (V', E').

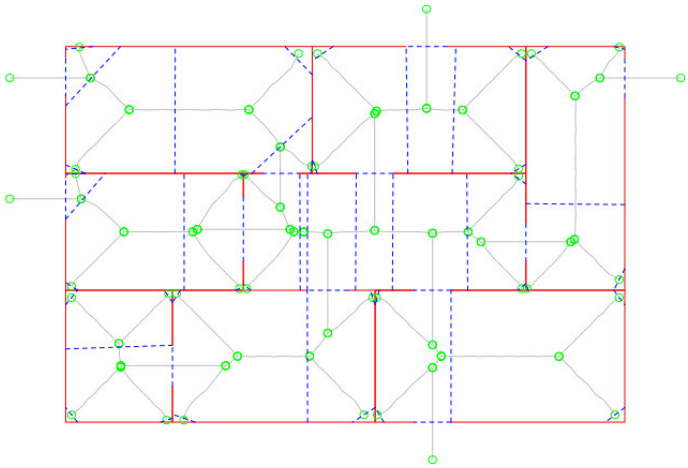


Fig. 9. Selected edges (dashed, blue) from Delaunay triangulation determine widths of paths described by edges in the graph. Delaunay triangulation also provides a mapping of points on the floor plan to corresponding vertices (though some triangles may contain a few vertices $v \in V$). Vertices in corners have very small areas assigned to them and the ratio of width to area size of such vertices is relatively high, thus they are not bottlenecks.

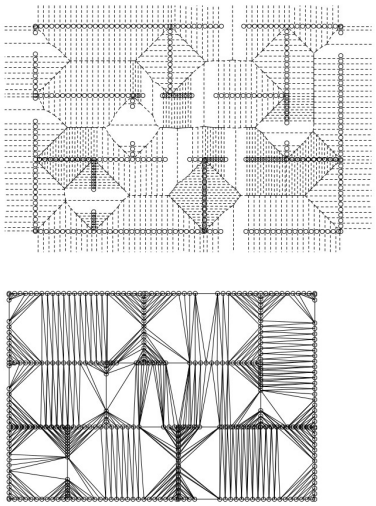


Fig. 6. Voronoi tessellation and Delaunay triangulation of a floor plan approximated by a set of points.

the commander in specifying initial plans of action better than passive information delivery. On rare occasions some hints may be also provided by the operator of Fire Command Center, e.g. unavailability of certain exits.

- We consider performing MAT or S-MAT calculations for floor plans consisting of line segments. In our preliminary experiments we have approximated the floor plan by a set of points so as to utilize the duality of Voronoi tessellation and Delaunay triangulation. The approximation by a discrete set may not be necessary.
- Addressing the problem of counter-flows, i.e. the interaction of fire-fighters getting into the building with the occupants trying to get out of the building.

REFERENCES

[1] E. D. Kuligowski, R. D. Peacock, and B. L. Hoskins, "A review of building evacuation models, 2nd edition," National Institute of Standards and Technology, Technical Note 1680, 2010. [Online]. Available: http://evacmod.net

[2] S. M. V. Gwynne and E. Rosenbaum, in SFPE Handbook of Fire Protection Engineering, 4th ed., P. J. DiNenno, D. Drysdale, and C. L. Beyler, Eds. Quincy, MA: National Fire Protection Association, 2008, ch. Employing the Hydraulic Model in Assessing Emergency Movement, pp. 373–395 of section 3.

[3] The application of fire safety engineering principles to fire safety design of buildings: part 6 : Human factors: Life safety strategies : occupant evacuation, behaviour and condition (sub-system 6). London: BSI, 2004.

[4] L. M. Zheng Xiaoping, Zhong Tingkuan, "Modeling crowd evacuation of a building based on seven methodological approaches." Building and Environment, vol. 44, pp. 437–445. [Online]. Available: http://dx.doi.org/10.1016/j.buildenv.2008.04.002

[5] J. Lee, "A spatial access-oriented implementation of a 3-D GIS topological data model for urban entities." GeoInformatica, vol. 8, no. 3, pp. 237–264, 2004. [Online]. Available: http://dblp.uni-trier.de/db/journals/geoinformatica/geoinformatica8.html#Lee04

[6] N. P. M. Meijers and S. Zlatanova, "3D geo-information indoors: Structuring for evacuation." in Proceedings of the Joint International ISPRS, EuroSDR, and DGPF Workshop on Next Generation 3D City Models., 2005, pp. 21–22. [Online]. Available: http://www.gdmc.nl/publications/2005/3D_indoor_geoinformation.pdf

[7] H. Blum, "A Transformation for Extracting New Descriptors of Shape," in Models for the Perception of Speech and Visual Form, W. W. Dunn, Ed. Cambridge: MIT Press, 1967, pp. 362–380.

[8] Y. Lee and S. Zlatanova, Geospatial Information Technology for Emergency Response: International Society for Photogrammetry and Remote Sensing, 1st ed. Bristol, PA, USA: Taylor & Francis, Inc., 2008, ch. A 3D data model and topological analyses for emergency response in urban areas., pp. 143–168.

[9] L. Liu and S. Zlatanova, "A 'door-to-door' path-finding approach for indoor navigation," in International Archives ISPRS XXXVIII, 7th Gi4DM, B. . Z. Antalya, Backhause, Ed., Jun. 2011.

[10] S. Pu and S. Zlatanova, "Evacuation route calculation of inner buildings," in Geo-information for Disaster Management, P. Oosterom, S. Zlatanova, and E. Fendel, Eds. Springer Berlin Heidelberg, 2005, pp. 1143–1161. [Online]. Available: http://dx.doi.org/10.1007/3-540-27468-5_79

[11] I. Afyouni, C. Ray, and C. Claramunt, "Spatial models for context-aware indoor navigation systems: A survey." J. Spatial Information Science, vol. 4, no. 1, pp. 85–123, 2012. [Online]. Available: http://dblp.uni-trier.de/db/journals/josis/josis4.html#AfyouniRC12

[12] L. Liu and S. Zlatanova, "A semantic data model for indoor navigation," in Proceedings of the Fourth ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness, ser. ISA '12. New York, NY, USA: ACM, 2012, pp. 1–8. [Online]. Available: http://doi.acm.org/10.1145/2442616.2442618

[13] B. Lorenz, H. Ohlbach, and E.-P. Stoffel, "A hybrid spatial model for representing indoor environments," in Web and Wireless Geographical Information Systems, ser. Lecture Notes in Computer Science, J. Carswell and T. Tezuka, Eds. Springer Berlin Heidelberg, 2006, vol. 4295, pp. 102–112. [Online]. Available: http://dx.doi.org/10.1007/11935148_10

[14] E. Galea, P. Lawrence, S. Gwynne, L. Filippidis, D. Blackshields, and D. Cooney, buildingEXODUS v6.0, University of Greenwich, Greenwich, London, October 2013.

[15] T. Korhonen and S. Hostikka, Fire Dynamics Simulator with Evacuation: FDS+Evac – Technical Reference and User's Guide, Julkaisija-Utgivare, VTT Working Papers 119, 2009.

[16] E. Ronchi, E. D. Kuligowski, P. A. Reneke, R. D. Peacock, and D. Nilsson, "The process of verification and validation of building fire evacuation models," National Institute of Standards and Technology, Tech. Rep. 1822, 2013.

[17] L. R. Ford and D. R. Fulkerson, "Maximal Flow through a Network." Canadian Journal of Mathematics, pp. 399–404. [Online]. Available: http://www.rand.org/pubs/papers/P605/

[18] T. H. Cormen, C. Stein, R. L. Rivest, and C. E. Leiserson, Introduction to Algorithms, 2nd ed. McGraw-Hill Higher Education, 2001.

[19] Y.-H. Lin, Y.-S. Liu, G. Gao, X.-G. Han, C.-Y. Lai, and M. Gu, "The IFC-based path planning for 3D indoor spaces." Advanced Engineering Informatics, vol. 27, no. 2, pp. 189–205, 2013. [Online]. Available: http://dblp.uni-trier.de/db/journals/aei/aei27.html#LinLGHLG13

[20] C. de Haas and M. Boysen, "The journey from IFC files to indoor navigation," Master's thesis, Aalborg University, 2013.

[21] A. Pilvinytė, "Middleware-free approach for indoor space shortest path queries," Master's thesis, Aalborg University, 2013.

[22] G. L. Dirichlet, "Über die Reduktion der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen," Journal für die Reine und Angewandte Mathematik, vol. 40, pp. 209–227, 1850.

AAIA'14 Data Mining Competition at the Knowledge Pit

KEY RISK factors for Polish State Fire Service is organized within the framework of the 9th International Symposium on Advances in Artificial Intelligence and Applications (AAIA'14), and is an integral part of the 1st Complex Events and Information Modelling workshop (CEIM'14) devoted to the fire protection engeneering. The task is related to the problem of extracting useful knowledge from incident reports obtained from The State Fire Service of Poland. Prizes worth over 3,000 USD will be awarded to the most successful teams. The contest is sponsored by Dituel Sp. z o.o. and F&K Consulting Engineers, with a support from The University of Warsaw and ICRA project.

Introduction

Incident Data Reporting Systems (IDRS) are used by public safety services across the globe to gather information about the incidents which required their actions. This information is used not only to simply document the events but it can also be incorporated into the training of new officers. Moreover, the knowledge extracted from such reports can help in better identification of threats and in planning of more effective procedures. An example of such a reporting system is EWID which is used by the State Fire Service of Poland. In the proposed competition, we would like to raise the problem of extracting useful knowledge from the reports generated in the EWID system, represented in a form of a data table. In particular, we would like to ask the participants to identify key factors influencing the risk of serious injuries among firefighters and people involved in various incidents. The contest will be hosted on a web platform called Knowledge Pit, designed especially for supporting organization of data mining competitions associated with scientific conferences.

Special session at CEIM'14 workshop

A special session devoted to the competition will be held at 1st Complex Events and Information Modelling workshop (CEIM'14) which is a part of 9th International Symposium on Advances in Artificial Intelligence and Applications (AAIA'14). We will invite authors of selected reports to extend them for publication in the conference proceedings (after reviews by Organizing Committee members) and presentation at the conference. The invited teams will be chosen based on their final rank, innovativeness of their approach and quality of the submitted report.

Awards

Authors of the top ranked solutions will be awarded with valuable prizes:

1. First Prize: computer hardware worth 2,000USD + one free FedCSIS'14 registration,
2. Second Prize: computer hardware worth 1,000USD + one free FedCSIS'14 registration,
3. Third Prize: one free FedCSIS'14 conference registration.

The award ceremony will take place during the FedCSIS'14 conference (September 7-10, Warsaw). Additionally, authors of all papers accepted for presentation at the CEIM'14 workshop, who decide to attend the conference will receive a diploma and a competition T-shirt.

Contest Organizing Committee

Andrzej Janusz (Chairman), University of Warsaw
Adam Krasuski, Main School of Fire Service & University of Warsaw
Dominik Ślęzak, University of Warsaw & Infobright Inc.
Hung Son Nguyen, University of Warsaw
Sebastian Stawicki, University of Warsaw
Guillermo Rein, Imperial College London
Stanisław Łazowy, Main School of Fire Service

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 345–354
DOI: 10.15439/2014F507
ACSIS, Vol. 2

Key Risk Factors for Polish State Fire Service: a Data Mining Competition at Knowledge Pit

Andrzej Janusz*, Adam Krasuski†, Sebastian Stawicki*, Mariusz Rosiak,
Dominik Ślęzak*‡ and Hung Son Nguyen*

*Institute of Mathematics, University of Warsaw
ul. Banacha 2, 02-097 Warsaw, Poland
{janusza,slezak,son,stawicki}@mimuw.edu.pl
mariusz.rosiak@gmail.com

*Section of Computer Science, The Main School of Fire Service
ul. Słowackiego 52/54, 01-629 Warsaw, Poland
krasuski@inf.sgsp.edu.pl

†Infobright Inc.
ul. Krzywickiego 34, lok. 219, 02-078 Warsaw, Poland

***Abstract*—In this paper we summarize AAIA'14 Data Mining Competition: Key risk factors for Polish State Fire Service which was held between February 3, 2014 and May 5, 2014 at the Knowledge Pit platform http://challenge.mimuw.edu.pl/. We describe the scope and background of this competition and we explain in details the evaluation procedure. We also briefly overview the results of this analytical challenge, showing the way in which those results can be beneficial to one of our other projects which is related to the problem of improving firefighter safety at a fire scene. Finally, we reveal some technical details regarding the architecture and functionalities of the Knowledge Pit competition platform, which we are developing in order to facilitate solving of practical problems that require advanced data analytics.**

***Keywords*-data mining competition, risk factors, attribute selection, EWID system**

I. Introduction

INCIDENT DATA REPORTING SYSTEMS (IDRS) are used by public safety services across the globe to gather information about the incidents which required their actions. The information is gathered in order to calculate statistics within the groups of incidents, identify peculiar cases and to improve the procedures [1]. Results of a thorough analysis of incident reports can also be utilized by decision support systems to increase safety of firefighters at a fire scene [2].

EWID is an example of such a reporting system. It is used by the State Fire Service of Poland [3]. A report submitted to the system by Incident Commander (IC - a coordinating officer) after a fire and rescue action (F&R) consists of two parts: a quantitative description, where facts regarding the action are expressed by numerical or categorical characteristics and a description in a natural language. The first part is often called the attribute section and the second is the descriptive section.

The attribute section is represented in a form of structured and quantified characteristics. Among over 500 attributes, it contains information about incident type, its severity or size and resources involved in the response. The descriptive section can be treated as an extension to the attribute section. It contains a natural language description of probable causes, conditions at the event scene and the course of the action. Figure 1 depicts a chunk of a report submitted to the EWID system.

It is assumed that the descriptive section should contain all the relevant information which could not be expressed in the attribute section. However, due to the fact that there are no instructions regarding what information is relevant in a context of a particular incident type, the descriptive section sometimes contains irrelevant and useless fragments of text. A quality of the textual descriptions in the system also variates, depending on a personality and attitude of IC who writes the report. For instance, a content of the part devoted to the course of the action may range from very useful information concerning the consecutive decisions of IC, applied techniques and their consequences, to very cursory and ambiguous sentences such as: *rubbish lit*.

On the other hand, the attribute section is unable to reflect all information regarding a very large spectrum of possible incidents. All the above mentioned shortcomings make it challenging to extract useful information from the EWID reports, especially when this information is only indirectly related to the set of characteristics from the attribute section [4]. One example of a task that requires such information is the problem of recognition of risk factors which affect the possibility of a serious injury or death among firefighters and other people involved in various incidents. A similar problem, i.e. the identification of threats, has been already investigated by several researchers [3], [5], [6].

In the research presented in this paper we address the above mentioned challenge. We decided to ask the machine learning community to identify characteristics extracted from the EWID reports, which are useful for predicting whether any people were harmed during a given incident. For this

978-83-60810-58-3/$25.00

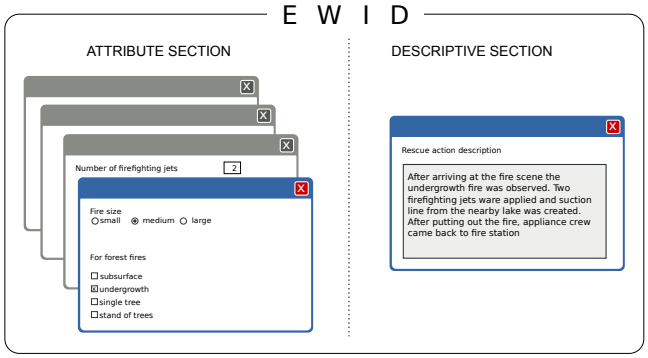


Fig. 1. The chunk of the EWID report.

purpose, we devised a data mining competition, titled *AAIA'14 Data Mining Competition: Key risk factors for Polish State Fire Service*. A form of this challenge was similar to the competitions which our team organized in the past [7], [8] on the TunedIT platform [9]. This time, however, we organized it on a novel web platform called Knowledge Pit (www.challenge.mimuw.edu.pl) hoping that the participants will be able to enrich our understanding of the EWID data and point at attributes that describe the most relevant information to the stated problem.

In the following sections we reveal details regarding the architecture of Knowledge Pit (Section II) and then, in Section III, we describe the proposed challenge. Next, in Sections IV and V, we present an overview of results obtained by participants of the competition and analyze those results with respect to the semantic types (i.e. the meaning) of attributes that were frequently appearing in the submitted solutions. Finally, we conclude the paper by drawing our plans for a continuation of this study.

II. The Knowledge Pit Platform

Knowledge Pit is a platform created to support organization of data mining challenges. It is designed in a modular way, on top of an open-source e-learning platform Moodle.org [10], to follow the best practices of a software development. Therefore, the platform with its current modules, including user accounts, challenges and resources management subsystems, time and calendar functionalities, communications features (i.e. forums and messaging subsystems), and a flexible interface for connecting automated judging services prepared to evaluate contestants' submissions, is conceptually ready to introduce new features or enhance the existing ones.

A more detailed architecture overview requires to describe two main parts of the system. All elements that are available to the users interested in participating in a data mining competition, together form a web user interface. To fulfill this requirement, Knowledge Pit utilizes a very popular solution stack Apache/MySQL/PHP – a set of software components that is sufficient to provide web solutions ranging from simple to complex ones [11], [12], [13]. The second part of the system concerns competition handling from the point of view of evaluating the submitted solutions. This functionality is separated from the remaining part of the platform to cope with the requirement for high flexibility (with regard to a programming language or a framework, parallelization of expensive calculations, etc.) of the judging software setup. The general architecture of the system is presented in Figure 2.

From the point of view of Knowledge Pit system there are several roles which can be assigned to a user – a guest, a contest participant or a contest organizer role. Therefore, the front-end engine consists of several modules which provide the functionalities to the users, depending on their role in a given moment. The main modules of the system are as follows:

- user management and user privileges
- challenge maintenance
- challenge Leaderboard
- challenge submissions manager
- calendar
- forum module
- internal messaging system
- chat
- private resources (files) repository

The above modules can be thought of as pieces of software that implement specific elements of the system. When combined, they constitute the higher-level features described below in this section.

A. User interface

Knowledge Pit implements users and user groups management, an advanced privileges support and an enhanced context handling, e.g. a user can be a guest in a given challenge, a participant in other and also a creator and manager of another one. The site administrators can manually promote or demote users access corresponding to any of the given contexts, e.g. a context of the page a user can browse. This means that the administrators can grant privileges in a local context (e.g. a forum of a specific contest, a chat, etc.) leaving the access privileges to the other parts of the site unchanged. Moreover, a special registry is used to administer the users and the user groups. If necessary, a new user type or users group can be created with selected privileges granted, thereby facilitating the task of managing large number of users in some particular contexts.

Each user is assigned to a set of assets such as a private file repository, a dedicated calendar with adjustable scope and event levels, a public profile shown to others in contexts of chats or forums, and a personalized site appearance – a set of settings that allows to adjust the way how the site looks, e.g. a user can hide or move specific parts of menus and navigation modules, or use predefined site themes, all accordingly to his own choice.

Each site visitor can view a calendar on which events are displayed accordingly to the access level and the site context. The calendar is fully customizable and has events ordered according to scopes:

- global

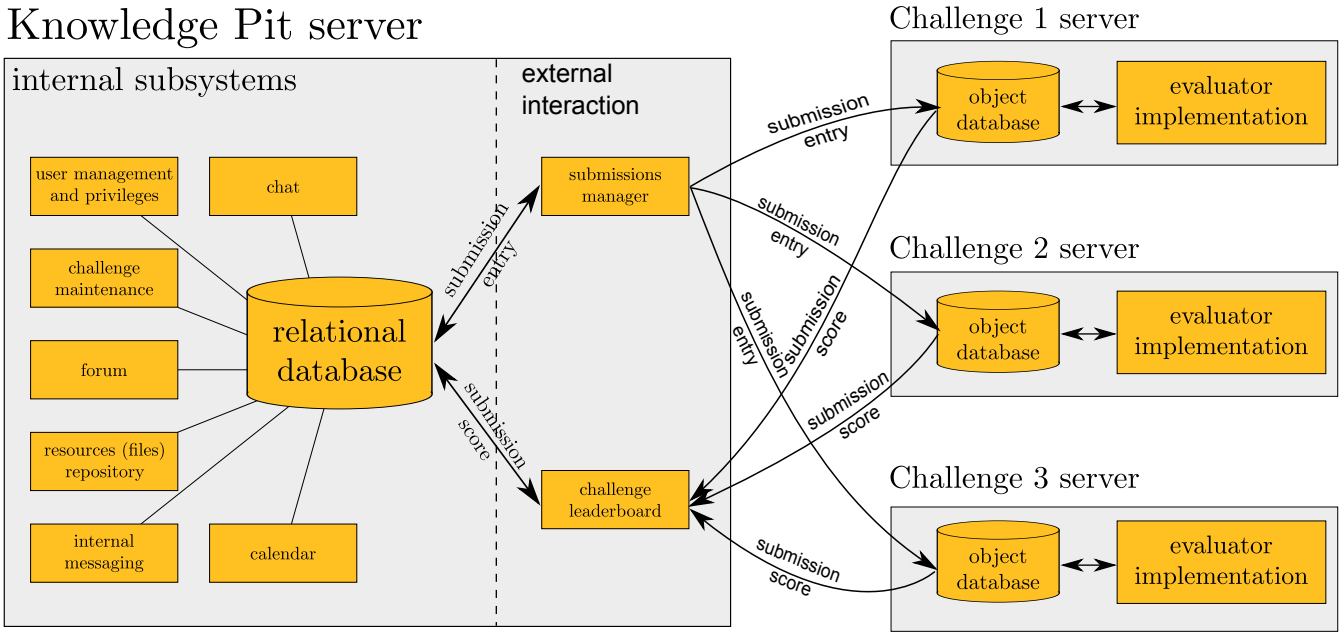


Fig. 2. A system architecture of the Knowledge Pit web platform.

- competition
- group
- user

Apart from events generated automatically with respect to each of the categories mentioned above, users can add their own events and bind them to the given levels. This functionality may facilitate cooperation between users of Knowledge Pit that decide to work in teams.

Calendars can be exported to standard exchange formats (to be imported into other calendar solutions) or can be subscribed to by RSS. This helps users to stay up to date and somehow automate their time management.

Each user has a dedicated storage space (or private resources repository) with an access and management abilities available via the web interface.

Users can communicate using the built-in chat and forum engines. They can exchange messages at various levels according to the context they are enrolled in. Each competition includes a dedicated internal chat and a forum. There are also global forums and chat rooms available to the site users, which can be enabled or disabled by the competition managers. This is yet another idea on how the Knowledge Pit platform can be useful – it may help scientists to get know new people with the same interests and stimulate their cooperation.

B. Definition and management of competitions

Data mining challenge support and maintenance are the main objectives of the site. Each contest is an individual entity in the system and it requires time and care to be well defined and created. There are two ways to achieve the goal of starting the competition. The site administrators create, run and hook a dedicated evaluator to a competition, due to theirs responsibility, knowledge and appropriate privileges, bypassing a standard acceptance procedure. The second (more common) way involves the operation of a regular user who prepares a project of the competition using the forms and tools available in Knowledge Pit. Then, a request to the site administrators is sent to revise the proposal of the contest. If it is accepted, a new hidden challenge is automatically created and the user obtains manager privileges to it. All standard modules are initialized by the default values of parameters. That includes data description pages, data files folder, news panel and submissions upload interface. Initially, a newly created challenge is hidden from other users until all necessary information is filled in and the dates are defined. After providing the necessary data (including a task description, input data files, etc.), the challenge may be published and becomes visible to anyone visiting the site.

C. A course of a competition

To each challenge there are associated three important dates:

1) a start date
2) a submissions end date
3) a contest end date

The first two dates define the actual time in which participants may compete by submitting their solutions to the challenge task. During this period the users are supposed to upload and manipulate their solution files using the available submission manager. The submitted files are automatically pre-evaluated by a dedicated service (e.g. a script or an external application that are compatible with a Knowledge Pit evaluator protocol), accordingly to the settings defined by the challenge authors. The users may upload multiple result data files among which they mark one as their target

solution that would take part in the final evaluation. The pre-evaluation is meant to generate scores that are placed on the competition's Leaderboard. The preliminary scores may serve as a rough estimation of how good the submitted solution is in comparison to results of the other users and therefore may stimulate the competitive spirit of the contest. The second period which ends the contest is the time needed to finally evaluate users' submissions (previously marked by the participants as their target solutions) and determine the winners. Each submission is scored and accordingly placed in a table called Final Leaderboard which is published within the challenge summary.

Organizers of a competition may require additional reports that describe solutions provided by the participants. In this case a competition manager may set a condition which restrains the final evaluation to the submissions with an attached report.

Each challenge, if it is already published and its visibility is not restricted by the organizers, is available to any site user. In that case, a guest has an access to the contest's general information, including a task description, a list of important dates, and the Leaderboard. Unless contest access restrictions are in effect, every user can enroll in a chosen challenge accessing all the contest's resources, including the provided data files.

D. Evaluation of the uploaded solutions

The above description refers only to the features available via the web user interface. It also presents the general flow of events starting from the contest organization, the users enrollment, the submission of solutions, their preliminary and final evaluation, ending on the publication of the challenge summary and announcement of the winners. However, not much was said about the method of defining the evaluators.

It would be very difficult to build and share a general evaluator for all possible types of data mining competitions. Usually, the stated data mining tasks are very different in many aspects, including:

- the category of performed data analysis (clustering, classification, multi-label classification, etc.)
- the solutions representation (a single file vs. multiple files placed together in one archive file) and their formats (multicolumn answers, the way of describing clusters, etc.)

Another important aspect of the data mining solution evaluation is that it often requires a lot of resources (memory, CPU time, disc I/O or database connections), e.g. when it is associated with a predictive model creation. This could result in malfunction of the competition platform due to its resources limitation. In Knowledge Pit the responsibility of the evaluation is delegated to the competition organizers. They need to provide an object database and a working evaluator. The responsibility of Knowledge Pit is limited to interaction with the object database where all the solutions are uploaded and stored. The submission scores are downloaded from it and propagated to internals of the system. The evaluator may be implemented in any suitable programming language, as a script, a stand alone compiled application or a utilization of available libraries. The only requirement is that it should maintain correct protocol of information exchange by means of changing the objects inside the database in a predefined way. The proposed flow of responsibilities frees Knowledge Pit from the things which it cannot cope with in a generic way. It also gives the organizers a very flexible method of expressing their data mining task in a form of a fully customizable evaluation procedures. For example, in AAIA'14 Data Mining Competition which is described in the following sections, MongoDB [14] was utilized as the object database and the evaluation system was implemented in the R programming language [15].

III. The Task Description

The Knowledge Pit was inaugurated with AAIA'14 Data Mining Competition which took place between February 3, 2014 and May 7, 2014. In this challenge the focus was on the feature selection problem and the data came from the public safety domain.

Our team obtained a data set containing nearly 260,000 reports from the EWID system. The reports corresponded to actions carried out by the Polish State Fire Service within the city of Warsaw and its surroundings (the Mazovia district) in years 1992 – 2011. We preprocessed a subset of this data and transformed it into a table in which each of the reports is described by nearly 12,000 attributes. Additionally, we distinguished three target attributes that correspond to information whether in the described incident there were casualties among firefighters, children or other involved people, respectively. The task in AAIA'14 Data Mining Competition was to identify attributes that can be used to robustly assign the reports to the corresponding decisions labels. We hoped that participants would come up with solutions which improve our understanding of the risk factors associated to various types of accidents.

The competition data set was provided to participants in two different formats. The first one was a traditional tabular representation of data as a comma-separated values file. Each row of this file represented a single EWID report and, in the consecutive columns, it contained values of its characteristics (the attributes). The attributes in this table could be divided into two groups. The first one contained the features extracted from the quantitative part of the report and the second group corresponded to a document-term matrix obtained from the natural language description sections. In total, the training data available to participants contained descriptions of 50,000 incident reports. Each report was characterized by 11,852 conditional attributes. All the attributes were discrete and only a few had more than two possible values. We thought about those attributes as indicators of the risk factors corresponding to the incidents.

The same data set was made available in a sparse matrix format as an EAV file [16]. In every row, the file contained exactly three integer numbers: an identifier of an object, an identifier of an attribute and the corresponding attribute value.

Since the EAV file stored exactly the same information as the traditional tabular representation of the data, this file was provided only for convenience of participants.

To each of the reports from the training data there were also assigned values of three binary decision attributes. The first decision attribute indicated incidents in which occurred a serious injury or death of one of the firefighters or members of the rescue team. The second decision attribute indicated cases for which there were children among the injured people. The third decision identified situations where any civilians were hurt. Values of those decision attributes were made available for all participants of the competition in a separate file.

It is worth noting that, by its nature, the provided data set was highly dimensional. The total number of conditional attributes corresponded to the number of distinct words in the textual part of the reports (after lemmatization), plus several hundreds of attributes from the quantitative part. Additionally, the data was sparse since only a small fraction of the attributes had a non-zero value for a particular report. On top of that, all three decision attributes were highly imbalanced – the positive classes corresponded to relatively rare events. The proportions of the positive cases for the rescuers, children and civilians were ≈ 0.004, ≈ 0.007 and ≈ 0.059, respectively. There was also a separate test data set which was used for the evaluation of submissions. It had similar characteristics to the training data but it was not available for the participants during the competition.

The competitors were asked to indicate sets of attributes that allow to accurately classify the incidents using an ensemble of Naive Bayes models [17], [18] and upload their solutions using the on-line submission system. We required that in each solution there were exactly ten attribute sets. The sets were ought to contain at least three integer numbers corresponding to indexes of attributes from the training data set. There was no upper limit for the number of attributes indicated in a single set, however, the evaluation system penalized solutions that use a large number of features.

The submitted solutions were evaluated on-line and the preliminary results were published on the competition Leaderboard. The preliminary score was computed for each submission on a random subset of the test set, which was fixed for all participants. This subset corresponded to approximately 10% of the test data. The final evaluation was performed after completion of the competition using the remaining part of the test data. Those results were also published on-line. In order to be considered for the final evaluation, each participating team had to provide a short report describing their approach.

Quality of the submissions was assessed by measuring performance of a classifier ensemble composed of Naive Bayes models. Those models were constructed using attribute sets indicated by the submitted solution, separately for each decision attribute. An output of the ensemble was computed by averaging probabilities of the positive classes returned by the individual Naive Bayes models. During the evaluation, all the training data was used for the construction of the models. The performance of a single ensemble was measured by taking Area Under the ROC Curve (AUC) [17], [18] of the probability predictions for the corresponding decision attribute and the result was averaged over all three decision attributes. Finally a penalty was applied for using a large number of conditional attributes.

In more details, if we denote by:

s	–	a submitted solution,
$\|s\|$	–	a total number of attributes used in the solution (counted with repetitions),
$AUC_i(s)$	–	Area Under the ROC Curve (AUC) of a classifier ensemble for the i-th decision attribute,

then the quality measure used for the assessment of submissions can be expressed as:

$$score(s) = F\left(\frac{1}{3}\sum_{i=1}^{3} AUC_i(s) - penalty(s)\right)$$

where the penalty is equal to:

$$penalty(s) = \left(\frac{|s| - 30}{1000}\right)^2$$

and the function F is defined as:

$$F(x) = \begin{cases} x & \text{for } x > 0 \\ 0 & otherwise \end{cases} .$$

All the data sets utilized in the competition, including the test set with the corresponding decision values, were made available after completion of the challenge at the competition web page: http://challenge.mimuw.edu.pl/contest/view.php?id=83. We are convinced that the public availability of the data will facilitate future research in this area by other members of the machine learning community.

IV. Results of the Competition

AAIA'14 Data Mining Competition attracted many skilled participants from around the world. In total there were 116 registered teams, from which 57 actively participated in the challenge by submitting at least one solution to the stated task. We received nearly 1,300 solutions and 290 of those submissions obtained a score higher than 0.94. Additionally, 46 teams provided a short report describing their approach.

The participants utilized diverse machine learning techniques in order to come up with their final attribute sets. A large share of the solutions was devised by combining the attribute filtering approach for reducing the initial feature subset with well known wrapper-based techniques. The final solutions were commonly tuned using evolutionary algorithms or the hill climbing method. The best results, however, were obtained by using algorithms optimized specifically for finding attribute sets that improve the AUC of Naive Bayes prediction models.

The wide spectrum of solutions submitted by the participants during the challenge makes it possible to perform a comprehensive study of the factors that have the biggest impact on predictions of the positive classes in the data. However, since

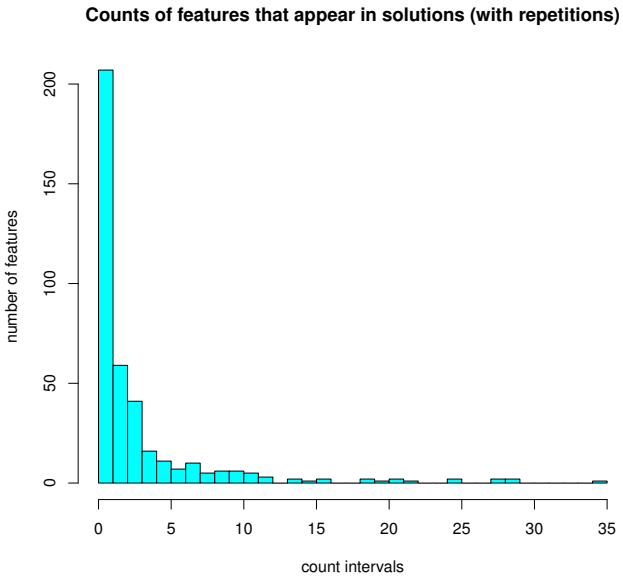


Fig. 3. Frequencies of individual attributes in the final solutions submitted by the twenty top-scored participants of AAIA'14 Data Mining Competition. The attributes were counted with every repetition. The most frequent attribute was present 35 times in the considered solutions. It corresponded to the term "during".

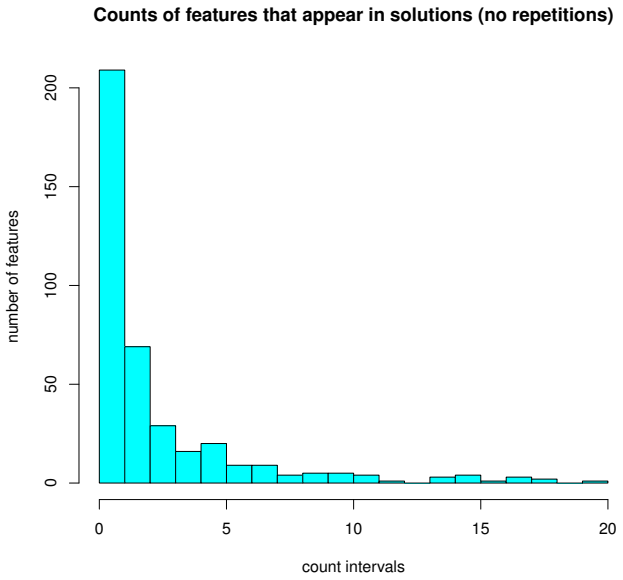


Fig. 4. Frequencies of individual attributes in the final solutions submitted by the twenty top-scored participants of AAIA'14 Data Mining Competition. The attributes were counted without repetitions (an attribute is counted once for every solution in which it appears). The most frequent attribute was present in all of the considered solutions. It corresponded to the term "during".

the considered phenomena are complex and diverse, and there was a short period between the submissions of the solutions and the preparation of this paper, the analysis described here is limited to a set of hypotheses. Those hypotheses try to explain some of the observed regularities, however, we need to stress that most of them require further investigation to be supported with strong statistical evidences. Therefore, all the explanations which we present in the following sections are just a starting point for a more detailed analysis leading to a better understanding of dangerous situations and the related threats faced by firefighters.

A. *Analysis of frequent features*

We start our analysis with reviewing the features which most commonly appeared in the top-scored solutions. Each of the features in question corresponds to a specific attribute from the attribute section, a word or a word-set (see Section III) from the descriptive section of the reports.

We analyzed the attributes indicated by the final solutions submitted by the twenty participants who obtained the highest ranks on the Final Leaderboard. In total (counted with repetitions) they constituted a set of 1,538 attributes. The number of different attributes in this set is 413. In most of the cases an attribute was used only in a single solution. However, there is a small subset of features that are more frequently used than the others. We focused our investigation on these attributes and verified their relation with the occurrence of serious injuries. Figures 3 and 4 depict the frequency of individual attributes, computed by considering every occurrence of attributes and by counting the solutions in which an attribute is present, respectively. In the second case, an attribute is counted only once for every solution in which it appears, regardless of the number of its appearances in that solution.

We analyzed the most frequent attributes with regard to their relation to the fire safety domain. We arranged them in several groups. Each of those groups defines a different type of relation to the cases of serious injuries or deaths in incidents.

The first group consists of features that correspond to nouns related to the injured parties. Incident Commanders who report the casualties tend to use a set of specific terms for describing the victims. Those terms are not used to refer to other people present at the emergency scene. In particular, examples of such nouns are: "a boy", "a kid", "a girl", "male", "female" or "a private". The usage of these words in a report indicates that something happened to the described person. There is also some interesting regularity in this group: the term "boy" appears nearly three times more often than the word "girl". The other interesting thing observed among the attributes in this group is that the terms referring to the common participants of the interventions change in situations when they get injured. For example, a firefighter who performs his/her activities according to the schedule, is usually anonymous (e.g. "two firefighters set the ladder"). However if something wrong happens, a full rank, the name and surname are reported in the descriptive section of the report. In the most of cases the injuries concern lower-ranked firefighters. Therefore, attributes represented by terms such as "str" (in Polish it is an abbreviation for "a private") were often selected as good predictive features.

TABLE I
THE FEATURE SETS SUBMITTED BY THE WINNERS OF AAIA'14 DATA MINING COMPETITION.

#Sub	First place	Second place	Third place
1	action_people_evacuation, to_sting, longitude, thought, ignition, action_people_release, losses_overall_dollars, hospital_name, private	local_threat_type_road_transport, kid, knee-joint, coal, head, to_transport, observation, bite	palm, during, head, action_people_evacuation, icicle, bar, man
2	road, observation, hand, man, hit, transport, rinse	private, action_people_evacuation, to_drive_away, passenger, warm, carbon_monoxide_poisoning, first_fire_engine_drove_km	thumb, withdraw, blast_off, losses_overall_dollars, grating, boy, observation
3	year, right, to_transport, leg, kid, during, used_equipment_chainsaw, to_go_somewhere, light, hospital_name, daughter, ankle	resources_fire_rescue_unit, person, ankle, condition, foot, action_used_extinguishing_aid_attack, hospital_name	girl, to_bite, carbon_monoxide_poisoning, to_lead, to_sting, action_people_release, to_hit_a_pedestrian, local_treat_medium
4	action_door_opening, decease, compartment, allotment garden, to_transport	light, used_equipment_chainsaw, to_faint, carbon_monoxide, to_sting, corpse, apply	kid, resources_fire_rescue_unit, hospital_name, driver, volume_of_incident_scene, finger, daughter
5	carbon_monoxide, oxygen_therapy, local_threat_cause_gas_device_fault, local_threat_medium, extinguishing_in_attack, burn, victim	kid_or_kids, eye, minibar, street_name, officer_name, latitude	work, apparatus, drive, hospital_name, deceased, light, action, delay
6	knee, head, rescuers_fire_rescue_unit_people, district	suffer, leg, grating, deceased, knee, extricate, palm	kid, make_of_a_car, injury, hospital_name, extract, hand, morning
7	action_smoke_extraction, palm, latitude, local_threat_type_road_transport, person, connector_or_a_switch, hospital_name, state_property, deceased	truck, to_set_fire, star_or_a_whistle, to_hit_or_run_over_someone, longitude, extract, compartment	truck_type, ankle, to_slip, private, hospital_name, to_swell, to_bite, corpse
8	resources_police_cars, team, technical, kid, water, to_drive_a_driver_or_a_steering_wheel, to_lead, cause_of_a_local_threat_act_of_terror, to_do_or_break	action_people_release, to_hit_or_crash, agricultural, immediately, department_branch_or_division, withdraw, boy	year, knee, passenger, suffer, personal_details, foot, firefighter
9	explosion_any_type, delay, face, homeless, grating, forearm	during, hand, to_fall_asleep, local_threat_medium, functioning, explosion_any_type, oxygen_therapy	leg, to_force, to_hit_or_crash, oxygen_therapy, coal, to_hit_or_run_over_someone, carbon_monoxide, to_wash, technical
10	corpse, girl, to_hit_or_run_over_someone, suffer, boy, burn_down	darkness, to_hit_or_knock_someone_off, ankle, to_twist, man, action_inside_chimneys, delay	team, local_threat_cause_careless_driving, orthopedic, face, to_carry_or_transport, mean_of_transport, compartment, person

The second group is related to descriptions of the injuries and mostly consists of names of human body parts. In this group, the most commonly used words are: "leg", "palm", "hand", "side", "body", "foot", "twinkle", etc.

The next group represents the attributes that describe activities undertaken by firefighters when they faced an injured person. This group consists of attributes from the attribute section such as: "action_people_evacuation", "localizing_people", "oxygen_therapy", etc. or words (mostly verbs) from the descriptive section such as: "to transfer" (to an ambulance), "to transport", "observation", "to cut", "to open", etc.

Another group represents features related to terms which describe a cause of the injuries or fatalities. In this group we find the following words: "intoxication", "to hit" (a pedestrian), "to get" (a stroke), "sprain", "to twist" (an ankle), "to slip", "bite", "bump", etc.

All the groups described above, are examples of attributes which were used by ICs in order to address matters related to an injury or death. They can be useful for a post-incident analysis of the causes, since the identification of such key phrases may boost performance of information retrieval systems that work on EWID data. However, those terms alone do not reveal any specific risk factors related to the fire safety domain. The knowledge resulting from the identification of those attributes does not have a direct impact on the safety of firefighters and incident victims. Moreover, it does not reveal interesting aspects of the rescue actions, apart from the words or phrases which are used in the reports in order to describe the casualties or fatalities.

There is, however, one group of features which are likely to correspond to important risk factors. By obtaining information regarding those factors during a real-life F&R action we may potentially improve the safety of involved people. This group consists of attributes corresponding to terms such as: "carbon monoxide", "darkness", "single-family terraced buildings", "mart", "electrocution", "bite off" and some specific geographical coordinates. A further analysis of a role and a context of these attributes in the reports may shed light on the factors that affect the possibility of serious incidents. Nevertheless, a thorough investigation is required in order to explain their role in the generation of the unwanted events.

B. Analysis of frequent attribute sets

Due to the fact that usefulness of knowledge obtained by the analysis of individual attributes was limited, we performed an additional investigation of frequent attribute sets. In this analysis we distinguish global and local sets of attributes.

As the first type we consider the whole attribute sets that correspond to individual models in the submitted solutions. The second type refers to subsets of attributes that commonly co-appear in the top-scored solutions.

In our first attempt, we analyzed the global feature sets, i.e., those which turned out to have the best predictive abilities for the whole data. These global feature sets were submitted by the winners of the competition. Table I gives the names of attributes from the sets submitted by the three best teams.

In those groups we indicated a few interesting types of sets that should undergo a further analysis. The first one can be summarized by terms or quantitative attributes such as: "activities_opening_doors", "decease", "compartment", "allotment", "garden", "to transport". Features from this set were often present in reports describing incidents caused by homeless or youngsters who illegally occupy cottage-gardens. This may indicate that there is a considerably large fraction of fatalities resulting from fires started in such conditions.

Another interesting type of attribute sets contains terms such as: "carbon monoxide", "oxygen therapy", "caused_by_heating_device_fault", "extinguishing", "fatality". This group may indicate that a large number of deaths is caused by carbon monoxide poisonings or fires started as a result of malfunctioning heating devices.

The next of the interesting feature set types can be characterized by terms: "explosion", "corps", "face", "homeless", "rate" and "forearm". It seems that there is a considerable number of incidents that involve homeless and some explosions. This set is very difficult to explain without a deeper analysis of the reports describing specific incidents.

A different attribute set type can be represented by the terms: "light", "used_equipment_chainsaw", "wood", "wane", "body", "to sting" and "girl". Combination of those terms often indicates a subset of incidents related to light injuries caused by an inappropriate usage of sharp tools, such as a chainsaw.

There is also a type of attribute sets which may be related to the incidents that happens after a nightfall, in a situation when somebody or something fell into a hole or a chimney. This set is represented by the attributes: "nightfall", "man", "cat", "twist", "shorten", "hit", "action_inside_chimneys".

The last of the identified types of interesting attribute sets is once again related to the problem of carbon monoxide poisonings. However, if the terms "oxygen therapy" or "carbon monoxide" appear along the terms such as "coal", "technical" or "functioning" it may indicate that the problem of poisoning is often related to malfunctioning coal furnaces.

C. Analysis of the local attribute sets

The analysis of the attribute sets submitted by the winning teams was an attempt to identify the most significant factors that have an impact on the occurrence of the cases from the positive decision classes. However, due to a large diversity of interventions of Fire Services – ranging from fires, through road traffic accidents, to natural disasters – finding the globally most affecting features is a very complex task. Therefore, we need to face a problem of finding attribute sets which have an impact on subclasses of incidents such as fires in residential buildings [3]. To accomplish this challenging task we applied a frequent item set mining technique, i.e. the *Apriori* algorithm [19].

We computed frequent attribute sets from the solutions submitted by the twenty top-scored participants (i.e. every line in the solution files was treated as a transaction) and we ranked them according to their support. The utilization of Apriori resulted in finding millions of frequent attribute sets. Due to our limited human processing abilities, we reduced the number of the sets for the analysis to the top 351 with the highest value of the support. Below we present a few examples of the interesting attribute sets which were revealed by this analysis.

As in the previous analysis, the most commonly appearing attribute sets are related to the expressions used by IC in order to report the injuries or fatalities. However, as in the previous cases, we were able to identify a few interesting attribute chunks. All of them should be further analyzed by experts from State Fire Service. Examples of such sets include: "used_equipment_chainsaw" and "light" – it indicates that there is a group of incidents related to an unfortunate use of a chainsaw by fire-fighters. Even though it seems reasonable that in the most of such cases the inflicted injuries are superficial, those results indicate that a proper handling of this type of tools should be better stressed during the firefighter training.

Another group consists of terms: "firefighter" and "sprain". It may indicate that there is a significant number of limb injuries during the rescue actions. The next of the interesting attribute sets is composed of terms: "firefighter", "releasing people" and "bite" which may indicate that there is a number of cases where firefighters are bitten by animals during a rescue activities. The last example of a common attribute set is "to slip" and "hand". It may be considered similar to the group of attributes related to limb injuries. It requires a further investigation in order to be associated with a specific type of firefighter actions.

D. Attribute cluster analysis

After the investigation of attributes and attribute subsets that frequently appear in the best solutions, we decided to check whether there is any redundancy among them. We were also interested in finding pairs of attributes that can be regarded semantically similar in the context of the fire safety. Successful identification of such pairs or groups would be beneficial for the further analysis of the EWID data. It would also be very useful for the risk assessment purposes, in situations when a part of information about an incident is unavailable or unreliable.

In order to find groups of closely related attributes we performed an attribute cluster analysis [20]. Intuitively, two attributes can be considered similar if they often co-occur in the solutions with the same groups of other attributes. However, if a pair of attributes commonly appears in the same sets submitted by the highest scored participants, it means that those features are complement in some way and they should

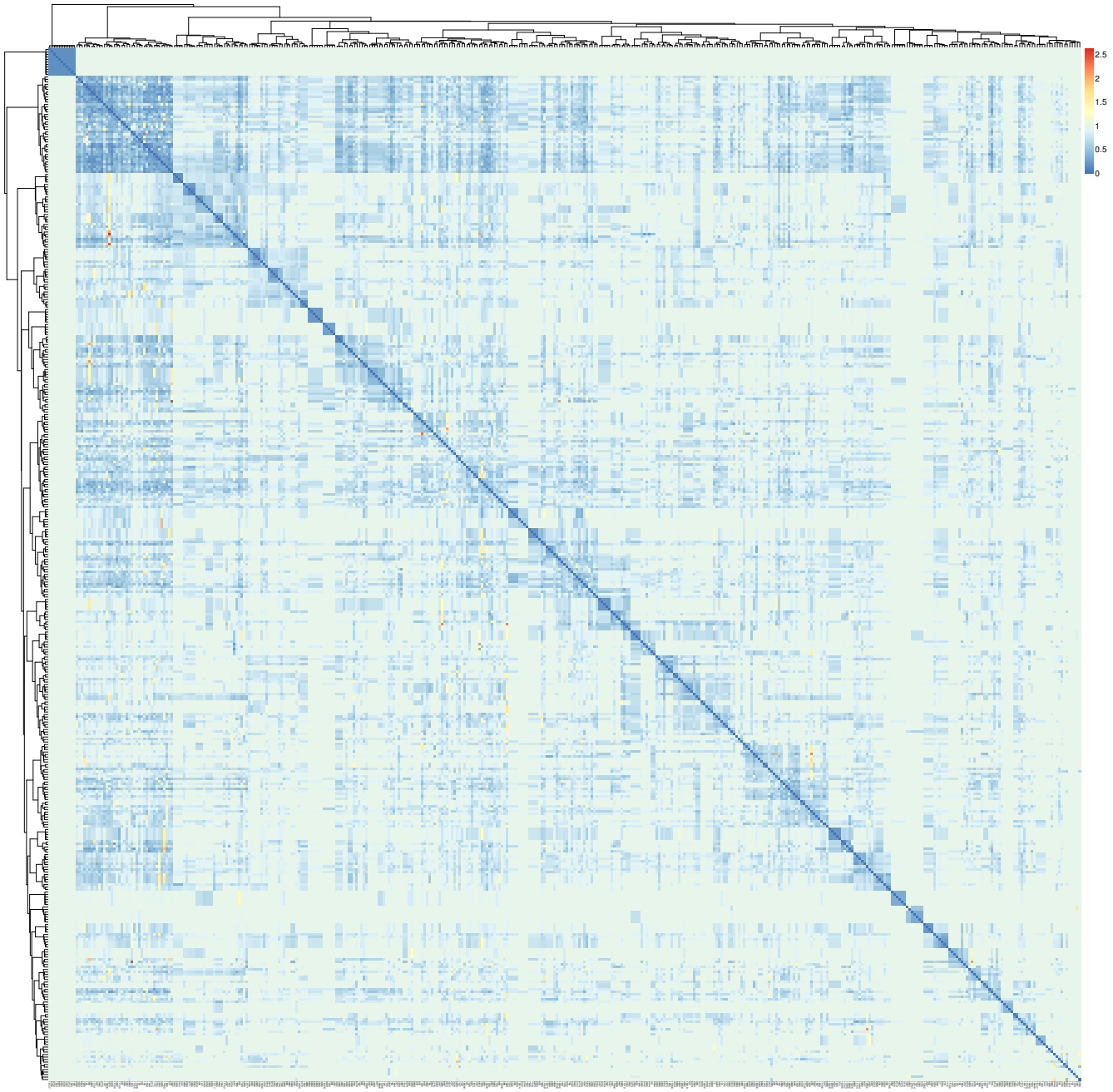


Fig. 5. A co-occurrence-based heat map of the attributes appearing in the top 20 solutions. Rows and columns of the matrix correspond to the attributes and the color of the spots symbolizes their dissimilarity. Additionally, on the top and the left side of the plot, a dendrogram of a hierarchical clustering of the selected attributes is given. The darker squares along the diagonal correspond to clusters of closely related (i.e., potentially exchangeable) attributes.

not be regarded similar. For this reason, we computed the dissimilarity between each attribute pair twofold.

First, we constructed a co-occurrence matrix whose rows and columns correspond to the 413 attributes from the solutions. For every attribute (a row of the matrix), we iterated over the attribute sets in which it was present and increased the matrix entries in the columns corresponding to the co-occurring features, by the inverse of the attribute set cardinality. In this way we constructed a new representation of the attributes. In the second step, we created a dissimilarity matrix taking the values from the co-occurrence matrix and subtracting from them the corresponding values of cosines between the new attribute representations.

Using the dissimilarity matrix we were able to perform the attribute cluster analysis. We did it using the hierarchical agglomerative approach with the Ward's linkage function [21]. The clustering results are depicted by the heat map in Figure 5. The darker spots correspond to pairs of similar

attributes. They are likely to be exchangeable in the context of classification and thus can be interpreted as semantically related. In the plot, potential clusters of attributes are represented by dark squares aligned along the diagonal of the matrix. For example, in the first cluster there were attributes such as "type_of_building_standalone_compartment", "one_story_high", "single_family_houses" and "action_inside_buildings_at_ground_floor".

In the future we plan to extend this analysis by considering decision rules which may be constructed from the frequent attribute sets. Such rules may compose a useful tool for supporting ICs at a fire ground, which is the main task of our ICRA project [2].

V. Summation

In this paper we focused on introducing a web platform, called Knowledge Pit, created in order to support organization of data mining competitions. On the one hand, this platform is appealing to members of the machine learning community for whom competitive challenges can be a source of new interesting research topics. Solving real-life complex problems can also be an attractive addition to academic courses for students who are interested in practical data mining. On the other hand, setting up a publicly available competition can be seen as a form of outsourcing the task to the community. This can be highly beneficial to the organizers who define the challenge, since it is an inexpensive way to solve the problem which they are investigating. Moreover, an open data mining competition can become a bridge between domain experts and data analysts. In a longer perspective, it may leverage a cooperation between industry and academic researchers.

We also described *AAIA'14 Data Mining Competition: Key risk factors for Polish State Fire Service* which was the first analytic challenge organized at Knowledge Pit. We presented the scope of this competition and briefly summarized its results. In addition, we discussed the results of our initial analysis of the best of the submitted solutions, highlighting their potential practical applications.

The conducted analysis is by no means complete. In future, the results of the competition will be thoroughly investigated by a team composed of experienced Incident Commanders and data mining experts. We hope that the results of this research, conducted as a part of a larger project ICRA [2], will have a noticeable impact on the fire safety domain. We also hope that our competition will revive a discussion on this topic among researchers with different backgrounds and expertise.

Acknowledgments

This work was partly supported by Polish National Science Centre (NCN) grants DEC-2011/01/B/ST6/03867 and DEC-2012/05/B/ST6/03215 and by National Centre for Research and Development (NCBiR) grant No. O ROB/0010/03/001 in the frame of Defence and Security Programmes and Projects: "Modern engineering tools for decision support for commanders of the State Fire Service of Poland during Fire&Rescue operations in the buildings".

References

[1] H. Johansson, *Decision Making in Fire Risk Management*. Dept. of Fire Safety Engineering, Lund University, 2001.

[2] A. Krasuski, A. Jankowski, A. Skowron, and D. Ślęzak, "From sensory data to decision making: A perspective on supporting a fire commander," *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, vol. 3, pp. 229–236, 2013.

[3] A. Krasuski and A. Janusz, "Semantic tagging of heterogeneous data: Labeling fire & rescue incidents with threats," in *FedCSIS*, 2013, pp. 77–82.

[4] K. Bąk, A. Krasuski, and M. Szczuka, "Searching for Concepts in Natural Language Part of Fire Service Reports," in *Concurrency Specificaton and Programming*, 2013.

[5] B. Gilbert, D. Nichols, B. Aisbett, M. Phillips, M. Sargeant *et al.*, "Fighting with fire: how bushfire suppression can impact on fire fighters' health," *Australian family physician*, vol. 36, no. 12, p. 994, 2007.

[6] M. Zaksek and J. L. Arvai, "Toward improved communication about wildland fire: mental models research to identify information needs for natural resource management," *Risk analysis*, vol. 24, no. 6, pp. 1503–1514, 2004.

[7] M. Wojnarski, A. Janusz, H. S. Nguyen, J. Bazan, C. Luo, Z. Chen, F. Hu, G. Wang, L. Guan, H. Luo, J. Gao, Y. Shen, V. Nikulin, T.-H. Huang, G. J. McLachlan, M. Bošnjak, and D. Gamberger, "RSCTC'2010 discovery challenge: Mining DNA microarray data for medical diagnosis and treatment," in *Proceedings of RSCTC'2010*, ser. LNAI, M. S. Szczuka et al., Ed., vol. 6086. Heidelberg: Springer, 2010, pp. 4–19.

[8] A. Janusz, H. S. Nguyen, D. Ślęzak, S. Stawicki, and A. Krasuski, "JRS'2012 Data Mining Competition: Topical Classification of Biomedical Research Papers," in *Proceedings of RSCTC'12*, ser. LNAI, J.T. Yao et al., Ed., vol. 7413. Springer, Heidelberg, 2012, pp. 417–426.

[9] M. Wojnarski, S. Stawicki, and P. Wojnarowski, "TunedIT.org: System for automated evaluation of algorithms in repeatable experiments," in *Proceedings of RSCTC'2010*, ser. LNAI, vol. 6086. Springer, 2010, pp. 20–29.

[10] J. Cole, *Using Moodle*, 1st ed. O'Reilly, 2005.

[11] Lee and B. Ware, *Open Source Development with LAMP: Using Linux, Apache, MySQL and PHP*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2002.

[12] E. Rosebrock and E. Filson, *Setting Up LAMP: Getting Linux, Apache, MySQL, and PHP Working Together*. Alameda, CA, USA: SYBEX Inc., 2004.

[13] P. C. Isaacson, "Building a simple website using open source software (gnu/linux, apache, mysql, and python)," *J. Comput. Sci. Coll.*, vol. 19, no. 1, pp. 286–288, Oct. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=948737.948777

[14] E. Plugge, T. Hawkins, and P. Membrey, *The Definitive Guide to MongoDB: The NoSQL Database for Cloud and Desktop Computing*, 1st ed. Berkely, CA, USA: Apress, 2010.

[15] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013. [Online]. Available: http://www.R-project.org/

[16] J. Wróblewski and S. Stawicki, "Sql-based kdd with infobright's rdbms: Attributes, reducts, trees," in *RSEISP*, ser. LNCS, M. Kryszkiewicz, C. Cornelis, D. Ciucci, J. Medina-Moreno, H. Motoda, and Z. W. Raś, Eds., vol. 8537. Springer, 2014, pp. 28–41.

[17] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[18] T. M. Mitchell, *Machine Learning*, ser. McGraw Hill series in computer science. McGraw-Hill, 1997.

[19] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. I. Verkamo *et al.*, "Fast discovery of association rules." *Advances in knowledge discovery and data mining*, vol. 12, no. 1, pp. 307–328, 1996.

[20] A. Janusz and D. Ślęzak, "Rough set methods for attribute clustering and selection," *Applied Artificial Intelligence*, vol. 28, no. 3, pp. 220–242, march 2014.

[21] L. Kaufman, P. Rousseeuw, and E. Corporation, *Finding Groups in Data: an Introduction to Cluster Analysis*. Wiley Online Library, 1990, vol. 39.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 355–359
DOI: 10.15439/2014F496
ACSIS, Vol. 2

Parsimonious Naive Bayes

Marc Boullé
Orange Labs,
2 avenue Pierre Marzin, 22300 Lannion, France
http://perso.rd.francetelecom.fr/boulle
Email: marc.boulle@orange.com

***Abstract*—We describe our submission to the AAIA'14 Data Mining Competition, where the objective was to reach good predictive performance on text mining classification problems while using a small number of variables. Our submission was ranked 6^{th}, less than 1% behind the winner. We also present an empirical study on the trade-off between parsimony of the representation and accuracy, and show how good performance can be obtained quickly and efficiently.**

I. Introduction

THE AAIA'14 Data Mining Competition[1] is related to a problem of text classification. A corpus of 50,000 texts coming from reports of the Polish State Fire Service is provided, with a representation consisting of 11,852 variables (mainly based on documents words). The objective is to classify the texts into three binary classes by the mean of an ensemble of ten Naive Bayes classifiers, using as few variables as possible. In this paper, we present the method we used at the AAIA'14 Data Mining Competition. It mainly exploits the results of a Selective Naive Bayes classifier (summarized in Section II), trained for each of the three challenge class variables. The best subset selections of variables are collected and grouped together to form our submission to the challenge (see Section III).

Interestingly, the challenge evaluation criterion combines the test performance and the number of selected variables in a single formula. Whereas feature selection [1] aims at improving interpretability, test accuracy and deployment time, most papers in the literature focus on test performance only. Still, a small number of selected variables is often a requirement in practical data mining studies. In Section III, the trade-off between test performance and number of selected variables is investigated, using the challenge dataset as a case study. Finally, Section IV summarizes the paper.

II. Selective Naive Bayes Classifier

We summarize the Selective Naive Bayes (SNB) classifier introduced in [2]. It extends the Naive Bayes classifier owing to an optimal estimation of the class conditional probabilities, a Bayesian variable selection and a Compression-based Model Averaging.

[1]http://challenge.mimuw.edu.pl/mod/page/view.php?id=565

A. Optimal discretization

The Naive Bayes (NB) classifier has proved to be very effective in many real data applications [3], [4]. It is based on the assumption that the variables are independent within each class, and solely relies on the estimation of univariate conditional probabilities. The evaluation of these probabilities for numerical variables has already been discussed in the literature [5], [6]. Experiments demonstrate that even a simple equal width discretization brings superior performance compared to the assumption using a Gaussian distribution per class. In the MODL approach [7], the discretization is turned into a model selection problem and solved in a Bayesian way. First, a space of discretization models is defined. The parameters of a specific discretization are the number of intervals, the bounds of the intervals and the class frequencies in each interval. Then, a prior distribution is proposed on this model space. This prior exploits the hierarchy of the parameters: the number of intervals is first chosen, then the bounds of the intervals and finally the class frequencies. The choice is uniform at each stage of the hierarchy. Finally, the multinomial distributions of the class values in each interval are assumed to be independent from each other. A Bayesian approach is applied to select the best discretization model, which is found by maximizing the probability $p(Model|Data)$ of the model given the data. Owing to the definition of the model space and its prior distribution, the Bayes formula is applicable to derive an exact analytical criterion to evaluate the posterior probability of a discretization model. Efficient search heuristics allow to find the most probable discretization given the data sample. Extensive comparative experiments report high performance.

The case of categorical variables is treated with the same approach in [8], using a family of conditional density estimators which partition the input values into groups of values.

B. Bayesian Approach for Variable Selection

The naive independence assumption can harm the performance when violated. In order to better deal with highly correlated variables, the Selective Naive Bayes approach [9] exploits a wrapper approach [10] to select the subset of variables which optimizes the classification accuracy. Although the Selective Naive Bayes approach performs quite well on datasets with a reasonable number of variables, it does not scale on very large datasets with hundreds of thousands of instances and thousands of variables, such as in marketing applications or text mining. The problem comes both from the

978-83-60810-58-3/$25.00

search algorithm, whose complexity is quadratic in the number of variables, and from the selection process which is prone to overfitting. In [2], the overfitting problem is tackled by relying on a Bayesian approach, where the best model is found by maximizing the probability of the model given the data. The parameters of a variable selection model are the number of selected variables and the subset of variables. A hierarchic prior is considered, by first choosing the number of selected variables and second choosing the subset of selected variables. The conditional likelihood of the models exploits the Naive Bayes assumption, which directly provides the conditional probability of each label. This allows an exact calculation of the posterior probability of the models. Efficient search heuristic with super-linear computation time are proposed, on the basis of greedy forward addition and backward elimination of variables. The classifier resulting from the best subset of variables is the MAP (maximum a posteriori) Naive Bayes, which we call MNB in the rest of the paper.

C. Compression-Based Model Averaging

Model averaging has been successfully exploited in bagging [11] using multiple classifiers trained from re-sampled datasets. In this approach, the averaged classifier uses a voting rule to classify new instances. Unlike this approach, where each classifier has the same weight, the Bayesian Model Averaging (BMA) approach [12] weights the classifiers according to their posterior probability. In the case of the Selective Naive Bayes classifier, an inspection of the optimized models reveals that their posterior distribution is so sharply peaked that averaging them according to the BMA approach almost reduces to the MAP model. In this situation, averaging is useless. In order to find a trade-off between equal weights as in bagging and extremely unbalanced weights as in the BMA approach, a logarithmic smoothing of the posterior distribution, called Compression-based Model Averaging (CMA), is introduced in [2]. The weighting scheme on the models reduces to a weighting scheme on the variables, and finally results in a single Naive Bayes classifier with weights per variable. Extensive experiments demonstrate that the resulting Compression-based Model Averaging scheme clearly outperforms the Bayesian Model Averaging scheme. In the rest of the paper, the classifier resulting from model averaging is called Selective Naive Bayes (SNB).

D. Training Time Complexity

The algorithm consists in three phase: data preprocessing using discretization or value grouping, variable selection and model averaging. The preprocessing phase is super-linear in time and requires $O(KN \log N)$ time, where K is the number of variables and N the number of instances. In the variable selection algorithm, the method alternates fast forward and backward variable selection steps based on randomized reorderings of the variables, and repeats the process several times in order to better explore the search space and reduce the variance caused by the dependence over the order of the variables. The number of repeats is fixed to $\log N + \log K$, so that the overall time complexity of this phase is $O(KN(\log K + \log N))$, which is comparable to that of the preprocessing phase. The model averaging algorithm consists in collecting all the models evaluated in the variable selection phase and averaging then according to a logarithmic smoothing of their posterior probability, with no overhead on the time complexity. Overall, the train algorithm has an $O(KN(\log K + \log N))$ time complexity and $O(KN)$ space complexity.

III. Challenge Submission

A. Preliminary experiments

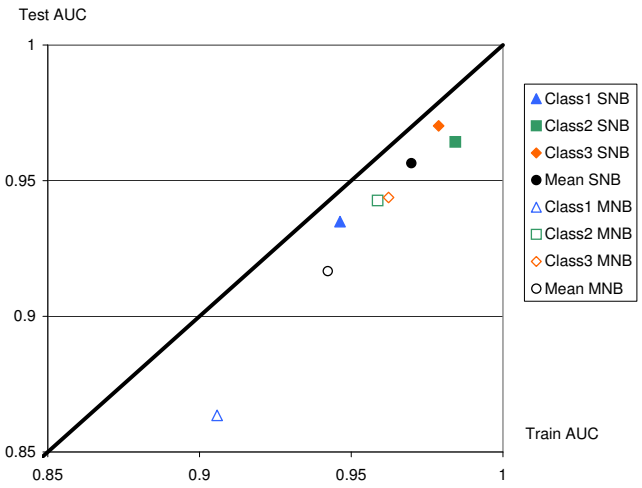


Fig. 1. Train versus test AUC for the MNB and SNB classifiers

The SNB classifier outputs several indicators for each variable:

- Level: evaluation of the predictive importance of the variable taken individually, based a normalized estimation of class conditional entropy. The level is between 0 (variable without predictive interest) and 1 (variable with optimal predictive importance).
- MAP: indicates that the variable belongs to best subset of variables, related to the MNB classifier.
- Weight: weight of the variable in the SNB classifier that exploits the model averaging method summarized in Section II.

We consider the standard NB classifier that exploits all the predictive variables; the non-informative variables with Level 0 are discarded. We also consider the 1NB classifier, which uses only one variable, the one with the highest Level. The MNB classifier is based on a subset of variables with fewer redundancy problems than the NB classifier. The SNB classifier exploits the same variables as the NB classifier, with weights per variable: it cannot be considered as a true Naive Bayes classifier.

As the challenge requires few variables, we focus on the MNB classifier which is very parsimonious compared to the SNB classifier, although it is both less accurate and less robust.

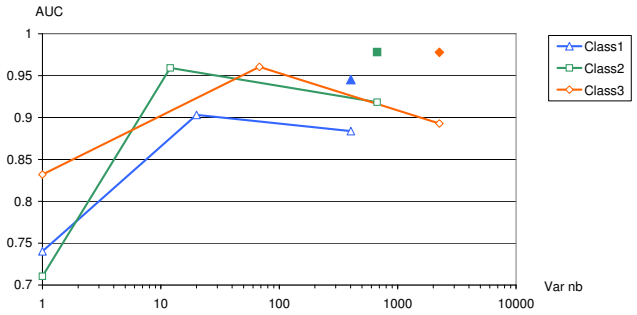


Fig. 2. AUC versus number of variables for the 1NB, MNB, NB and SNB classifiers

We trained the classifier [2] on a $70\% - 30\%$ split of the challenge dataset to evaluate its performance and robustness. We obtained a mean 96.9 train AUC and 95.6 test AUC for the SNB classifier, 94.0 train AUC and 91.6 test AUC for the MNB classifier, which confirms the usual behavior of both classifiers. This is illustrated in Figure 1, where the mean train and test AUC are presented as well as the detailed AUC per class.

We then trained the classifier using all the available data in order to reach better accuracy and robustness, with an expected decrease in accuracy on the challenge hidden dataset of about 1% for the SNB and 2% the MNB. In Figure 2, we report the AUC versus the number of variables for the 1NB, MNB, NB (empty shapes on the curves) and SNB (plain shapes) classifiers. As a reminder, the default AUC (with no variables) is 0.5. The results show that using one single variable, the 1NB classifier gets very good AUC, between 0.70 and 0.85. The MNB obtains 0.903 AUC with 20 variables for the first class, 0.959 AUC with 12 variables for the second class, 0.961 AUC with 68 variables for the third class. The NB classifier that keeps between 400 and 2000 informative variables out of the 11,852 input variables suffers from redundancy between the variables, which is harmful w.r.t. the independence assumption. All together, the NB uses far more variables than the MNB and gets a lower AUC. As expected, the SNB classifier obtains the higher accuracy using variable weights.

B. First submission

To get familiar with the challenge evaluation protocol, we grouped together the MAP variables of our three MNB classifiers (without removing the duplicates) and obtained a set of 100 variables. As the variable number is greater than that of each MNB, a better accuracy and robustness can be expected. The challenge rules states that the variable set is evaluated using and ensemble of ten Naive Bayes classifiers, each consisting of at least three variables. We chose to partition our set of 100 variables into ten random subsets of equal size, with the hope that the resulting ensemble classifier would behave similarly to a single Naive Bayes with 100 variables. This first submission was settled within a few hours after the download of the challenge data and got a score of 0.9468 on the leaderboard, second behind the leader (0.9476) on 2014-04-18.

[2] Available as a shareware at www.khiops.com

C. Additional experiments

As this first result was promising, we decided to proceed with further optimizations. The challenge evaluation criterion (score) is a mean AUC minus a penalty. The mean AUC should be above 0.5 (default performance). The penalty in the challenge is quadratic w.r.t the number of selected variables $|s|$ according to

$$penalty(s) = (\frac{|s| - 30}{1000})^2.$$

It is 0 with 30 variables and reaches 0.5 with 737 variables. With 100 variables, we got a penalty of 0.005 and therefore a leaderboard AUC of about 0.952, which is in line with our expectation. There might be room for some improvement, by optimizing directly the challenge evaluation criterion.

Algorithm 1 Bachward variable selection

Require: $X = (X_1, X_2, \dots X_K)$ {Set of input variables}
Ensure: S_{Best} {Best subset of variables}
$S = X, S_{Best} = X$ {Start with all the input variables}
{Backward selection}
while $|S| > 30$ **do**
  {Select best variable to remove}
  for $X_k \in S$ **do**
    if $(score(S - \{X_k\}) < score(S))$ **then**
      $X_{Remove} = X_k$
    end if
  end for
  {Update selection}
  $S = S - \{X_{Remove}\}$
  if $(score(S) < score(S_{Best}))$ **then**
    $S_{Best} = S$
  end if
end while

We then started from our subsets of MAP variables for each of the three classes, augmented with MAP variables resulting from the training of the three classes simultaneously ($AllClass = Concat(Class1, Class2, Class3)$). We obtained a starting set of 103 distinct variables. We then used a standard variable backward elimination algorithm based on a direct optimization of the challenge criterion on all the dataset. This variable selection method is summarized in Algorithm 1. At each step, it evaluates each variable elimination, then removes the variable that brings the best score. The algorithm returns the best subset of variables found during optimization.

In Figure 3, we report the AUC per class and the mean AUC obtained along the optimization path, from 103 variables down to 30 variables. The very few first optimization steps eliminate redundant variables, and improve the mean AUC from 0.953 to 0.957 with 95 variables. We then have a long plateau until getting 65 variables, and finally a slow decrease in

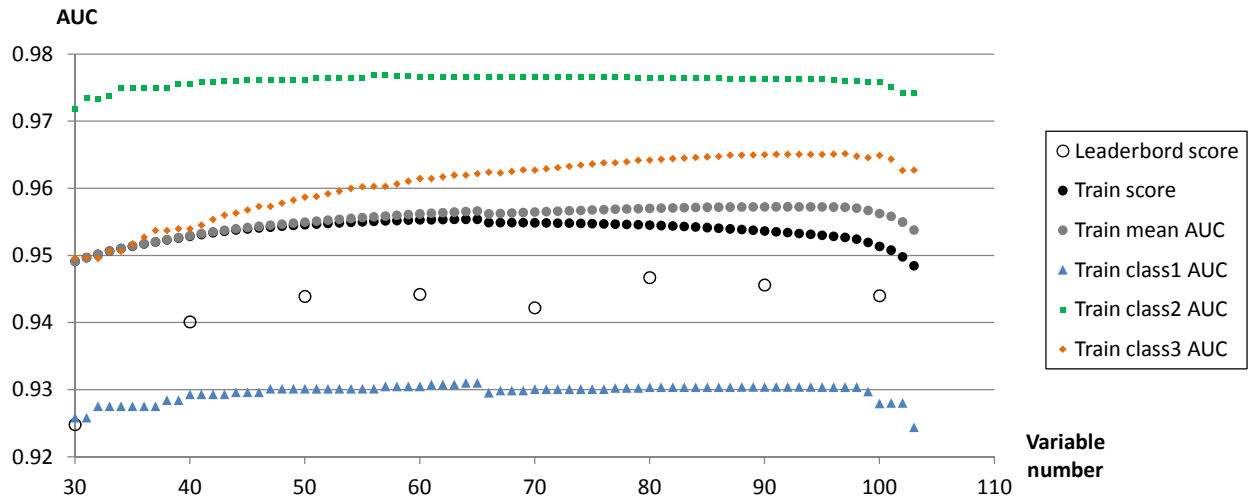


Fig. 3. Sensibility analysis: AUC versus number of selected variables

AUC, with a final AUC of 0.949 with 30 variables. The shape of the AUC curves is similar to that of Figure 2, which shows three AUC results with selections from one single variable to thousands of variables. The AUC increases quickly with very few variables (in Figure 2, one single variable is sufficient to goes from an AUC of 0.5 to around 0.75). Then, after a few tens of variables, a plateau is reached, and finally, for large numbers of variables, the performance decreases significantly down to the performance of the NB classifier (0.89 on average in Figure 2). In this bi-criterion problem, the beginning of the curve presented in Figure 3 can be interpreted as a Pareto curve, where each point corresponds to an optimal AUC given a max number of selected variables. In real data mining projects, this kind of curve might be helpful to find the best trade-off between accuracy and number of selected variables, according to the requirement and constraints of the project.

In the challenge, the score includes a penalty to choose the best trade-off. The challenge score is reported with black circles in Figure 3. The best score shown in Figure 3 gets a 0.5% improvement (up to a train score of 0.955 with 65 variables). This improvement is rather small and might be prone to overfitting, with an expected increased variance as the number of variables decreases. We got a score of 0.9452 on the leaderboard with this optimized solution. We tried other random partitions of the same variables into ten subsets (for the ten Naive Bayes ensemble classifier) and obtained score variations of about 0.3%. We also submitted a series of variable sets of increasing size along our optimization path, from 30 variables up to 100 variables by steps of ten. The resulting leaderboard scores (reported using white circles in Figure 3) shows that the improvements obtained during the train optimization vanish withing the variance of the results on the challenge leaderboard dataset. Furthermore, within a same set of selected variables, different random partitions in ten subsets produce a variance of the results of similar magnitude.

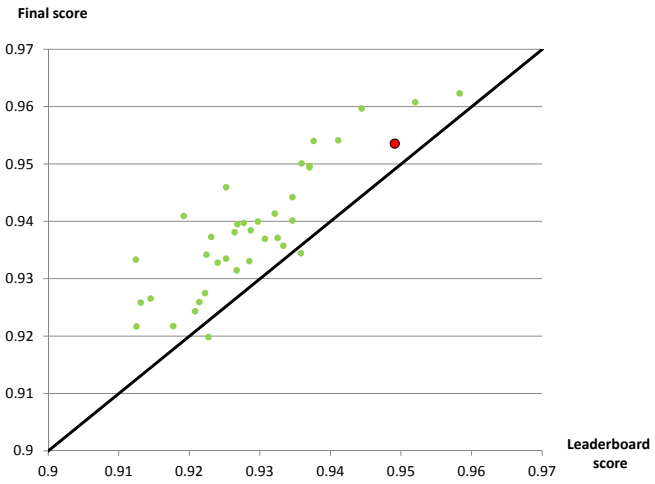


Fig. 4. Final challenge results

D. Final submission

The sensibility analysis of performance versus number of selected variables in Section III-C shows that the initial submission (see Section III-B) is competitive w.r.t. the objective of the challenge. Given the expected small and unreliable improvement in challenge score, the potential risk of using too few variables (larger expected variance) and the ignorance regarding the behavior of the ensemble classifier used in the challenge, we finally came back to the first submission. We removed the duplicate variables (keeping 97 variables), and obtained a challenge leaderboard score of 0.9491.

E. Challenge results

The last day of the challenge, our submission was ranked 3rd on the leaderboard. Usually, in data mining challenges, the participants tend to overfit the leaderboard score with many submissions, and we anticipated to get lower scores in the final results. Figure 4 shows the leaderboard versus final scores of all participants that obtained a score beyond that of the organizer's baseline; our score is represented by the red circle. Our final score (0.9536) was improved by 0.5% compared to the leaderboard score, which was a good surprise. Overall in this challenge, the final scores where improved on average by 1%. Participants ranked 4^{th} to 6^{th} on the leaderboard dataset got a 1.5% improvement of their final score and we finally got ranked 6^{th} in the final evaluation, 0.9% behind the winner.

IV. Conclusion

In most data mining projects, specific business requirements and constraints must be fulfilled. Several criterions must be taken into account, such as the time spend for the project, the training time, the deployment time, the intepretability of the models, the predictive accuracy. The AAIA'14 Data Mining Competition was an interesting challenge that focused on predictive accuracy versus number of selected variables. We have shown that using the Selective Naive Bayes classifier allows to quickly and efficiently obtain a competitive solution. We have also presented a sensitivity analysis between the two challenge criterions, that presents all possible trade-offs along a Pareto curve. This kind of analysis might be helpful to fulfill requirements in real world data mining projects.

References

[1] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[2] M. Boullé, "Compression-based averaging of selective naive Bayes classifiers," *Journal of Machine Learning Research*, vol. 8, pp. 1659–1685, 2007.

[3] P. Langley, W. Iba, and K. Thompson, "An analysis of Bayesian classifiers," in *10th National Conference on Artificial Intelligence*. AAAI Press, 1992, pp. 223–228.

[4] D. Hand and K. Yu, "Idiot bayes ? not so stupid after all?" *International Statistical Review*, vol. 69, no. 3, pp. 385–399, 2001.

[5] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," in *Proceedings of the 12th International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 1995, pp. 194–202.

[6] H. Liu, F. Hussain, C. Tan, and M. Dash, "Discretization: An enabling technique," *Data Mining and Knowledge Discovery*, vol. 4, no. 6, pp. 393–423, 2002.

[7] M. Boullé, "MODL: a Bayes optimal discretization method for continuous attributes," *Machine Learning*, vol. 65, no. 1, pp. 131–165, 2006.

[8] ——, "A Bayes optimal approach for partitioning the values of categorical attributes," *Journal of Machine Learning Research*, vol. 6, pp. 1431–1452, 2005.

[9] P. Langley and S. Sage, "Induction of selective Bayesian classifiers," in *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, 1994, pp. 399–406.

[10] R. Kohavi and G. John, "Wrappers for feature selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[11] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[12] J. Hoeting, D. Madigan, A. Raftery, and C. Volinsky, "Bayesian model averaging: A tutorial," *Statistical Science*, vol. 14, no. 4, pp. 382–417, 1999.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 361–367
DOI: 10.15439/2014F499
ACSIS, Vol. 2

Building an Ensemble from a Single Naive Bayes Classifier in the Analysis of Key Risk Factors for Polish State Fire Service

Stefan Nikolić, Marko Knežević, Vladimir Ivančević, Ivan Luković
University of Novi Sad, Faculty of Technical Sciences,
Trg Dositeja Obradovića 6, 21 000 Novi Sad, Serbia
Email: {stefan.nikolic, marko.knezevic, dragoman, ivan}@uns.ac.rs

***Abstract*—In this paper, we describe our solution in a competition that required performing data mining to identify key risk factors for the State Fire Service of Poland. The goal was to create an ensemble of Naive Bayes classifiers that could predict incidents involving firefighters, rescuers, children, or civilians. To this end, we first created a single Naive Bayes classifier and then partitioned the set of attributes used in that classifier. The attribute subsets were used to create new Naive Bayes classifiers that would form an ensemble, which generally performs better than both the single classifier and ensemble obtained by searching over all attributes considered when creating the single classifier. The application of our approach yielded a solution that ranked third in the competition.**

I. Introduction

THE main problem in our study is how to use data from incidence reports to identify key factors influencing the risk of serious injuries in actions carried out by the Polish State Fire Service. The dataset and the task description were provided by the organizers of the data mining competition hosted within the framework of the 9th International Symposium on Advances in Artificial Intelligence and Applications (AAIA'14), which is a part of the Federated Conference on Computer Science and Information Systems (FedCSIS) 2014. Additional information about the competition and its propositions is available in .

In accordance with the competition instructions, we performed data mining on the incidence reports. These reports are structured as a single table, in which each row represents one report and each column represents one attribute, i.e., a potential risk factor for injury during interventions and rescue operations. The result of this study is an ensemble of Naive Bayes classifiers that could be used to predict injuries of involved rescuers or civilians based on the most important risk factors.

We describe how we created such an ensemble and compare its performance to that of a single classifier and ensemble obtained by another method that searches over larger set of attributes. The utilized approach represents the main contribution of this paper.

The research presented in this paper was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, Grant III-44010

In addition to Introduction and Conclusion, the paper features four sections. In Section II, we present the dataset and the task in the competition. In Section III, we present our approach to creating both a single Naive Bayes classifier and a Naive Bayes ensemble. In Section IV, we offer the comparison of performance of the single classifier and ensembles formed using the proposed approach. In Section V, we review work related to predicting fires and associated injuries.

II. Problem Description

In this section, we provide an overview of the dataset and outline the shared competition task.

A. Dataset

The provided dataset is extracted from 50,000 reports, which correspond to actions carried out by the Polish State Fire Service within the city of Warsaw and its surroundings between 1992 and 2011. The dataset is highly dimensional and given in form of a table, in which each report is described by 11,852 attributes. All of these attributes are discrete and only a few have more than two possible values. The data are also sparse, since only a small fraction of the attributes has a non-zero value for a particular report. There are three binary decision attributes that describe whether there were casualties among firefighters, children or other involved people, respectively. All three decision attributes are highly imbalanced, since the positive classes correspond to relatively rare events.

B. Task

The task is to identify attributes that could be used to robustly assign reports to corresponding decisions labels. As defined by the organizers, the quality of a solution is assessed by measuring performance of a classifier ensemble composed of Naive Bayes models. Those models are constructed using ten attribute sets, separately for each decision attribute. An output of the ensemble is computed by averaging probabilities of the positive classes returned by

978-83-60810-58-3/$25.00

individual Naive Bayes models. The performance of the ensemble is measured by taking an average Area Under the ROC Curve (AUC) over the probability predictions for each decision attribute, decreased by a penalty for using a large number of conditional attributes. Namely, if we denote the chosen set of attributes by s, the total number of attributes used (with repetitions) by $|s|$, and AUC of a classifier ensemble for the i-th decision attribute by $AUC_i(s)$, then the quality measure for the assessment of a chosen set can be expressed as:

$$score(s) = F\left(\frac{1}{3}\sum_{i=1}^{3} AUC_i(s) - penalty(s)\right) \quad (1)$$

$$penalty(s) = \left(\frac{|s| - 30}{1000}\right)^2 \quad (2)$$

$$F(x) = \begin{cases} x & \text{for } x > 0 \\ 0 & otherwise \end{cases} \quad (3)$$

C. Approach to building an ensemble

In this research, we base our work on the task defined in the previous subsection. However, we are also interested in finding a single classifier, using a similar evaluation function, which is calculated as an average of three AUC values decreased by the penalty. Then, we examine whether the single classifier could be divided so to obtain an ensemble yielding higher accuracy compared to the original classifier. We also try some other, more common method for creating an ensemble. Finally, all of these approaches are compared.

III. Methodology

Our approach to ensemble creation was motivated by the task defined in the previous section. We were first interested in finding a single classifier. Then, we examined how to obtain an ensemble. We used two approaches for creating an ensemble - a division of the single classifier and a search over the same attributes that were considered (searched) when creating the single classifier. In this section, we describe how we select attributes to train a single classifier and how an ensemble may be formed using a custom boosting-based algorithm.

A. Attribute ranking

As already mentioned, the used dataset is highly dimensional. Therefore, in order to find the best classifier (model), it is not practical to search over the entire set of attributes. Hence, we first ranked all attributes according to their relation to the decision attributes. Since the used dataset is large, in order to speed up the process of testing different ranking methods, we manually constructed a small dataset that had similar characteristics as the original dataset, i.e. it included discrete (usually binary) attributes and sparse data. Using this small dataset, we could analyze the results of different ranking methods. The chi-square test [3] gave very good results, so we chose it as our method for attribute ranking.

For each attribute, using the chi-square test, we calculated three weight coefficients, each representing a degree to which an attribute is related to one of the three decision attributes. Finally, for each decision attribute, we ranked all attributes according to their weight.

B. Creating a single classifier

After the ranking, we started a search for a single classifier. Once the search is finished, we should have an approximate accuracy degree that can be reached with one classifier, but also the approximate limit in the number of attributes for a classifier (single or ensemble), since we have a penalty for using a large number of attributes. In order to find the set of attributes that represents the best classifier we used algorithms based on forward and backward search [4].

First, for each of the three decision attributes, we selected the top 50 attributes with respect to their weight rank. The union of these three groups of attributes yielded a set of 121 attributes. We then executed forward search on this set in order to find the best subset, i.e. the one offering the highest accuracy when predicting the values of three decision attributes. The function responsible for evaluating these subsets uses 5-fold cross validation. For each fold, it creates three Naive Bayes models, one for each decision attribute. It calculates predictions using each model and assesses the individual model accuracy using the AUC value. We then calculated the average of these three AUC values and decreased it by the penalty. The overall accuracy is calculated by averaging these values over all five folds.

The forward search process narrowed down the set of 121 attributes to a starting set of 41 attributes, which yield a model whose overall accuracy is 0.950, as calculated by our evaluation function. To improve accuracy, we had to increase the number of attributes being searched. One option was to include more than 50 attributes for each decision attribute and start a forward search from the beginning. However, due to a lack of computing power, we had to gradually increase the number of attributes.

First, for each of the three decision attributes, we selected attributes that were initially ranked between 51 and 75. We created a union of the three sets of attributes with these rankings. This union was used in a forward search that started not from the empty set but from the starting set. This forward search led to the inclusion of new attributes. Next, we performed a backward search, which resulted in the removal of some attributes.

We repeated the whole process and gradually increased the number of attributes to 350. The latest iteration in attribute selection resulted in 72 attributes, which yield a model whose overall accuracy is about 0.965, as calculated by our evaluation function.

The forward search is an iterative process. In each iteration, all attributes that are not already in the model are tested to be included. The attribute that yields the biggest improvement in accuracy is added to the model. Usually, attributes that show improvement in one iteration showed improvement in the previous iteration too. Considering this, in order to speed up the process of searching, we implemented forward search so that in every new iteration it searches only over the attributes that showed improvement in the previous iteration. Also, backward search is implemented in the same manner, so that in every new iteration it only considers attributes whose potential removal showed improvement in accuracy in the previous iteration. In the rest of this paper, whenever we mention the terms forward or backward search, we refer to variants described in this subsection.

It is worth noting that we tried to include additional ranking to speed up the process of searching even more. We tried to select some large number of attributes from each ranking (based on the chi-square test), unite selected attributes and rank them according to the metric that measures how good predictors they are. The metric measures the predictive quality of a single attribute as the accuracy of the model formed using just that one attribute. This additional ranking did not lead to any improvement, i.e. the forward search process did not include attributes with higher weight (according to this metric) as often to justify this method.

C. Creating an ensemble

The main idea behind ensemble systems is to create many classifiers and combine them so that the combination improves upon the performance of a single classifier. Generally, good ensembles should demonstrate diversity, i.e. each classifier should make errors on different examples, so that in combination these classifiers could reduce the total error. Algorithm that we used in order to create a diverse set of classifiers is based on the idea of boosting method, more concretely on the idea behind the AdaBoost algorithm [5].

We want to examine two approaches when creating ensembles. The first one is to try to use one strong single classifier that may be constructed after searching over a set of attributes and to split it into a set of classifiers. The other is to search through that entire set of attributes in order to find a different ensemble. Finally, we may observe differences between two ensembles. In order to make a comparison, these ensembles will have similar structure, i.e. they will include the same number of attributes and the cardinalities of the corresponding classifiers will be the same.

The advantage of the first approach is that it consumes less time, because it searches over a smaller set of attributes. Even if we take into consideration the time needed to find a single classifier, this method is still faster, since our search algorithm needs less time to find one big classifier than ten smaller classifiers where for each of them the search starts from the beginning. We can expect the first method to yield lower accuracy, because it searches over a limited set of attributes. However, it may be less prone to overfitting since it uses only significant attributes selected for a single classifier, but this could be the subject of further analysis.

In this research, we will only consider ensembles in which classifiers have balanced cardinality. Moreover, we will not examine cases when different classifiers in an ensemble may include the same attributes. These two constraints may be part of future research.

D. Boosting-based algorithm for ensemble creation

Our algorithm for ensemble creation follows the general idea of the well-known AdaBoost algorithm. It is an iterative method, where in each iteration we construct one classifier that focuses on examples that are misclassified by previous classifiers.

The algorithm uses ten iterations, as we need ten classifiers. At the beginning of each iteration we construct a training set, by sampling with replacement 50,000 examples from the original training set. Every example has the assigned probability to be sampled. The probability depends on the accuracy with which the example was classified by the previous classifiers (constructed in previous iterations). The more misclassified an example, the higher is its probability to be sampled for the next iteration. Therefore, each new classifier focuses on examples that were misclassified. Initially, the probabilities (p) have uniform distribution, so that in the first iteration every example has the same probability of being sampled.

In each iteration, we search for a classifier that maximizes accuracy, which is calculated as average AUC value for three decision attributes. The selection process is based on the same forward and backward search algorithms that we used when building a single classifier. Once the classifier is found, the probability for each example is updated in accordance with the classification error (err). When classifying, the Naive Bayes method calculates probabilities, so for each example it yields probabilities for three decision attributes. The average of these three probabilities indicates accuracy, and the classification error is calculated as a complement. Classification error is in the interval [0,1]. Here, we introduce a small constant c, which is close to 0, so that probabilities are multiplied with value from interval [c,1]. This constant is used because we do not want the probability of an example classified with error 0 to be multiplied by 0. All the steps executed in one iteration are as follows:

1. Create a training set, using the probabilities p.
2. Using this training set, find the most accurate classifier.
3. Update the probabilities according to the formula: $p = p * (c + (1 - c) * err)$
4. Normalize the probabilities so their sum is 1.

TABLE I
NUMBER OF USED ATTRIBUTES AND RESULTS FOR EACH STEP OF MODEL SELECTION

Number of attributes from rankings	Total number of attributes	Number of attributes in the model	Number of inserted + removed attributes	Overall accuracy of the model
50	121	41	41	0.9500514
75	175	53	14 + 2	0.9561488
100	233	59	6 + 0	0.9571192
150	335	66	9 + 2	0.9586754
200	430	67	7 + 6	0.9605470
250	540	69	6 + 4	0.9630121
300	628	70	2 + 1	0.9638735
350	727	72	4 + 2	0.9648648

If the repetition of attributes among classifiers is not allowed, this algorithm may yield imbalanced classifiers, i.e. classifiers that have very different accuracies. For instance, the first classifier could contain the most significant attributes, because it is created using the uniform distribution of probabilities. Ensembles usually demonstrate better performance if the classifiers are more balanced. Hence, if necessary, we may try to balance these classifiers to improve the overall accuracy.

IV. RESULTS AND DISCUSSION

In this section, we provide an overview of the performance of the single classifier, and ensembles formed by dividing the single classifier and searching over all attributes considered when creating the single classifier. Finally, all of these approaches are compared with respect to the predictive quality of models obtained.

A. Creating a single classifier

As discussed above, in every step of the (model) selection process, we increased the number of attributes being searched. In Table 1, we present the results obtained in each of these steps. Each row contains data about one step in the process. In the first column, we show the number of attributes from each of the three rankings considered in the search process. The second column indicates the total number of attributes considered in the search process, i.e. the union of all attributes obtained from the three rankings. The cardinality and accuracy of the obtained model are presented in the third and fifth column, respectively. Furthermore, the fourth column indicates the number of attributes inserted and removed from the model during one step.

We stopped at the step that included 350 attributes from each of the rankings. Certainly, we could include additional steps in order to improve our model. Each step terminates after a relatively short period. In the future, we can execute further steps, and we expect further improvement in accuracy. We do not expect the cardinality of the model to change very much because of the penalty for using a large number of attributes.

B. Creating an ensemble

Since some of the described methods for creating ensembles are time-consuming, we decided to first test our methods on the initial set used in the process of creating a single classifier (the first row in the Table 1). Later, chosen methods are applied on the final set (the last row in Table 1). The initial set has 121 attributes and yields a classifier that includes 41 of these attributes.

In the first approach, we want to use the classifier and try to split it into ten smaller ones, which are further combined into an ensemble. The cardinalities of these classifiers will be balanced, so there will be nine classifiers with 4 attributes and one with 5.

In order to check the probability that this kind of split yields an ensemble that is more accurate than the single classifier, we made 100 random splits (each split yielding nine subsets with 4 attributes and one with 5). For each of these 100 ensembles, we observed the value that represents accuracy, calculated by the formula defined in the task and using 5-fold cross validation. In the next table (Table 2) we give some quantitative measures regarding these values for all of the 100 ensembles (minimum, maximum, quartiles and mean). The histogram (Fig. 1) shows how the values are distributed.

The accuracy of a single classifier of 41 attributes is 0.9500514. From Table 2 and the accompanying histogram, we may see that there is a large proportion of ensembles whose accuracy is higher than that of the single classifier.

We aim to find a method that yields an ensemble with high accuracy, ideally higher than all of the ensembles obtained after random splits. Here, we test our boosting algorithm to split the classifier (41 attributes). We executed this algorithm with different values of constant c. We present these results in Table 3.

TABLE II
QUANTITATIVE MEASURES FOR ACCURACY OF 100 GENERATED ENSEMBLES

Min.	1st Qu	Median	Mean	3rd Qu	Max.
0.9477	0.9490	0.9496	0.9496	0.9501	0.9515

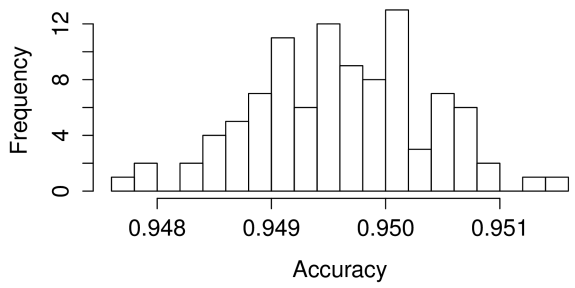


Fig. 1 Accuracy of generated ensembles

TABLE III
ACCURACY OF ENSEMBLES OBTAINED FOR DIFFERENT VALUES OF CONSTANT C

c value	Ensemble accuracy
0.01	0.9494519
0.001	0.9495294
0.0001	0.9498423

As expected, lower values of *c* yield higher accuracy. When we decrease this constant, the examples, with error whose order of magnitude is same as for *c*, have more accurate probabilities. It is possible to decrease the constant further. However, we do not expect this to offer significant improvement, since the number of examples with error lower than 0.0001 is small. We tried to modify our algorithm so that for sampling we use squared (and normalized) probabilities to additionally emphasize differences in probabilities, but this did not produce an ensemble with higher accuracy.

It is important to mention that, when executing our algorithm, we also tried to calculate probabilities according to the original AdaBoost algorithm, but this gave lower accuracy. Most probably, this is because the calculated values used for updating probabilities do not emphasize differences between correctly classified examples and misclassified examples, as much as our algorithm does.

We may see that, when *c* is 0.0001, our algorithm produced an ensemble only slightly more accurate than half of the ensembles obtained by random splits. This does not present a satisfactory result.

One of the reasons why the ensemble produced by the boosting algorithm has low accuracy could be that the classifiers from the ensemble are imbalanced with respect to accuracy. The analysis indicates that the first classifier contains the most significant attributes, because it was created using the uniform distribution of probabilities for sampling. Every subsequent classifier usually has lower accuracy when compared to the previous one.

Among ensembles obtained by random splitting, when we compare those with high accuracy to those with low accuracy, ensembles with higher accuracy usually contain more balanced classifiers. For each ensemble, we may observe the values of accuracy for its ten classifiers and determine standard deviation for these ten values. Lower standard deviation means having more balanced classifiers. In the next Table 4, we show standard deviations for five most accurate and five least accurate ensembles.

From Table 4, it is clear that ensembles with higher accuracy have lower standard deviation. However, this is a characteristic that should be considered, but it is not a rule. When we calculate standard deviation for an ensemble obtained by our boosting algorithm with constant 0.0001, we get a value of 0.1259726. This is much higher deviation when compared to all ensembles from Table 4, and probably from all other ensembles obtained by random splitting. However, the accuracy of this ensemble is better than that of one half of the ensembles obtained by random splitting, as we showed earlier.

Therefore, we will try to balance classifiers in the ensemble obtained by boosting algorithm, with respect to their accuracy. Balancing will be done in iterations. In each iteration, we select classifiers with maximum and minimum accuracy. Then, we try to swap two attributes from those classifiers. We examine each pair of attributes, until the swap yields an improvement greater than some threshold value we set. When the swap is executed, we continue to the next iteration. This process improves accuracy and usually decreases standard deviation. Also, it usually stops after a few iterations. When this method is executed on an ensemble obtained by the boosting algorithm, it stopped after three iterations and improved accuracy to 0.9512469. A small number of iterations is favorable since we do not want to change our ten classifiers too much, considering that boosting algorithms grouped attributes so that each classifier is focused on different set of examples. Accuracy yielded after balancing is high, and only two ensembles from those obtained from random splits have better accuracy.

TABLE IV
STANDARD DEVIATION FOR TEN ENSEMBLES WITH HIGHEST AND LOWEST ACCURACY

ensemble (rank)	standard deviation	ensemble accuracy
1st	0.06989024	0.9515106
2nd	0.07783254	0.9513274
3rd	0.08413265	0.9509978
4th	0.07679123	0.9508524
5th	0.08046396	0.9507565
96th	0.1000637	0.9482818
97th	0.09956434	0.9482052
98th	0.1126992	0.9479637
99th	0.08952922	0.9479142
100th	0.09684023	0.9477181

We also executed balancing on other ensembles that have similar accuracy as ensemble obtained by the boosting algorithm, but none yielded as high improvement. This could indicate that the way in which attributes are grouped by boosting is important, i.e. that the diversity of ensembles is important, and that it may not be easily compensated by balancing. However, we would need to do more testing to confirm that. We also experimented with executing the boosting method on the same set of 41 attributes, where we allowed repetition of attributes among classifiers, but this yielded ten similar classifiers and much lower accuracy.

In our second approach, we executed our boosting algorithm on the set of all 121 attributes. As we aim to compare this approach to the first one, we constructed an ensemble with the same characteristics. That ensemble included classifiers with same cardinalities (nine classifiers with cardinality 4, and one with cardinality 5), with no attribute repetitions. The set of attributes included in this ensemble was much different when compared to the set of 41 attributes from the single classifier. The accuracy of the obtained ensemble is 0.9379888. This accuracy is much lower compared to all ensembles constructed by random splitting.

Finally, considering results presented above, we chose only to test our first approach on the single classifier obtained in the final step. This classifier contains 72 attributes, hence we constructed ensemble containing eight classifiers of seven and two of eight attributes. We tried both boosting with balancing and random splitting. Again, some of the ensembles obtained by random splitting outperformed both the single classifier and the ensemble obtained by boosting and balancing. Our final result was an ensemble yielding the highest accuracy.

V. Related Work

There are many attempts to use fire related data for predicting, managing and reducing injuries caused by fire outbreaks. One group of such studies estimates future wildfire activities in order to reduce negative effects and facilitate its management, limiting the most destructive aspects of fire. Beckage and Platt [6] explored potential of a time series model that considers the area burned in previous years and the Southern Oscillation Index (SOI) condition to predict the area burned in the current years. However, the authors raise attention that their prediction is not true for "out of sample" predictions since model selection, parameter estimation, and the modeling process itself used the data that were to be predicted. Furthermore, Yue, et al. [7] managed to improve their regression and parameterization models for predicting wildfire activity by combining them with an ensemble of climate model projections for 2050 conditions. Moreover, they observed importance of considering meteorological attributes other than temperature in predicting changes in area burned.

The second group of research focuses on predicting injuries and identifying key factors influencing the risk of injuries in various incidents. One such analysis of data from population-based case-control study in King County, Washington [8] showed that there is association between smoking and residential fire injuries. By using relative risk estimation as a uniform method of comparison Warda, et al. [9] summarized house fire injury risk factor data retrieved from fifteen relevant articles. The authors stress that the research is relevant for developing, targeting, and evaluating preventive strategies. In their study, Burgess, et al. [10] compared injury rates among various fire departments and observed that the rates vary substantially. Variations in work practice and risk management regulations as well as in reporting practices were proposed as a potential explanation for the study findings.

Selection of significant attributes is deemed a prerequisite for constructing an accurate prediction model. In this sense, Shai [11] examined social and demographic predictive factors using multiple regression. Besides identifying significant attributes in predicting fire injuries for the civilian population, the author observed interaction effect between age of housing and income. The results of such research showed that older housing, low income, the prevalence of vacant houses, and ability to speak English (native language for area of Philadelphia) have significant effects on fire injury rates. In this research, we also aimed to find features that make accurate predictions regarding injuries in fire incidents. As opposed to Shai, we used Naive Bayes (NB) models, feature ranking and feature subset selection methods. The combination of multiple feature selection methods allowed us to capture complex feature interactions and select the model with the highest predictive capacity. Moreover, we used these methods to create both a single classifier and classifier ensemble, and to compare the two approaches. It was shown that ensemble models can yield higher accuracy than single models [12].

VI. Conclusion

In this paper, we present an approach for creating an ensemble of Naive Bayes classifiers from a single classifier. The creation of this approach was motivated by the need to identify key risk factors for the Polish State Fire Service as part of a data mining competition. The use of the proposed approach yielded an ensemble that generally performs better than both a single classifier and ensemble obtained by searching over all attributes that were considered when creating the single classifier. Moreover, the trained ensemble ranked third in the competition.

The advantage of this approach is that the model can be incrementally improved relatively quickly. We can include more steps in the process of selection of a single classifier in order to improve the classifier. Every step requires relatively small amount of time to finish. After executing these steps, we can make a division of the single classifier in order to obtain a new ensemble. The process of division does not consume much time.

In the future research, we would like to improve our methodology. As our results suggest, one single classifier can be divided into smaller ones which in combination yield higher accuracy. Our goal is to find a method for classifier division that would give a good ensemble, i.e., an ensemble that would outperform all of the ensembles generated by random divisions. Our method did not give completely satisfactory solution, but it is on the line of achieving the goal. So far, we tested only an algorithm based on boosting. It was used for division and when constructing an ensemble by searching (all attributes considered when creating the single classifier). We could try to improve this algorithm, but also to try other different algorithms for creating ensembles, since there are many others known in literature. Later, we should make further comparisons between approaches where we construct an ensemble from a single classifier and by searching. Finally, we could experiment with the possibility of attribute repetitions and different cardinalities among classifiers in an ensemble in order to obtain higher accuracy.

REFERENCES

[1] (2014, 02. June). *AAIA'14 Data Mining Competition*. Available: https://fedcsis.org/2014/dm_competition

[2] (2014, 02. June). *AAIA'14 Data Mining Competition: Key risk factors for Polish State Fire Service*. Available: http://challenge.mimuw.edu.pl/contest/view.php?id=83

[3] P. Romanski, "FSelector: Selecting Attributes," *Vienna: R Foundation for Statistical Computing,* 2009.

[4] E. Cantu-Paz, S. Newsam, and C. Kamath, "Feature Selection in Scientific Applications," in *Proc. ACM International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 788-793, http://dx.doi.org/10.1145/1014052.1016915

[5] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *Computational learning theory*, 1995, pp. 23-37, http://dx.doi.org/10.1007/3-540-59119-2

[6] B. Beckage and W. J. Platt, "Predicting severe wildfire years in the Florida Everglades," *Frontiers in Ecology and the Environment,* vol. 1, pp. 235-239, 2003, http://dx.doi.org/10.1890/1540-9295(2003)001[0235:PSWYIT]2.0.CO;2

[7] X. Yue, L. J. Mickley, J. A. Logan, and J. O. Kaplan, "Ensemble projections of wildfire activity and carbonaceous aerosol concentrations over the western United States in the mid-21st century," *Atmospheric Environment,* vol. 77, pp. 767-780, 2013, http://dx.doi.org/10.1016/j.atmosenv.2013.06.003

[8] J. E. Ballard, T. D. Koepsell, and F. Rivara, "Association of smoking and alcohol drinking with residential fire injuries," *American journal of epidemiology,* vol. 135, pp. 26-34, 1992.

[9] L. Warda, M. Tenenbein, and M. E. Moffatt, "House fire injury prevention update. Part I. A review of risk factors for fatal and non-fatal house fire injury," *Injury Prevention,* vol. 5, pp. 145-150, 1999, http://dx.doi.org/10.1136/ip.5.2.145

[10] J. L. Burgess, M. Duncan, J. Mallett, B. LaFleur, S. Littau, and K. Shiwaku, "International comparison of fire department injuries," *Fire Technology,* pp. 1-17, 2013, http://dx.doi.org/10.1007/s10694-013-0340-y

[11] D. Shai, "Income, housing, and fire injuries: a census tract analysis," *Public health reports,* vol. 121, 2006.

[12] A. Tsymbal, M. Pechenizkiy, and P. Cunningham, "Diversity in search strategies for ensemble feature selection," *Information fusion,* vol. 6, pp. 83-98, 2005, http://dx.doi.org/10.1016/j.inffus.2004.04.003

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 369–373
DOI: 10.15439/2014F497
ACSIS, Vol. 2

Identification of Key Risk Factors for the Polish State Fire Service with Cascade Step Forward Feature Selection

Piotr Płoński
Institute of Radioelectronics,
Warsaw University of Technology,
Nowowiejska 15/19,00-665 Warsaw, Poland
Email: pplonski@ire.pw.edu.pl

***Abstract*—The Polish State Fire Service gathers information about incidents which require their intervention. This information is stored to document the events. However, it can be very useful for new officers training, better identification of threats and planning of more effective procedures. The identification of key risk factors for casualties among firefighters, children or other involved people was a topic of data mining competition organized as a part of 1st Complex Events and Information Modelling workshop devoted to the fire protection engineering. The task of the competition was to find ten subsets of features for ten Naive Bayes classifiers. The ensemble output was used to predict occurence of casualities. Herein, the solution description that took 5th place is presented. The proposed method used cascade step forward feature selection procedure to find features subsets.**

***Index Terms*—key risk factors, fire service, Naive Bayes, feature selection, cascade step forward**

I. Introduction

THE POLISH EWID [2] reporting system is the Incident Data Reporting System (IDRS) used by Polish State Fire Service to gather information of their interventions in incidents. This data documents historical events. However, useful knowledge could be extracted from them, which can be later used for new officers training, preparation of safer and more effective procedures, and better understanding of danger factors in incidents [5], [4], [8]. The identification of key risk factors for casualties among firefighters, children and other people involved was a topic of data mining competition organized within the 9th International Symposium on Advances in Artificial Intelligence and Applications (AAIA) and was an integral part of the 1st Complex Events and Information Modelling workshop devoted to fire protection engineering. The competition results will bring data-driven insights into key risk factors in incidents and contribute to safety improvement, which is important for Fire Service supporting systems [6],[7].

The competition dataset comes from reports of the EWID system, which documents actions carried out by the Polish State Fire Service within the city of Warsaw and its surroundings in years 1992–2011. Each report obtains a feature vector descriptor after preprocessing [4], [5]. The competitors task was to find ten subset of features among over 11,000 discrete attributes describing 50,000 reports, which are relevant to the safety of people in incidents. Based on selected features, the ensemble of ten Naive Bayes classfiers [3] was created for each of three decisions variables:

1) injured firefighter in the action,
2) injured children in the incident,
3) other injured people involved.

They were used to evaluate the competition score metric, which considered the performance of the classfiers on each of the decision variable and penalizes large feature subsets. It is worth to note, that the same ten subsets of features were used in Naive Bayes construction for all decisions variables. The additional obstacle in analysis was sparsity of training data and rare occurance of positive values in decision variables.

The task of the competition can not be simplified to a sole feature selection problem. It is a problem of feature selection for ensemble of classifiers which should have the highest average accuracy in predicting three various dependent variables simultaneously with the smallest possible number of features. The proposed method used a cascade step forward selection of features that maximize the competition score metric on cross validation (CV) on training dataset. In each selection step, previously chosen features subsets were considered, therefore the proposed method is called 'Cascade Step Forward' (CSF) feature selection. The CSF procedure was speeded-up by initial features filtering and storing information about values occurences in CV folds.

The atricle is organized as follows: firstly detailed description of competition dataset, task and score metric are desribed; secondly, the proposed method is presented; then, obtained results are shown; finally, the conclusions and directions for future research are presented.

II. Methods

A. Data description

The training dataset available for participants consists of 50,000 incident reports. Each report was described using 11,852 discrete features. The majority of features were binary, with only few features with more distinct values (up to 5 values). The details of number of discrete values in features are

978-83-60810-58-3/$25.00

presented in Table I. The major values in the training dataset were zeros. From all 592,600,000 available values in training dataset only 5,217,892 have non zero values, which is only 0.8805% of all values. The sparsity and high dimensionality of data was implied by the nature of considered problem. The features correspond to the number of distinct words in the textual part of the reports (after lemmatization) and to several hundreds of features from the quantitative part of the reports [4], [5].

TABLE I
NUMBER OF DISCRETE VALUES IN FEATURES.

Discrete values	2	3	4	5
# of features	11826	9	7	10

For each report there were associated three binary decision varibles. The first decision attribute indicates incidents resulting in injury or death of a firefighter or a member of rescue team. The second decision variable indicates cases in which there were children among injured people and the third attribute identifies situations where civilians were hurt. All three decision attributes are highly imbalanced, since the positive classes correspond to relatively rare events. The details of positive values occurence in decision variables are presented in Table II

TABLE II
NUMBER OF POSITIVE VALUES IN DECISION VARIABLES.

Decision variable	# of positive values	Percentage
1	199	0.40%
2	366	0.73%
3	2955	5.91%

Let's denote dataset as $D = \{X_1, X_2, ..., X_N, Y_1, Y_2, Y_3\}$, where $N = 11,852$ is a feature number, X_i is a i-th feature vector and Y_1, Y_2, Y_3 stand for three decision variables, injury of firefighter, children, other involved people, respectively.

B. Task description

The competition task was to select ten subsets of features. They were used to build an ensemble of ten Naive Bayes classifiers for each decision variable. The sum of the output of classifiers ensemble was used to predict the occurence of positive values in each of decision variables. The accuracy of the selected features was computed with competition metric described below. It is worth to note, that there was a lower bound limit equal 3 for number of features in each subset.

C. Evaluation metric

The competition score metric can be expressed as:

$$score(s) = F\left(\frac{1}{3}\sum_{1}^{3} AUC_i(s) - \left(\frac{|s| - 30}{1000}\right)^2\right), \quad (1)$$

where

- $s = \{s_1, s_2, ..., s_{10}\}$ is a selected ten subsets of features,
- $|s|$ is a total number of selected features with repetitions,
- AUC_i is Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC) [3] computed for i-th decision variable,
- $F(x) = \begin{cases} x, & \text{if } x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$

The first term of eq.1 computes the average performance of classifier ensemble on all decision variables, whereas the second term penalizes the solutions with large number of selected features. It is worth to note, that penalization term vanishes when exactly three features are selected in each subset.

D. Proposed Method

The competition used a Naive Bayes classifier (NBC) [3] to evaluate the metric. The NBC is a classification method, which for a given sample $\mathbf{x} = \{x_1, ..., x_K\}$, with K features, calculates the posterior probability for all $y \in Y$, $p(Y = y|X_1 = x_1, ..., X_K = x_K)$, and assigns the class with the highest posterior probability. This can be expressed as:

$$y = \operatorname*{argmax}_{y \in Y} p(Y = y|X_1 = x_1, ..., X_K = x_K). \quad (2)$$

The posterior probability can be rewritten with Byes rule, the eq.2 becomes:

$$y = \operatorname*{argmax}_{y \in Y} \frac{p(Y = y)p(X_1 = x_1, ..., X_K = x_K|Y = y)}{p(X_1 = x_1, ..., X_K = x_K)}. \quad (3)$$

The evidence probability in denominator is the same for all classes and what is more, the NBC assumes that all features are conditionally independent given decision, thus the eq.4 can be written as:

$$y = \operatorname*{argmax}_{y \in Y} p(Y = y) \prod_{i=1}^{K} p(X_i = x_i|Y = y). \quad (4)$$

For discrete features the prior and likelihood can be computed as follows:

$$p(Y = y) = \frac{M_y}{M}, \quad (5)$$

and

$$p(X_i = x_i|Y = y) = \frac{M_{x_i,y}}{M_y}, \quad (6)$$

where

- M is total number of samples,
- M_y is number of samples with class label equal y,
- $M_{x_i,y}$ is number of samples with class label equal y and X_i feature equal to x_i.

The feature selection for single NBC can be done with greedy step forward (SF) procedure [3] with maximization of score with CV. The SF algorithm starts selection with empty subset of features $S_0 = \{\}$. Afterwards it checks the performance of the classifier with addition of each of the available features. The performance is computed on repeated (R_{cv} times) CV with drawing training and testing split for each repetition. The feature X_j which maximizes the quality metric is added to the subset, $S_1 = S_0 \cup X_j$. The whole procedure is

repeated till the required number of features L is selected or the required score value is achieved. The pseudocode of SF selection for single classifier is described in the Algorithm 1 listing.

Algorithm 1: The step forward feature selection procedure for single classifier.

input : $D = \{X_1, X_2, ..., X_N, Y_1, Y_2, Y_3\}$,
N number of available features,
L number of features to select,
R_{cv} repeats in cross validation.
output: The selected optimal subset S of features.
begin
  Set $S_0 = \{\}$
  for l *in 1 .. L* **do**
    for i *in 1 .. N* **do**
      Build a classifier H_i using as a feature subset $S_{l-1} \cup X_i$
      for c *in 1..* R_{cv} **do**
        Draw training and testing split of data
        Compute perfomance of classifier H_i on testing subset;
    Select classifier H_j with the highest average accuracy
    Set $S_l = S_{l-1} \cup X_j$

The SF procedure is applicable for selecting features for single classifier. It is inefficient for selecting features for ensemble of classifiers because for every classifier the similar subset of features will be assigned. The classifier ensemble requires a diverse subset of features for each classifier to obtain high accuracy [9]. To overcome this obstacle the 'Cascade Step Forward' feature selection procedure is proposed. The CSF algorithm, contrary to SF, searches for subsets of features for each classifier in the ensemble. It applies the SF procedure to find a subset of features for each classifier. However, in candidate feature scoring the performance is computed for ensemble instead of single classifier. The CSF procedure returns a set of feature subsets $S_{all} = \{S^1, ..., S^J\}$, where J is a number of classfiers in the ensemble. The pseudocode for CSF procedure is presented in Algorithm 2 listing.

E. Implementation Details

The greedy feature selection procedure has high computational cost. However, it can be decreased with filtering the features with low likelihood values. In feature selection only attributes with likelihood values greater than threshold value t for at least one decision variable were considered. The filtering condition can be expressed as:

$$p(X_i|Y_1) > t \vee p(X_i|Y_2) > t \vee p(X_i|Y_3) > t. \qquad (7)$$

The threshold value used was $t = 0.02$. After applying the eq.7 from inintial 11852 there remained 2333 features. The CSF procedure run only on remaining features.

Algorithm 2: The cascade step forward feature selection procedure for ensemble of classifiers.

input : $D = \{X_1, X_2, ..., X_N, Y_1, Y_2, Y_3\}$,
J number of classifiers in ensemble,
N number of available features,
L number of features to select,
R_{cv} repeats in cross validation.
output: The set of feature subsets for each classifier in ensemble $S_{all} = \{S^1, ..., S^J\}$.
begin
  Set $S_{all} = \{\}$
  for j *in 1 .. J* **do**
    Set $S_0^j = \{\}$
    for l *in 1 .. L* **do**
      for i *in 1 .. N* **do**
        Build a classifier H_i^j using as a feature subset $S_{l-1}^j \cup X_i$
        for c *in 1..* R_{cv} **do**
          Draw training and testing split of data
          Compute perfomance of ensemble of classifiers $\{H^1, .., H^{j-1}, H_i^j\}$ on testing subset
      Select classifier H_i^j with the highest average accuracy
      Set $S_l^j = S_{l-1}^j \cup X_i$
    Set $S_{all} = S_{all} \cup S^j$

In the proposed solution splitting dataset into training and testing subsets was performed many times during cross validation. Therefore, the counts of values occurences were stored in each fold to speed-up process of computing priors and likelihoods. The available dataset was splitted into $F = 500$ equally sized folds, from which $F_{tr} = 50$ and $F_{te} = 450$ were drawn for the training and testing respectively. Such uncommon partition provides a quite good matching between local CV scoring and public leaderboard score. The CV scoring was repeated $R_{cv} = 20$ times for each new feature testing. The i-th fold stores information M_y^i about samples number with class label equal y, and $M_{x_i,y}^i$ about number of samples with values equal x_i and class label y for all of considered features. Therefore, the probabilites needed for NBC construction can be computed as:

$$p(Y = y) = \frac{\sum_{i=1}^{F_{tr}} M_y^i}{F_{tr}\frac{M}{F}}, \qquad (8)$$

and

$$p(X_i = x_i|Y = y) = \frac{\sum_{i=1}^{F_{tr}} M_{x_i,y}^i}{\sum_{i=1}^{F_{tr}} M_y^i}. \qquad (9)$$

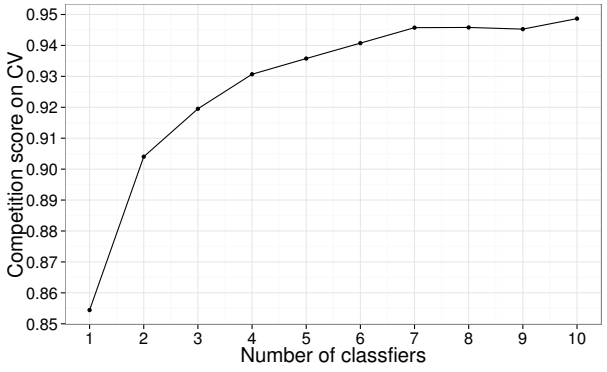


Fig. 1. The local CV score obtained for classifier ensemble in CSF feature selection.

The CSF feature selection was implemented in C++ to achieve high speed of computations.

III. Results

To omit the penalization term in the score metric (eq.1) there were selected exactly three features for each classifier. The selected features for each classfier are presented in Table III. It is worth to note, that selected features are only 0.25% of all available features. The obtained local CV scores during CSF selection for ensemble with different number of classifiers are presented in the Fig.1. It can be observed that the score is increasing when adding up to 7 classifiers into ensemble. For greater number of classifiers in the ensemble the score is stable. The local CV score was 0.9487, the public leaderboard score computed on approximately 10% of testing data was 0.9376, whereas score computed on full testing set was 0.9540. The solution that scored the 1st place achieved 0.9623 on full testing dataset, so there is only 0.0083 difference between proposed solution and the best one. The dependency between scores computed on public leaderboard and full testing dataset for solutions of all participants, with score on full testing dataset greater than 0.9, are presented in the Fig.2. It can be observed that for almost all solutions the score on public leaderboard was lowered with respect to score on the full testing dataset.

TABLE III
SELECTED ATTRIBUTES FOR EACH CLASSIFIER.

Classifier	Attribute 1	Attribute 2	Attribute 3
1	11701	5270	675
2	143	142	2182
3	691	5909	3735
4	10446	3492	2924
5	2887	8853	8914
6	7980	7148	72
7	11463	10882	1509
8	3963	258	4313
9	3596	8872	8249
10	7755	5270	6534

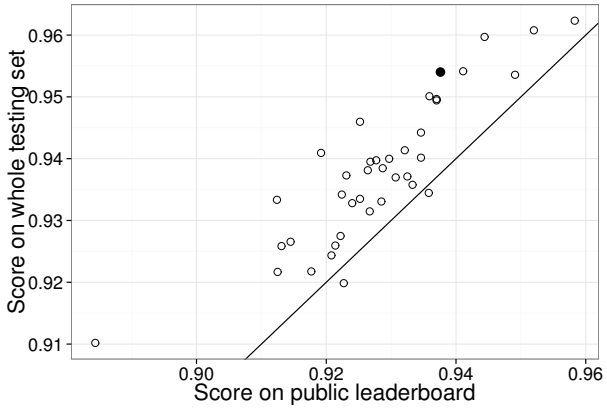


Fig. 2. The dependency between scores computed on public leaderboard (10% of testing set) and whole testing set for solutions of all participants with score on full testing set greater than 0.9. The solution presented in this paper is marked as filled black circle.

IV. Conclusion

The solution description that took 5th place in AAIA'14 Data Mining Competition: "Key risk factors for Polish State Fire Service" was presented. The proposed solution used a cascade step forward feature selection to select feature subsets for classifiers in the ensemble. The CSF maximize the competition score on cross validated training dataset in each step. To speed-up the selection process the initial filtering out of features with low likelihood were performed and number of occurence of feature values and class labels were stored in folds of training dataset. The identified key risk factors can be useful for Polish State Fire Service in new officers training, preparation of safer and more effective procedures and awareness of threats in actions.

The proposed CSF method can be applied for feature selection for other domains, for example in neuroimaging data analysis where data sets are highly dimensional and only small fraction of features are usable [10]. The performance of CSF procedure can be improved by cosidering several best features in each step instead of just one.

Acknowledgement

The author has been supported by the European Union in the framework of European Social Fund through the Warsaw University of Technology Development Programme.

References

[1] K. Bąk, A. Krasuski, M. Szczuka, "Searching for Concepts in Natural Language Part of Fire Service Reports," In: Proceedings of Concurrency, Specification and Programming; XXIII-th International Workshop, CS&P 2013, Warsaw, Poland, September 25-27, 2013.

[2] Collective Work (2001) Ewidencja zdarzen%■EWID99. Technical report, Abacus. http://www.ewid.pl/?set=rozw_ewid&gr=roz. Accessed date 23 April 2007

[3] T. Hastie, J. Friedman, R. Tibshirani, "The elements of statistical learning," Springer, 2009, DOI: 10.1007/978-0-387-84858-7

[4] A. Janusz, A. Krasuski, M. Szczuka, "Improving Semantic Clustering of EWID Reports by Using Heterogeneous Data Types," Lecture Notes in Artificial Intelligence, vol. 8170, 2013, pp. 304-314, DOI: 10.1007/978-3-642-41218-9_33

[5] A. Krasuski, A. Janusz, “Semantic Tagging of Heterogeneous Data: Labeling Fire & Rescue Incidents with Threats,” 8th International Symposium Advances in Artificial Intelligence and Applications, 2013, pp 77-82

[6] A. Krasuski, A. Jankowski, A. Skowron, D. ́lęzak, “From Sensory Data to Decision Making: A Perspective on Supporting a Fire Commander,” Web Intelligence/IAT Workshops, 2013, pp 229-236, DOI: 10.1109/WI-IAT.2013.188

[7] A. Krasuski, K. Kreński, S. Łazowy, “A Method for Estimating the Efficiency of Commanding in the State Fire Service of Poland,” Fire Technology vol.48, 2012, pp 795-805, DOI: 10.1007/s10694-011-0244-7

[8] A. Krasuski, P. Wasilewski, “The Detection of Outlying Fire Service’s Reports. The FCA Driven Analytics,” In Processings of the 11-th International Conferene on Formal Concept Analysis, 2013, pp 35-50

[9] B. Krawczyk, G. Schaefer, “A hybrid classifier committee for analysing asymmetry features in breast thermograms,” Applied Soft Computing, vol. 20, 2014, pp 112–118, DOI: 10.1016/j.asoc.2013.11.011

[10] P. Płoński, W. Gradkowski, K. Jednoróg, A. Marchewka, P. Bogorodzki, “Dealing with heterogeneous multi-site neuroimaging data sets: a study on discrimination of children dyslexia,” In: Ślęak, D., et al. (Eds.) Brain Informatics and Health, 2014, Lecture Notes in Artificial Intelligence, vol. 8609, 2014, pp 471-480

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 375–380
DOI: 10.15439/2014F502
ACSIS, Vol. 2

Robust Method of Sparse Feature Selection for Multi-Label Classification with Naive Bayes

Dymitr Ruta
Etisalat, British Telecom Innovation Centre
Khalifa University, Fatima F302, PO Box 127788
Abu Dhabi, UAE
Email: dymitr.ruta@kustar.ac.ae

***Abstract*—The explosive growth of big data poses a processing challenge for predictive systems in terms of both data size and its dimensionality. Generating features from text often leads to many thousands of sparse features rarely taking non-zero values. In this work we propose a very fast and robust feature selection method that is optimised with the Naive Bayes classifier. The method takes advantage of the sparse feature representation and uses diversified backward-forward greedy search to arrive with the highly competitive solution at the minimum processing time. It promotes the paradigm of shifting the complexity of predictive systems away from the model algorithm, but towards careful data preprocessing and filtering that allows to accomplish predictive big data tasks on a single processor despite billions of data examples nominally exposed for processing. This method was applied to the AAIA Data Mining Competition 2014 concerned with predicting human injuries as a result of fire incidents based on nearly 12000 risk factors extracted from thousands of fire incident reports and scored the second place with the predictive accuracy of 96%.**

I. Introduction

The unmanageable scale of big data comes in many forms symbolically paraphrased by *5Vs*: Volume, Velocity, Variety, Veracity and Value [1]. Huge volume defined by both the size or dimensionality of big data is one such "V" that particularly adversely affects computational complexity of the process of learning from data. The hype about big data may be therefore elusive while its possible value very difficult to extract. There are many examples reported in the literature that demonstrate both very powerful and very ineffective exploitations of large data sets for predictive tasks [4], [1], [1], [3].

Inspired by the pioneering work in [4], however, there is a widespread belief that the more data the better and the inability of exploiting it all is just a reflection of the predictor's weakness [3]. We argue, however, that a blind admission of all big data into the predictive modelling may be wrong or at least inefficient approach for some class of problems. Although certain cognitive tasks may indeed require billions of data points to reveal the full explanative power of the data [4], [5], our experience indicates that the majority of data problems can be explained by the relatively small data sample, which might be buried under the masses of big data. For these problems the availability of big data for predictive analytics widens the choice and the opportunity for both novel data exploitations and the improvement of predictive performance of the existing models.

As a result, the emerging paradigm of working with big data appears to be centred around careful data filtering, preprocessing, features generation and selection. Very often these procedures eliminate most of the original data leaving only essential evidence that retains almost complete explanative power [3]. What is more, the evidence reported in the machine learning literature indicates that given a typical supervised learning problem the key drivers for performance lie predominantly in the discriminative power and the choice of the data features rather than in the complexity of the predictive model [6], [7], [8], [9], [10]. All these points lead to a conclusion that when faced with the problem of learning from big data the main challenge and effort should be directed towards extracting or generating the key explanative data features while the actual learning and predictive performance could be delivered with relatively simple and robust learning model [10].

In line with this approach we have entered AAIA'2014 data mining competition with the intend to demonstrate how effective could be feature selection for supervised learning problems with very high dimensional data. We proposed a relatively easy and fast, greedy feature selection method that works particularly well with the large number of sparse features. In the competitive environment we will demonstrate that it delivers very high performance with a very simple predictor like Naive Bayes. We also propose much faster yet nearly equally robust feature selection method that eliminates completely the need of predictor application, and for that as we argue it is a very strong contender for real-time applications of predictive analytics on big data.

II. Task Description

AAIA Data Mining Competition 2014 was concerned with extracting the risk factors and attributes of fire incidents that would allow the most accurate prediction of human injuries or casualties as a result of these incidents. The total of 11852 features extracted from 50000 fire incident reports were presented as input data and the objective of the competition was to select a subset of features that would achieve the best predictive accuracy of detecting simultaneously the following 3 binary class target outputs with the Naive Bayes model:

- serious injury or death of one of the firefighters or members of the rescue team
- children were among injured people
- civilians were among hurt/injured people

978-83-60810-58-3/$25.00

Additional constraint enforced by the competition was the format of the solution and its assessment. The format of the solution was enforced to be organised within 10 feature subsets of at least 3 features each and the performance metric was set to be the area under the curve (AUC) of the receiver-operator curve (ROC) obtained from averaging the outputs from Naive Bayes classifier ensembles across all 3 target variables. The performance metric additionally incudes the penalty term that penalises for using many features in the solution as in the following:

$$score(s) = max\left\{0, \frac{1}{3}\sum_{i=1}^{3} AUC_i(s) - \left(\frac{|s|-30}{1000}\right)^2\right\} \quad (1)$$

Note that the size penalty term reduces to 0 when all 10 selected feature subsets have exactly 3 features. The problem is challenging due to the fact that the input data is huge, high dimensional and sparse in nature, while the class target values are highly imbalanced. Further difficulty is that the evaluation considers the average performance of de-facto 3 distinct classification problems sharing the same features. A successful feature selection method needs to find the compromise in maximising the average performance of all the 3 models at the same time.

III. Feature Elimination

Given the very large feature dimensionality and the sparse nature of the input data the first natural step is to eliminate redundant features that have no chance of contributing to the performance of the target prediction. The approach taken was that given the feature sparsity and huge imbalance of the target class variables, the features which have all non-zero values occurring only at negative class outputs have completely zero predictive power in isolation or in combination with other features. Denoting by $X^{[N\times M]}$ the matrix of input data and by $Y^{[N\times 3]}$ the matrix of corresponding class outputs we can safely eliminate redundant features by applying the following simple filtering expressed in Matlab formulation:

```
F = find(sum(X(any(Y, 2), :))>0);    (2)
```

This simple filtering resulted in elimination of 1931 (16.3%) redundant features. An interesting observation is that if the three target class variables were to be predicted and assessed separately, the above filtering would have resulted in much deeper reductions of: 6418 (54.1%), 6174 (52.1%), and 2146 (18.1%), respectively. What is more, separating predictive tasks would further allow to identify feature redundancy through containment. Namely, we can further eliminate a feature *A* whose non-zero intersection with the positive target class (true positives) is fully contained by other feature *B*, while its non-zero intersection with the negative class (false positives) fully contains feature *B*. What it means is that prediction with feature *A* would always be less accurate than with feature B which is guaranteed to make more positive predictions (true positives) at a lower costs (false positives). Such further elimination through feature containment would achieve the reductions of 9754 (82.3%), 9313 (78.6%), and 5005 (42.2%) respectively. It suggests that it might be much more efficient to model all three predictive tasks independently rather than force them to share the same feature subset of input data.

Constrained by the evaluation criterion defined by eq. 1 the feature set to work with had to be left with the only lightly reduced size - down to 9921 features, to avoid the loss of information.

IV. Naive Bayes Classifier

Naive Bayes (NB) is a simple yet very effective and fast probabilistic classification method that naively assumes that features are conditionally independent given the class value [11]. Given a binary classification problem with n features F_i the NB model tries to give the estimate of the posterior class probability given feature observations: $p(C, F_1, ..., F_n)$. From the chain rule applied to conditional probability definition the searched likelihood becomes:

$$p(C, F_1, ..., F_n) = p(C)p(F_1|C)p(F_2|C, F_1)... \\ ...p(F_n|C, F_1, F_2, ..., F_{n-1}) \quad (3)$$

which after applying the naive assumption of conditional feature independence simplifies to:

$$p(C, F_1, ..., F_n) = \frac{1}{Z}p(C)\prod_{i=1}^{n} p(F_i|C) \quad (4)$$

where Z is a constant scaling factor that is fixed for known feature variables.

Since most of the features are binary or categorical we are dealing with the multinomial distribution here, and constructing a posterior class likelihood is just a matter of calculating a product of class conditional probabilities of specific feature values observed for every input data instance. This process is critical and most often repeated when evaluating classification performance hence it is reasonable to speed it up by precalculating the class conditional likelihoods of all feature values empirically from the training data. Calculating the posterior would be then reduced to taking the relevant class conditional feature probabilities from the lookup table and multiplying them together or adding log likelihoods in odder to avoid the precision loss for small numbers.

Since the competition performance metric was an AUC of the ROC curve, the class posteriors are all that is required for the score calculation.

V. Sparse Feature Selection Strategies

Given the training input data of 50000 examples composed of nearly 10000 sparse features (after filtering) and a well defined and fast predictive performance metric defined in eq. 1 the objective was to extract 10 subsets of features that would maximise the expected predictive performance on the unseen testing set. Our preliminary investigations revealed that separating a validation set out of training data appears to be a good method for comparing the generalisation robustness of the strategies. On the other hand setting any data aside for the validation reduces the evidence that the predictive model is learnt on and hence may not give the best performance on the testing set. The approach that was finally taken was to use the actual performance feedback to decide whether a validation set improves the predictive performance on the testing set.

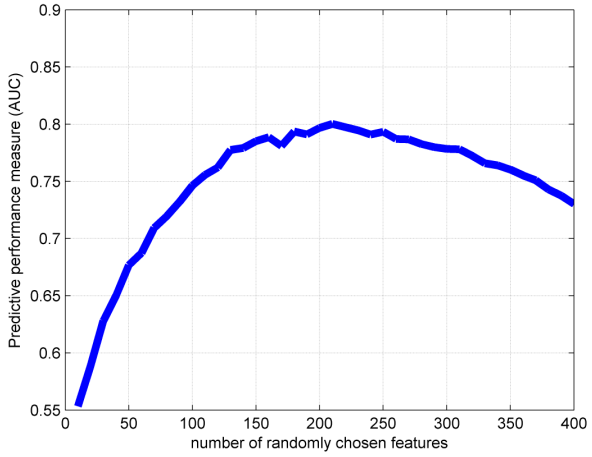


Fig. 1. Random subset method performance curve

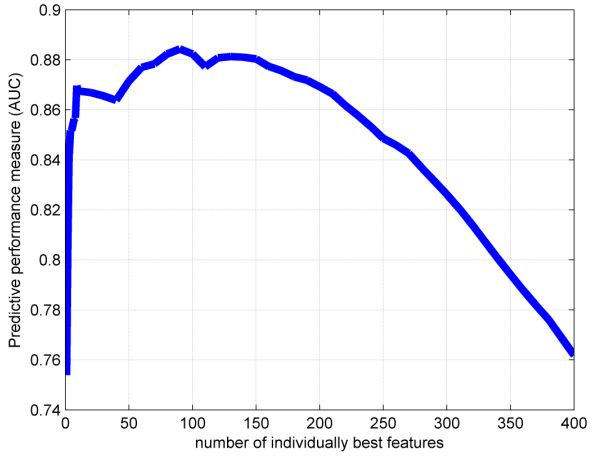


Fig. 2. Incremental single best performance curve

A. Random Feature Subsets (RFS)

Random feature subset method appears very naive for large feature space problems, however it was quickly developed to provide some intuition around the predictive value of the data and to set some baseline predictive performance levels. It was also useful for establishing the impact of the performance metric penalty term provided by eq. 1 and through some experiments draw an estimate of what might be the optimal size of the feature set that would maximise the competition performance criterion.

The performance of random feature subset selection method was evaluated using random features set sizes increasing from 10 to 500 at a step of 10 and obtaining corresponding performance measure inline with eq. 1. It has been repeated 50 times and the results averaged to build stable performance estimates for increasing number of random features included in the model. The resulting performance curve is presented in fig. 1.

As it can be seen from the figure, the performance of the random feature subset method is expectedly quite poor and peaks for roughly 200 features included in the model.

B. Incremental Single Best (ISB)

The random feature subset method performs quite poorly but it does not require any computation effort related to feature selection. Incremental single best method presented here goes a step further and shifts the balance towards improving the predictive performance at the relatively small prior computational cost of evaluating all individual features performance. Since each feature is evaluated in isolation, no conditional feature dependencies are considered, and the model build simply follows the greedy strategy of sequentially adding individually best available features until their combined predictive performance stops increasing. Fig. 2 illustrates the performance curve of such incremental single best selection strategy for the feature set sizes set from 1 to 500. Clearly the performance curve very quickly climbs to a much higher levels above 0.88 comparing to the random subset method and

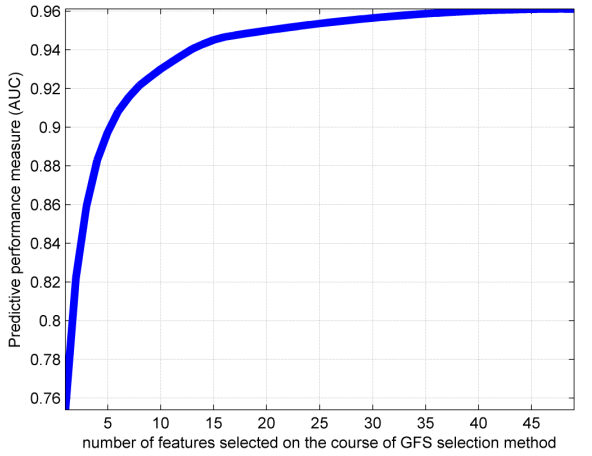


Fig. 3. Greedy forward search performance curve

it peaks for about 90 individually best features included in the model.

C. Greedy Forward Search (GFS)

The next feature selection strategy to consider is a traditional greedy forward search. This method would start from the same individually best feature. At each round it then checks the performance change of adding all remaining features one at the time to finally add a feature that results in maximum possible improvement of predictive performance. This search already introduces a significant processing cost as its computational complexity based in the number of performance evaluations is $O(N \times n)$ where N is the total number of available features and n stands for the number of features selected for the model.

The performance curve of the GFS is presented in fig. 4.

It might be surprising that it beats the performance of all previously presented selection methods with just 4 first features. The performance curve peaks with 49 features. The reported validation performance in excess of 0.96 was com-

parable to the training performance and on its own climbs to a competitive level. The advantage of GFS method is a rapid performance growth with just a few features yet the computational complexity heading towards quadratic order becomes a real issue here and caused the complete search to take few days on the standard PC. Further drawback of this method is that beyond few features the risk of falling into local shallow maxima grows really fast and affects the method ability to find robust solutions.

D. Diversified Greedy Backward-Forward Search (DGBFS)

Greedy forward search introduced in the previous section demonstrated really good potential for high predictive performance that is however hindered by the problem of local maxima trap. The proposed diversified DGBFS method tries to exploit the strengths of the forward search method while improving its flexibility to get out of local maxima traps, increasing the exposure to the diversity of the whole feature set and significantly improving the speed of the search.

The method starts from the same greedy forward search but rather than adding only the single feature that maximally improves the performance for every feature set scanning round it keeps adding all the features that improve the currently best performance. As a result a single forward scanning round could add hundreds of features instead of just one. What happens then is a backward search, in a sense that all features selected so far are attempted to be removed from the set and such removal is granted if it causes the performance improvement. Such backward search adds vital ability of the method to refine its earlier greedy choices by exploring latter additions that do not maximise the performance gain but lead to better longer term solutions.

The forward and backward scanning rounds follow each other in a sequence until for both not a single addition or removal is able to improve the performance. Since this method is dependent on the order of features presented for the scanning, feature indices are randomly permutated before each scanning round such that the whole feature space is equally exposed to the chances of being selected.

The complete performance curve across many rounds of additions and removals is visualised in fig. 4.

Forward moving sections represent the performance progression during forward search and backward sections reflect the corresponding performance gains during backward search. Notable is a big overshoot of the size of the selected feature set during the first forward search. This was the effect of the initial ease of improving the performance through additions. In fact most of the newly added features were later removed in the subsequent backward search since they were added not because they were very robust but because they were just better than random features initially populating the selected feature set.

The presented DGBFS feature selection method achieved the top expected performance of over 0.97 and was selected to generate solutions for the AAIA'14 Data Mining Competition. Both the initial feedback and the final assessment positioned its solution on a second place in the competition trailing just a fraction of a per cent behind the top wining solution.

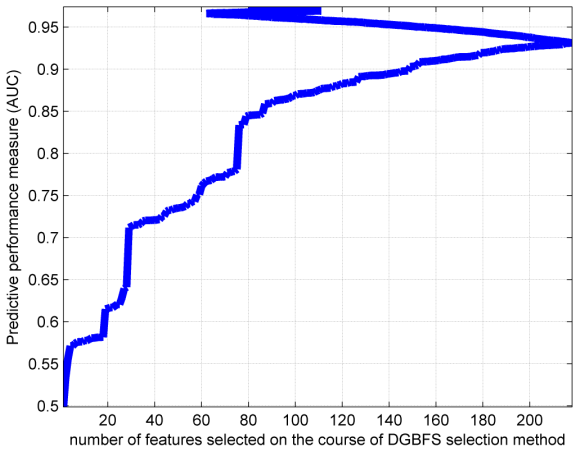


Fig. 4. Diversified greedy backward forward search performance curve

E. Fast Cumulative Sparse Feature Count Search (FCSFCS)

The greedy feature selection method presented in the previous section results in the best predictive performance as defined by eq. 1 that uses optimised Naive Bayes classifier to generate posterior class likelihoods. However, even such simple classifier additionally optimised for processing speed still absorbs non-negligible processing time and, due to the nature of Naive Bayes implementation, may require temporary expansion from the sparse to full representation making it impossible to evaluate the predictive performance with very many features. These issues significantly adversely affect the model scalability and might render its application impossible for larger scale problems, especially in the real-time operation.

To address this issue a further significant simplification is introduced which models the Naive Bayes posterior by just a simple sum of binarised features. We assumed that all the sparse features can be converted into a binary representation that indicate simply a presence of non-zero value. In case of the opposite enumeration of the features, binarisation should be preceded by the value conversion such that binarised "1/true" is always assigned to the sparse class i.e. unlikely set of feature values that has the highest joint probability with the positive target class. Once such binarisation is completed the posterior probability of the positive target class given the features can be simply modelled by the sum of positive binarised feature values which is equivalent to the voting count of true features for each input example. Such sum on binary features is extremely fast to calculate, is fully compliant with sparse feature representation and can swiftly evaluate the models with extremely large feature subsets. What is more it is actually performing very well as a classification method just slightly trailing the Naive Bayes classifier.

The performance of such method has been explored due to its very attractive properties of scalability and speed crucial for applications of huge high-dimensional data for predictive analytics purposes. Using this method allowed to explore normally prohibitive search strategies of greedy backward search (starting from the whole set) and multiple greedy ensemble search that now we managed to carry out in a matter of minutes

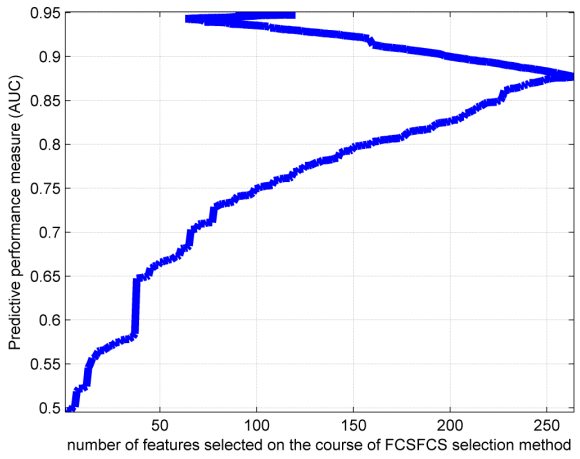


Fig. 5. Fast cumulative sparse feature count performance curve

on a standard PC.

Fig. 5 presents the performance curve achieved by the FCSFCS method using the same penalised AUC of ROC performance measure defined in eq. 1. Although it is just over 1% behind the top DGBFS method, given its simplicity and swift processing taking just minutes it really is a very good and fast feature selection proposition method or in fact a complete very shallow yet robust predictor.

F. Individual vs Ensemble Models

For the evaluation purposes a clarification is required as to the way feature subsets were evaluated. The competition performance criterion defined in eq. 1 clearly enforces the construction of the ensemble of 10 feature subsets with at least 3 features in each subset. The question of whether to construct the ensemble of the subsets of features or use just a single flat subset in fact reflects a long standing dilemma of individual vs ensemble learning. One motivation for ensemble learning is that it is much more efficient and faster to build multiple models with different parts of the feature subset if the computational complexity of the learning process exceeds linear order. This is however not the case for Naive Bayes classifier that given N examples of M-dimensional features is linearly complex in both $O(NM)$. There are also many examples reported in the literature, how a combination of weak learners each built on a small subset of evidence outperforms the single predictor trained on the complete evidence [12]. This effect of ensemble robustness through synergic complementarity is well known and reported on in ensemble learning methods like boosting [13], [12], where the performance gain through combination of weak learners is probably the most exposed.

What we have seen in the context of predominantly greedy feature selection methods is that building the ensemble of feature subsets is much more prone to overfitting. In fact we have built the ensemble versions of the DGBFS and FCSFCS methods where greedy additions or removals were done in turns for all ensemble subsets and the methods terminated when it was not possible to improve the performance for neither addition nor removal from any of the ensemble feature subsets. For all such experiments we have observed a consistent pattern of training performance improved by more than 1% but the validation and testing consistently down by almost 2%. We have also observed a pattern of about 10% to 20% increase in the total number of features selected with the ensemble evaluation method. A possible explanation is that with 10 different feature subsets the ensemble search has many more degrees of freedom and appears to find many new ways to better fit the training data with more features despite the penalty term. In the validation or testing phase, those many unstable coincidences of feature values turn out to be just random noise while the penalty term hits back with the guaranteed decrease of predictive performance.

Therefore throughout the competition the single flat feature subset representation was used and then to meet the solution requirement of being represented in a form of exactly 10 feature subsets a simple yet robust feature distribution method was used. This method exploited the property of the greedy search models which tend to provide the solution in a form of items ordered inline with their quality or contribution to the group performance. What it means is that the features added first and "surviving" in the solution subset tend to be the best while items added last are likely to be individually the worst. To distribute the predictive power of features evenly among the ensemble subsets taking from the top (best) to bottom (worst) the ensemble subsets were appended in turns until all selected features were distributed. As a result the ensemble was composed of different feature subsets that shared similar predictive power and the size difference between the least and most populous subset was at most 1 feature.

VI. Summary of Experiments

The experiments followed the typical competition journey of trying initially simple models, reflecting on the results, and gradually adding more and more complexity in a search for performance improvement. There were many more feature selection methods tested beyond the one reported above. Among the most significant were feature selection with decision trees reported in [10] and the acclaimed fast binary feature selection with conditional mutual information reported in [8]. None of these alternatives came close to the performance achieved by our top DGBFS method.

Fig. 6 illustrates the comparison of performance curves corresponding to different feature selection methods investigated in the paper.

What is striking is how efficient greedy forward search initially was. With just a few features it achieved really impressive performance. However this effect is achieved at the price of really slow processing and in the longer run suboptimal performance caused by the traps of falling into local maximum. The sparse feature voting method was by far the fastest as it effectively eliminated the Naive Bayes classifier, yet still managed to deliver very high predictive performance. The diversified greedy backward forward method performed relatively fast as it absorbed many suboptimal but good features and stabilised with a very robust solution after just few forward and backward rounds. It had reported the best predictive performance and was submitted as proposed solution to the AAIA'14 data mining competition.

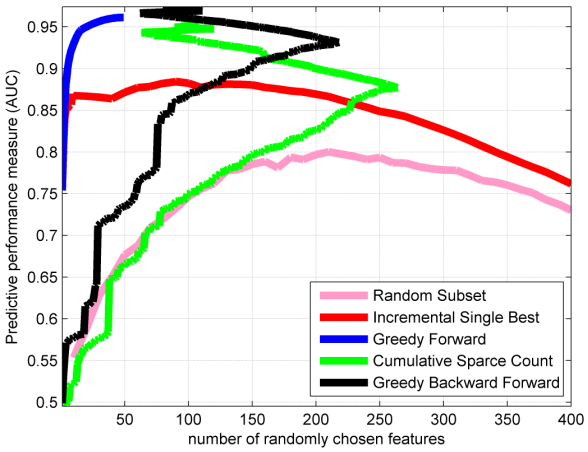


Fig. 6. Feature selection performance curves comparisons

VII. Conclusions

In this work we illustrated relatively simple yet very robust and generic feature selection method that appears to be particularly suitable for handling very large data sets of high-dimensional sparse features. The method employs highly diversified backward-forward search that is relatively fast yet allows to achieve deep features complementarity and very high and stable predictive performance. We have drafted the journey leading to the development of the top model and included informative and comparative examples of other feature selection models some of which could be good candidates for specific predictive requirements. Specifically we have shown that greedy forward search could be a very good model for a very limited number of features, while the binarised features voting method due to its extremely high speed looks to be particularly suitable for live, dynamic predictive system applications with the real-time requirement.

All of the presented feature selection models were considered for an entry in the AAIA'2014 Data Mining competition. Since the predictive performance is the only metric for the competition, the top performing model of diversified greedy backward-forward search has been applied to the data and its solution submitted as our entry in the competition. This solution scored the second place with the tested predictive performance in excess of 96%, just a quarter of the per cent behind the top scored solution. This achievement proves how important is the feature selection step and how much it can reduce the useful input data to provide huge improvements in predictive performance. In the end the model selected only 79 features from the total pool of nearly 12000 thereby elimination more than 99% of data.

The presented model could be used to better understand and prevent various accidents and complex hazardous situations. It really is a good example of how predictive analytics turned big data into small data to potentially save many lives.

References

[1] B. Franks. *Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics,* Wiley, Hoboken, NJ; 2012.

[2] V. Mayer-Schonberger and K. Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think,* John Murray Poblishers, London; 2013.

[3] T. Davenport. *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities,* Harvard Business Review Press, Boston; 2014.

[4] M. Banko and E. Brill. "Scaling to Very Very Large Corpora for Natural Language Disambiguation," In Proceedings. of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001), pp 26–33, 2001.

[5] Y. Bengio. "Learning Deep Architectures for AI," Foundations and Trends in Machine Learning 2(1): 1–127, 2009.

[6] E. Diaz-Aviles, W. Nejdl, L. Drumond and L. Schmidt-Thieme. "Towards real-time collaborative filtering for big fast data," In Proceedings of the 22nd International Conference on World Wide Web companion 2013, pp 779–780, 2013.

[7] L. Yu and H. Liu. "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," In Proceedings of the 20th International Conference on Machine Learning, pp 856–863, 2003.

[8] F. Fleuret and I. Guyon. "Fast Binary Feature Selection with Conditional Mutual Information," Journal of Machine Learning Research 5: 1531–1555, 2004

[9] H. Liu and Lei Yu. "Toward Integrating Feature Selection Algorithms for Classification and Clustering," IEEE Transactions on Knowledge and Data Engineering 17(4): 491–502, 2005.

[10] C. Ratanamahatana and D. Gunopulos. "Feature Selection for the Naive Bayes Classifier Using Decision Trees," Applied Artificial Intelligence 17: 475–487, 2003.

[11] T. Mitchell. "Generative and discriminative classifiers: naive bayes and logistic regression," in Machine Learning, McGraw Hill, 2010.

[12] J.H. Friedman. "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics 29: 1189–1232, 2000.

[13] Z. Zhi-Hua. *Ensemble Methods: Foundations and Algorithms,* Chapman & Hall / CRC Press, Boca Raton, FL; 2012.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 381–385
DOI: 10.15439/2014F498
ACSIS, Vol. 2

Feature Selection for Naive Bayesian Network Ensemble using Evolutionary Algorithms

Adam Zagorecki
Centre for Simulation and Analytics
Cranfield University
Defence Acedemy of the United Kingdom
Shrivenham, UK
Email: a.zagorecki@cranfield.ac.uk

***Abstract*—This document describes the winning method for the AAIA'14 Data Mining Competition: Key risk factors for Polish State Fire Service. The competition challenge was a feature selection problem for a set of three classifiers, each of them in a form of ensemble of naive Bayes classifiers. The method described in this paper uses a genetic algorithm approach to identify an optimal set of variables used by the classifiers. The optimal set of variables is found through a three-stage procedure that involves different settings for the genetic algorithm. The first step leads to reduction of attribute set under consideration from 11,582 to 200 attributes. The following two steps focus on finding an optimal solution by first exploring the solution space and then refining the best solution found in an earlier step.**

I. Introduction

THIS paper describes the winning method for the AAIA'14 Data Mining Competition: Key risk factors for Polish State Fire Service. The challenge was a feature selection problem for a set of three classifiers, each of them defined as ensemble of ten naive Bayes (NB) classifiers sharing the same set of features. The description of the competition and the task can be found at: http://challenge.mimuw.edu.pl/contest/view.php?id=83.

The method described here is based on genetic algorithm (GA) approach. I treated the challenge task as an optimisation problem, where the task was to find an optimal set of variables (it was divided into 10 sets of variables by the competition rules). The set was to optimise the objective function that was based on the score function defined for the competition. In my solutions, the objective function was slightly modified in order to avoid the overfitting phenomenon by using n-fold cross-validation (CV) in the process (with varying n).

The method described in this paper consisted of three different steps. For each step a different GA setup was used. The three steps can be summarised as follows:

- In the first step a small subset of *informative* variables from the original 11,852 variables was identified. For this task a single NB classifier was used rather than an ensemble of NBs.
- In the second step, solutions based on ensemble of NBs were identified, using only the informative variables subset identified in the first step.
- In the third step, the GA was used to improve the best solution obtained in the second step.

II. Genetic Algorithms

Genetic algorithms (GAs) introduced by Holland [1] are a category of evolutionary computation. Evolutionary computation is used to solve mathematical optimisation problems by means of heuristics, and therefore it can be viewed as a meta-heuristic optimisation algorithm. Evolutionary computation is concerned about developing algorithms and techniques that are inspired by the natural evolution. Concepts borrowed from the nature include generations, individuals and populations, genes, mating, natural selection, survival of the fittest, etc.

GAs are the most popular class of evolutionary algorithms [2] that uses *genes* and *chromosomes* to represent the individuals (which correspond to solutions). Genes are basically encoding of the solution by means of a string of numbers (typically binary, with possible other representations). New solutions are generated by means of combining typically pairs of individuals (existing solutions) form the population (working set of solutions). The combination mimics generic crossover with possible additional mutations. It has been believed that suitable chromosome representation can be critical to achieving satisfactory performance of the GA for the given problem. This premise was used in defining approach that I used in the competition.

III. The Challenge

The challenge was to predict three binary decision attributes based on a subset of 11,852 attributes. The challenge imposed a classifier model – for each of the three decision variables an ensemble of exactly 10 naive Bayes (NB) classifiers was to be used, with each NB having at least 3 attribute variables. The same set of NB classifiers was to be used for the three decision attributes.

The performance of the classifiers was determined by means of the receiver operating characteristic (ROC) curve. The goal was to maximise the average area under the ROC curve for the three classifiers. Additionally, a penalty for using a large number of attributes (beyond 30) was introduced. The penalty term p was defined as:

$$p = (\frac{n-30}{1000})^2$$

where n is the number of attributes in the solution. Fig. 1 visualises the effect of penalty as the function of the number

978-83-60810-58-3/$25.00

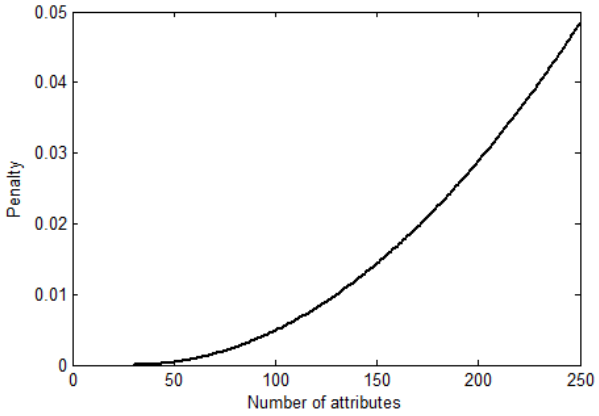


Fig. 1. Penalty as a function of the number of attributes included in the model

of attributes included in the ensemble. Taking into account that the area under the ROC curve should range between 0.5 and 1 and that the baseline solution provided by the competition organisers achieved the score of approximately 0.91, one should conclude that the number of attributes in the ensemble should not be greater than 200 to 300.

The data provided for competition was comprised of 11,582 variables and approximately 50,000 records. For each record a complete set of attributes is available. The distributions of values for the most of variables are heavily unbalanced and that applies to the decision variables as well.

IV. Cross-Validation

One of the basic challenges in data mining is overfitting [3]. To account for overfitting, I used k-fold cross-validation (CV) [4] embodied in the objective function. I used a simple k-fold CV schema, where data records were assigned to appropriate folds using modulo k operator and the original order of the records in the dataset. I used k-1 folds as a training set and the remaining fold as the test set, repeating it for each fold (round robin). The objective function was calculated based on the average of results for k folds.

I experimented with different values for k, as generally there is no widely accepted way to determine a *good* value of k. It is believed, that *good* values depend on the problem and particular dataset. In fact, I varied values for k for different steps. For the first step I used k=10, for the second step I used k=5, while for the third step I used k=3. While the change for the change between the first and second step was dictated just by the desire to avoid artefacts coming from the k selection, the reason for the change to k=3 was a result of simple investigation I decided to do after the second step: I evaluated performance of the same algorithm run with different CV settings: n=3,4,5,7, and 13 using the competition submission website that allowed to get feedback based on the evaluation set. I noticed that the results for n=3 and n=4 produced the same solution that scored better on the competition leaderboard than the solutions achieved using greater values of n. Therefore, for the final step I decided to use n=3.

V. The Method

In this section I will describe the method to find solutions to the competition challenge. The approach taken consisted of three steps that varied in the details of the experimental setup. The idea was to in the first step to identify attributes that can lead to good solutions. For the following steps I decided to use only a subset of attributes to improve the performance of the search heuristic.

My approach was dictated by two observations: (1) initial data analysis indicated that there are many variables that are in *present* state very rarely (and therefore arguably are not contributing much information or leading to overfitting) and (2) that initial experiments suggested that the *good* solutions quite consistently preferred certain variables over others. Based on those observations I decided first to narrow a subset of candidate attributes to 200.

In the three steps the same simple GA was used. However the chromosome and operator definitions, method of constructing initial population and the population size, and other parameters were varied throughout the steps. Below I provide the details of the algorithm for each step.

VI. The First Step – Identifying Informative Attributes

In the first step, the goal was to reduce an attribute set by identifying a subset of attributes that was *most informative*. In order to do that, I used a GA that would identify a NB (note: a single NB, not an ensemble of ten NBs) that otherwise would be optimising the problem as stated in the competition.

A. Chromosome Definition

Chromosome was defined as a list of integer values that were allowed to take values from 1 to 11,852. The length of chromosome was constrained to be at least 30, with upper limit set to 250 (but effectively solutions never exceeded 200 attributes anyway).

B. Crossover

As the crossover operator I used the uniform crossover method with mixing probability 0.5.

C. Mutation

Mutation had two operators, each of them applied with 0.5 probability:

- Remove – in that case the attribute would be removed from the list of attributes for that NB
- Replace – it would replace an attribute with a random attribute sampled from the uniform distribution. The addition of attributes was achieved by replacing an *empty* (denoted as 0) gene with a value related to one of the attributes (denoted as an integer 1 to 11,852).

The probability of mutation was set initially to the value 0.01 and was designed to decline after each generation. After

each iteration (generation), the probability of mutation was multiplied by a constant delta equal to 0.999. The parameters were set in such a way that a mutation would occur for a new individual with probability over 0.9.

D. Initial Population

Initial population consisted of 200 individuals. The initial number of attributes for individuals from the initial population was sampled from the uniform distribution with lower and upper limit 30 and 130 correspondingly. The attributes were sampled from the complete set of 11,582 attributes using the uniform distribution. The size of the initial population of 200 individuals was determined based on the size of the problem. The total number of possible attributes in the population (200 multiplied by 80, where 80 is an average chromosome size in the initial population) would exceed the total number of attributes in the data (11,852).

E. Simulations

In order to identify the *informative* attributes I run 200 independent simulations and collected the best result (an individual with the highest score at the end of simulations) for each of the 200 simulations. Each simulation terminated after 2500 generations. For each generation I generated 50 new individuals (25%) and subsequently rejected 50 individuals with the lowest score.

To perform simulations I used my own implementation of the algorithm written in C++ programming language. The implementation was intended to provide optimised code for the task. Rather than using the raw data, I used pre-calculated conditional probability tables for each attribute (given decision attribute and CV fold).

F. Result

As the result of simulations I obtained a set of 200 solutions. Please note that those solutions assumed a single NB model, rather than an ensemble of 10 NB models. To evaluate those solutions using the competition leaderboard I used a simple technique: I assumed that the first 27 attributes will be assigned for the 9 NBs in the ensemble with each contacting only 3 attributes (the minimum required). The remaining attributes were assumed to belong to the last NB in the ensemble. Additionally I sorted the attributes for the evaluation purpose. Solutions obtained using this method resulted in scores in the range of 0.93 to 0.945 on the competition leaderboard.

But the most valuable result for this step was the set of *informative* variables. It turned out that the 200 solutions shared a lot of common attributes, suggesting that there were clearly attributes that were more informative than the others.

The attributes regarded *informative* as were (sorted according to the ranking): 5270, 143, 2887, 8179, 10062, 460, 2924, 8914, 258, 2182, 7187, 7980, 5306, 1880, 7999, 11266, 1509, 11463, 7299, 5909, 3273, 5835, 1244, 72, 142, 8985, 6961, 10114, 6660, 2835, 7055, 8959, 304, 7148, 8039, 8107, 6779, 10880, 5446, 11359, 3294, 3492, 3951, 4323, 11459, 1335, 2684, 401, 4347, 3257, 7755, 8990, 675, 5519, 1772, 6949, 2883, 6196, 766, 6238, 9846, 6534, 5990, 5996, 7425, 10446, 1111, 3598, 3735, 8110, 8114, 9873, 691, 759, 10902, 2380, 3324, 4415, 8249, 40, 2971, 4959, 9926, 8972, 2635, 4962, 11802, 8772, 3290, 5572, 6163, 8529, 3402, 9715, 10019, 27, 6430, 7172, 8216, 8794, 9657, 9771, 1161, 7243, 7520, 10771, 271, 1540, 4153, 4469, 5655, 8034, 8853, 11196, 3166, 5229, 5350, 8159, 9848, 11701, 73, 799, 3890, 5861, 6474, 6653, 480, 2966, 3706, 6388, 11217, 11231, 327, 896, 3479, 5393, 5778, 9725, 8397, 648, 840, 1954, 2538, 7158, 10300, 11604, 833, 2697, 4594, 7046, 8314, 10740, 11088, 2273, 2402, 2868, 4137, 6162, 7304, 8553, 10014, 11483, 2539, 5262, 6750, 10638, 11421, 135, 220, 1307, 3751, 4443, 6222, 6498, 6507, 7103, 7414, 8093, 8663, 9055, 11594, 282, 1289, 1767, 2568, 4198, 5041, 5143, 5859, 6579, 6602, 7069, 7356, 8807, 11587, 11636, 364, 426, 958, 2218. I arbitrarily decided to use 200 attributes for the further experiments. The 200th attribute was close to 0.05 probability of occurring in the solution.

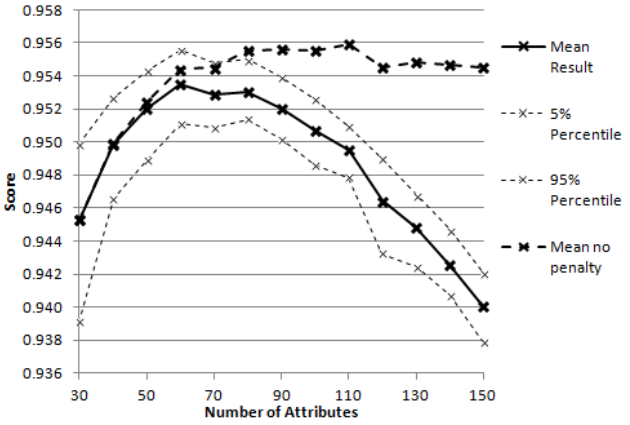


Fig. 2. Performance as a function of the number of attributes

As well, at this step I determined that the optimal number of parameters ranged between 60 and 80 (assuming single NB, not an ensemble). It should be noted that the definition of the chromosome and the simulation allowed the algorithm for adjusting the number of attributes in the solution, and therefore I did not need to be concerned about specifying the right number of the attributes in the solution.

In fact I run the experiments for which I used constant chromosome size, in order to ensure that the results I get are consistent with the fixed chromosome size experiments. Those experiments shed some light on the problem of the optimal number of attributes in the solution. The results are shown in Fig. 2. This analysis conformed that the optimal number of attributes should be approximately in range of 60 to 80 attributes if the penalty term is included.

VII. The Second Step – Finding Optimal Ensemble

In the second step, the task was to come up with a *good* solution for NB ensemble.

A. Chromosome Definition

For this step, the chromosome encoded the structure of the classifier ensemble, as stated in the competition requirements. The chromosome was defined as a vector of 10 lists of integers, where each list corresponded to one NB in the ensemble. The list was constrained to have at least three elements, with no upper limit on the number of elements.

B. Crossover

For both crossover and mutation, I treated each list of 10 chromosomes as individual chromosomes, applying crossover to i-th chromosome from the first individual with the i-th chromosome form the other individual. I indeed tried to crossover i-th chromosome from the first individual with j-th chromosome from the second individual, but performance of such algorithm was inferior to the one used, therefore I rejected the idea.

This time I used single point cross-over with the point randomly generated from the uniform distribution.

C. Mutation

The probability of mutation for a single NB classifier definition (repeated for each list in the vector) was set initially to be 0.05, and it was decreased at each generation with factor of 0.999. The mutation algorithm used one of three mutation operators applied with different probabilities:

- Add – adding a new element to the list (sampled from 200 attributes using uniform distribution), applied with probability 25%
- Remove – remove an attribute from the list (if there are more than 3 attributes in the list), applied with probability 25%
- Replace – replace an existing attribute with a randomly selected attribute (sampled from 200 attributes using uniform distribution), applied with probability 50%

D. Initial Population

Initial population consisted of 100 individuals. This time the chromosomes were randomly initialised using 200 attributes identified in the first step. The initial chromosome size was assumed to be 3 for all individual lists within a chromosome (hence started with 30 attributes in each solution).

E. Simulations

Each simulation included 5000 generations. For each generation I generated 25 new individuals (25%) and rejected 25 individuals with the lowest score. I run over 50 of simulations and collected the final best results for each of the simulation runs.

F. Results

The best solution identified at this step allowed me to achieve a score of 0.9512 at the competition leaderboard.

VIII. Step Three – Improving the Score

The final step was inspired by the idea of using evolutionary algorithms that alter the initial solution in order to improve it. But rather than using any specific approach, I decided to us the same GA framework I used previously, with the only difference that I decided to focus on the mutation operator as the search driving mechanism. The basic setup was the same as the algorithm in the second step, with some changes described below.

A. Chromosome Definition and Corssover

Chromosome definition and crossover operators were exactly the same as in the second step.

B. Mutation

The mutation operator was exactly the same as in the second step. The only difference was that the initial probability of mutation for each list element (individual NB in an ensemble) was increased to 0.15 to induce more mutations.

C. Initial Population

The key change was the initial population – this time I used seeding. I decided to use initial population that comprised of 50 copies of the same individual – the one that was achieving the highest score in the second step.

D. Simulations

Each simulation had 5000 generations, but in fact I used several simulations:

- I terminated the first simulation at around 500 generations, and I achieved the leaderboard score of 0.9522.
- I used the best result achieved so far as the seed for the initial population, and this time I allowed to run the complete 5000 generations. That allowed me to achieve the leaderboard result of 0.9561.
- Consequently, I used the best result so far to repeat the procedure. This time I decided to use several parallel runs with the same seeding. Most of the results were inferior to the initial (seeding) solution (0.9561) with most of leaderboard submissions ranging between 0.954 and 0.946. One outlier was the solution with the leaderboard result 0.9583 that I used at the final submission.

The solution that I used for the final submission was as follows:

- 11463, 11088, 8179, 6498, 2883, 460, 143, 6388, 2924
- 10880, 4415, 1880, 258, 2966, 5926, 6491
- 10446, 9771, 8249, 7980, 7187, 5270, 401, 5990, 8039, 8959, 8093, 3890
- 833, 72, 3751, 1266
- 8914, 7999, 6474, 5041, 4962, 10019, 947
- 10902, 5996, 2835, 748
- 8034, 7055, 6961, 5909, 1244, 27, 3324, 7172, 270
- 11701, 8110, 7356, 7187, 6222, 6162, 7158, 9873, 7560
- 11266, 10062, 220, 11231, 2971, 7521
- 11459, 7148, 7755, 3951, 2887, 766.

IX. Conclusion

In this paper I presented the description of the method used to optimise the set of variables used for classifiers based on ensembles of NB. I used the same GA framework, however in three different ways to address the challenge problem:

- First, I used GA to reduce the number of attributes in the problem in order to improve performance of the consequent simulations
- Then I used GA to find a set of optimal solutions to the competition problem,
- Finally, I used GA with focus on mutation to refine the solutions obtained from the previous step.

I would like to emphasise that I strongly believe that the winning solution I obtained can be improved. Similar applies to the algorithms used – they can be refined and improved in terms of convergence efficiency, chromosome encoding, etc.

Acknowledgment

I would like to thank my wife, Katarzyna Holownia, for encouragement and support during the competition.

References

[1] J. H. Holland, “Adaptation in natural and artificial systems”, Ann Arbor: The University of Michigan Press, 1975
[2] M. Mitchell, “An Introduction to Genetic Algorithms”, MIT Press, 1998
[3] T. Dietterich, Overfitting and undercomputing in machine learning, ACM Comput. Surv. 27, (3), 326-327, 1995
[4] S. Geisser, “The predictive sample reuse method with applications”, J. Amer. Statist. Assoc., 70:320–328, 1975

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 387–394
DOI: 10.15439/2014F500
ACSIS, Vol. 2

Feature selection and allocation to diverse subsets for multi-label learning problems with large datasets

Eftim Zdravevski
Faculty of Computer Science and Engineering
Ss.Cyril and Methodius University, Skopje, Macedonia
Email: eftim.zdravevski@finki.ukim.mk

Petre Lameski
Faculty of Computer Science and Engineering
Ss.Cyril and Methodius University, Skopje, Macedonia
Email: petre.lameski@finki.ukim.mk

Andrea Kulakov
Faculty of Computer Science and Engineering
Ss.Cyril and Methodius University, Skopje, Macedonia
Email: andrea.kulakov@finki.ukim.mk

Dejan Gjorgjevikj
Faculty of Computer Science and Engineering
Ss.Cyril and Methodius University, Skopje, Macedonia
Email: dejan.gjorgjevikj@finki.ukim.mk

***Abstract*—Feature selection is important phase in machine learning and in the case of multi-label classification, it can be considerably challenging. In like manner, finding the best subset of good features is involved and difficult when the dataset has significantly large number of features (more than a thousand). In this paper we address the problem of feature selection for multi-label classification with large number of features. The proposed method is a hybrid of two phases - preliminary feature selection based on the information value and additional correlation-based selection. We show how with the first phase we can do preliminary selection of features from tens of thousands to couple of hundred, and then with the second phase we can make fine-grained feature selection with more sophisticated but computationally intensive methods. Finally, we analyze the ways of allocating the selected features to diverse subsets, which are suitable for training of ensembles of classifiers.**

I. Introduction

MACHINE LEARNING provide means to automatically analyze enormous quantities of data and consequently to: derive various conclusions, make predictions for unseen data, find patterns within the data etc. As learning relies on the available data, its preprocessing is very important to such extent that most of the time of the project might be spent for this phase. During data processing various issues of the data can be addressed: feature modeling and construction [1] [2], outliers removal [3], noise detection and reduction [4], missing values imputation [5] [6], data normalization [7] [8], and data transformation [9] [10].

Many learning algorithms such as neural networks [11], Naive Bayes [12] [13], decision trees [14] notably experience degrading performance when the datasets contain redundant or irrelevant features. This phenomenon is confirmed with theoretical and empirical evidence in plenty of research papers, some of which are [15] [16] and [17]. The problem of feature selection [18] [16] [19] can be defined as the task of selection of subset features that describe the hypothesis at least as well as the original set. The representation of data instances is optimized with feature selection, which in turn can lead to:

- Improving the performance of learning algorithms.
- Reducing the training and execution times of algorithms.
- Improving the memory requirements and allow application of more algorithms.
- Improved robustness to over-fitting.
- Better understanding and visualization of the data.

Different methods for feature selection focus on various aspects of the above goals, or achieve the same goals but in different ways. In [20] are given guidelines for feature selection and are introduced the most widely used methods. It is important to note that finding the most useful and relevant features is not always the same task, as it is shown in [16] and [21].

II. Related work

There are two approaches for feature selection: filter and wrapper approach. The filtering approaches rank the features based on some metric. These methods are generally characterized by simplicity, scalability and solid empirical background. Because they rely on relatively simple metrics, they are memory and computationally efficient and can be applied on datasets with tens or even hundreds of thousands of features. Such application of these methods, as well as their empirical analysis, is further elaborated in [22], [23] and [24]. Filter methods are independent of the machine learning algorithm that is going to be applied later on.

Filter approaches for feature selection can further be categorized into two groups. The first group consists of methods that rank the features based on some measure of their individual predictive power: information value [25] [26] [27], information gain [28] [29], information gain ratio [28] [29], RELIEF [30] [31], entropy [32] etc. In [33] and [34] are described some filtering methods based on posterior probability. The common problem of all methods in this group is that they take into consideration only the individual usefulness of attributes in relation to the target classification and can not discover

This work was partially financed by the Faculty of Computer Science and Engineering at the Ss.Cyril and Methodius University, Skopje, Macedonia

redundancy, multicollinearity or interdependence between the chosen features.

The second type of filter approaches consists of methods which analyze the subset of features based on some metric that describes the performance of the whole subset and not only the individual features [1]. Namely, the correlation-based approaches described in [35] and [36] fall into this type of methods. Important to realize is that they search for subsets of features that have low inter-correlation between them and high correlation to the target classification [37]. Likewise, [38] proposes an approach for detecting stable clusters of features based on principal component analysis.

The wrapper approaches search for subsets of features that are useful for the classification or regression task at hand. They are based on evaluating the performance of different subsets of features using a machine learning algorithm [21] [17] [39]. When applying these methods the individual contribution of features is not being evaluated. In contrast, the contribution of the subset of features is taken into consideration and the whole process is black-box like. In other words, the method does not give exact information why that specific subset of features was selected. In order to apply a particular wrapper method, one has to define: how will be the space of all possible feature subsets traversed; how will be the performance of the learning algorithm evaluated in order to guide the search; and which learning algorithm to be used. If the number of features is small, then all combination of features can be evaluated, but this is rarely the case. The main problem of these methods is their computational complexity. Be that as it may, there are a lot of search techniques that mitigate this problem [19]. On the other hand, the main advantages of these methods is their universality and independence of the domain of the data and task. The research community has proposed various ways of making hybrid methods that combine filter and wrapper and [40] reviews them.

Our research presented in this paper focuses on feature selection in areas of application where datasets have tens or hundreds of thousands of variables. These areas include text processing, gene expression array analysis, and combinatorial chemistry. This paper is organized as follows: Section III describes the problem at hand and section IV gives overview to the proposed solution. In subsections IV-A and IV-B we describe the proposed hybrid approach for feature selection. Subsection IV-C presents various schemes for constructing diverse subsets of features that are suitable for ensembles of classifiers. In Section V we summarize our work.

III. Problem definition

This paper originated from our research during and after the AAIA'14 Data Mining Competition "Key risk factors for Polish State Fire Service". This competition is organized within the framework of the 9th International Symposium on Advances in Artificial Intelligence and Applications [41], and is an integral part of the 1st Complex Events and Information Modeling workshop devoted to the fire protection engineering. The task is related to the problem of extracting useful knowledge from incident reports obtained from The State Fire Service of Poland. With this in mind, our research goals were mainly guided within the task goals and requirements. Under those circumstances, during the following sections we will occasionally relate to some specifics for this task. Nevertheless, the proposed methods are not specific for this task and they can be applied to a variety of problems.

The organizers obtained a data set containing nearly 260000 reports describing the actions carried out by the Polish State Fire Service within the city of Warsaw and its surroundings in years 1992 - 2011. Each report consists of two parts. The first one contains a summary of resources utilized during the action in a form of structured and quantified characteristics. The second part contains a natural language description of the reported events, which is entered by the officer coordinating the action. They have preprocessed a subset of the reports and transformed it into a table in which each of the reports is described by almost 12000 attributes. The training dataset contains about 50000 instances. Additionally, they have distinguished 3 target attributes that correspond to information whether in the described incident there were casualties among firefighters, children or other involved people, respectively. The goal of the competition is participants to come up with solutions which will improve the understanding of the risk factors associated various types of accidents. Given these points, it seems that the problem is actually multi-label classification. As a matter of fact, after careful review of the training data we have observed that some instances (i.e. reports) are indeed classified to the positive classes in more than 1 of the decision attributes. The organizers have modeled the decision attributes in a way that actually transforms the multi-label problem into 3 binary classification problems. Such approach for tackling multi-label problems is, in essence, problem transformation method and is described in [42].

The task in this competition was to identify attributes that can be used to robustly assign the reports to the corresponding decisions labels. In particular, organizers decided to use ensemble of 10 Naive Bayes classifiers for each of the target classifications. Having 3 decision attributes, means that the selected features should be divided into 10 subsets and each subset should be used to train 3 individual Naive Bayes models. Every model assigns scores (i.e. probabilities) to test cases representing if that the case should be classified to the positive decision class or not. In this way, for every decision attribute and every test case there are 10 scores. The ensemble of predictions is constructed by taking the sum of the scores of the individual models.

The metric used to evaluate the performance of the selected attributes was the average AUC of the prediction ensemble for different decision attributes, decreased by a penalty for using a large number of attributes. We assume that the choice of metric is because the data is highly imbalanced and many papers confirm that this metric is best suitable for such cases [43] [44] [45].

Namely, if we denote by: s - submitted solution; $|s|$ a total number of attributes used in the solution (with repetitions); and

$AUC_i(s)$ Area Under the ROC Curve (AUC) of a classifier ensemble for the i-th decision attribute, then the quality measure used for the assessment of submissions can be expressed as:

$$score(s) = F\left(\frac{1}{3}\sum_{i=1}^{3} AUC_i(s) - penalty(s)\right)$$

where the penalty is equal to:

$$penalty(s) = \left(\frac{|s| - 30}{1000}\right)^2$$

and the function F:

$$F(x) = \begin{cases} x, \text{ for } x > 0 \\ 0, \text{ othewise} \end{cases}$$

From all the given task description and stated requirements the following challenges should be acknowledged:

- Evaluation of the usefulness of features in relation to the 3 target classifications.
- Selecting a small subset of features that will be contributing to all 3 target classifications.
- Optimal arrangement of the selected features in N subsets (N=10 in this case) in order to train ensemble of classifiers.

In order to overcome those challenges we propose a hybrid method which is described in the following section.

IV. Proposed method

Selecting the best subsets of features for this dataset is a challenging task because most of the feature selection algorithms cannot be applied due to the large number of features. Additionally some of the methods for feature selection are applicable only on binary classification problems. With this in mind and given that the task at hand has 3 decision attributes, the selection of features that are contributing to the 3 classification tasks at the same time gets even more difficult.

We propose a hybrid approach for feature selection consisting of three phases. The *first phase* performs preliminary feature selection in order to discard the features that are unlikely to contribute to any of the decision classes. The *second phase* applies more sophisticated feature selection algorithms on the dataset that after the first phase has significantly smaller number of features. As a result from the second phase the set of selected features is very concise and all of them contribute to the 3 classification tasks. If the goal was to create 1 model for each of the classification tasks, then we would use the selected features and we use some learning algorithm to build the models. In such case the feature selection would end here. Be that as it may, the contest rules described in III state that the goal is to train an ensemble of Naive Bayes classifiers. Having this in mind, we need a *third phase* that would optimally arrange the chosen features into subsets that will be later used by each individual classifier. We realize that it was not specifically forbidden to use one feature in more than 1 subset. Although this may be allowed, we believe that such approach is problem-specific and would require a significant effort for fine tuning, to the extent that the scientific contribution of the approach would diminish. For this reason we have decided to use diverse subsets of features for each individual classifier. In other words, each selected feature belongs to only 1 subset. The following subsections describe each of the phases in our approach.

A. Preliminary feature selection

The large number of feature in the original dataset presents a difficult task for most methods. The reason for this is because of the memory and/or computational complexity it imposes. The goal of this task is to overcome that problem by reducing the features to a significantly smaller number using some simple algorithm. Being able to do this clears the way for more sophisticated feature selection methods. As it was explained in section I, the prime candidates for a fast preliminary (i.e. coarse-grained) feature selection are the filter methods that assess that individual contribution of features. The following metrics can be used for feature selection are less demanding in terms of memory and computational power: information gain, information gain ratio [28] [29] and information value [25] [26] [27]. In spite of the slight differences between them in terms of computational complexity time, all of them can be computed in linear time ($O(mn)$) with 1 pass of the training dataset. We were not able to obtain results with the RELIEF method [30] [31] in reasonable time due to its higher complexity - $O(mnp)$. Here where m is number of training instances, n is the number of attributes and p is the number of randomly selected instances used for the RELIEF algorithm. We acknowledge that with proper tuning of the p parameter we might have been able to obtain results with it too, but since this phase performs only preliminary selection we believe that this is not worth the effort.

We have decided to use the information value for estimation of the predictive power of each of the features in relation to each of the decision attributes. It is widely adopted in industry especially for credit scoring problems [25] [26] [27]. The reason for this is because there are some widely adopted rules of thumb in terms that give simple guidelines of whether the feature is strong or weak predictor based on the information value. However, note that weak features may provide value in combination with others; or have individual values that could provide predictive power as dummy variables. As it has been suggested in [46], the following guidelines for evaluating the strength a predictor based on the information value can be used:

- Less than 0.02: unpredictive
- 0.02 to 0.1: weak
- 0.1 to 0.3: medium
- Greater than 0.3: strong

Although they are firmly grounded in good practice, how these guidelines be related to other metrics is discussed in [47]. At the same time, there are some drawbacks of this metric related to the some border cases that prevent using its original

definition (2). In [10] are proposed some enhancements of the weight of evidence (WoE) parameter, which in turn overcome the computational obstacles for the information value. With (1) is defined WoE, and it is further used for calculation of the information value (2). Here N_i^j and P_i^j represents the number of negative and positive instances labeled with the i-th value of the j-th feature, respectively. Also SN and SP denote the total number of negative instances and the total number of positive instances in the training dataset, respectively.

$$WoE_i^j = ln\left(\frac{N_i^j}{P_i^j}\right) - ln\left(\frac{SN}{SP}\right) \quad (1)$$

$$IV^j = \sum_{i=1}^{n}\left[\left(\frac{N_i^j}{SN} - ln\frac{P_i^j}{SP}\right) \times WoE_i^j\right] \quad (2)$$

From (2) it is obvious that the information value is applicable only to binary classification problems. When having multi-label classification with k possible positive labels, one needs to compute k information values for each of the features. For this specific task, having 3 decision attributes and almost 12000 features, means that we had to compute nearly 36000 information values. The computation for all of them takes less than 15 minutes on a regular laptop. As it turns out, some of the features are strong or medium predictors for one decision attribute, but are very weak predictors for the other one or two target attributes. The next challenge was how to aggregate the 3 information values of each feature into 1 value, so we can use it for feature selection. The following subsections describe the results of each aggregation type.

1) Average information value: When having multi-label classification with k positive classes we can average the k information values of each feature to get estimate of the its information value in relation to all positive classes. For this case in particular, we have tried averaging the 3 information values of each feature in order to use it for feature selection. We have examined various subsets containing 50 to 120 of the best features based on their average information value. By training ensembles of Naive Bayes classifiers as described in III we have obtained AUC performance on the leader-board dataset varying from 0.886 to 0.9, based on the number of features and the scheme of arrangement of them into diverse subsets. As a reference, the best results of the same test dataset were up to 0.94. The experiments showed that the performance of the ensembles build on the these selected features were fairly stable. However, they were worse than what we hoped to be achieved with more sophisticated methods. Nevertheless, it was notable that the average information value can be safely used for preliminary feature selection.

2) Maximum information value: The next obvious idea for aggregating the individual information values of a feature is to calculate their maximum. When we applied this logic on the current dataset and we selected the best different subsets containing 50 to 120 features based on their maximum information value, the performance of the ensembles was worse than with the approach in IV-A2. In fact, the AUC performance on the leader-board dataset was less than 0.8, regardless of the arrangement of the features in subsets. By looking into the selected features and their maximum information values we can explain this phenomenon. As it can be observed on Fig. 1, some features might have high information value for one of the decision attributes, but low for the other decision attributes. This in turn, translates to high maximum but low average information value.

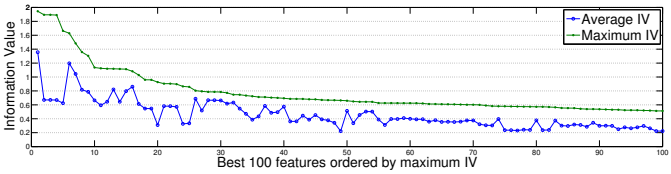


Fig. 1. Maximum vs. average information value of best 100 features

To conclude, this aggregation might be bad-performing for final feature selection, but one should not rush to avoid it. In fact, this aggregation may identify features for training models for each of the decision attributes separately and may also identify features that in combination with others might be very useful. The maximum information value is very useful for discarding features because it guarantees that the discarded ones are bad or week predictors for all decision attributes (positive labels).

3) Weighted average information value: Another approach for aggregating the individual information values of a feature is to calculate their weighted arithmetic average. The weight can be calculated based on the number of positive instances in the training dataset for each decision attribute (i.e. label). This idea has been applied for weighting averages of other statistics [48] [49]. However, for final feature selection wighted average type of aggregation does not seem to be suited, mostly because the most common positive label is always preferred. On the leader-board dataset the AUC of different subsets varying from 50 to 120 features was about 0.86. Maybe with a different weighting scheme, the performance would be improved. However, as this phase should not select final subset of features, we did not investigate other weighting schemes.

4) Dependent features: It has been extensively proved that features with high correlation have negative impact on performance for many machine-learning algorithms, among which is the Naive Bayes classifier. Some such papers are mentioned in section I. In order to address this issue, we have calculated the correlation coefficients and p values [50] between all features and used this information to find dependent features. By discarding a feature if we have already selected a dependent feature with higher aggregate information value, we were able to slightly improve the performance of the maximum and weighted average aggregations. In those cases, the performance was similar to the average aggregation. Despite that, the obtained results were not satisfactory for final feature selection.

5) Coarse-grained selection of features: We have applied the three aggregation methods for the information values and

then we selected sets of the best N features (where N is 400, 500, 600, 800, 1000, and 1500). By analyzing the sets obtained for different values for N we have concluded that for the same value of N but different aggregation methods most of the features (i.e. 70-90 %) are overlapping. That indicated the aggregation type will not have significant impact on the coarse-grained selection of features, except for their ranking. In order to decide how many features to select during this phase we have analyzed the max-aggregated information values and have noticed that after the best 500 features the maximum information value drops bellow 0.1, meaning that the discarded features are week predictors for all decision labels. The maximum information value is very useful for discarding features because it guarantees that the discarded ones are bad or week predictors for all decision attributes (positive labels).

Finally, as a result from this phase we have selected the best 500 features based on the maximum information value and we continued with the next phase to apply more sophisticated feature selection algorithms on the significantly simplified training dataset.

B. Fine-grained feature selection

After phase 1, the training dataset is significantly simplified. More specifically for the current dataset, we have 50000 instances and 500 features. It is significantly reduced than the original, so more intelligent feature selections algorithms can now be applied.

The number of features in a dataset should indicate whether it is possible to use wrapper methods for fine-grained feature selection. The experiments performed in [51] show that wrapper methods can be applied to relatively smaller datasets (containing less than 200 features and few thousand instances). In spite of the continued improvement in processing power during the recent years, trying out many combination of features especially with more complex learning algorithms is a very hard task. In this case, the reduced size of dataset is still quite large in order to be able to apply a wrapper method for feature selection on it. This fact limited the use of wrapper method to very simple internal learning algorithms. Again, having too many combinations of feature subsets still makes wrapper methods not adequate for this task.

One of the best performing methods for evaluation of subsets of features is the correlation-based feature subset selection [35]. It evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low inter-correlation are preferred.

If we were to apply this method on the 3 decision attributes separately, we would still need to aggregate the 3 selected subsets of features. As it was shown in subsection IV-A, this task can be very involved. Instead of doing that, we have decided to transform the problem space. The original task is multi-label classification which was transformed by the organizers to 3 separate binary classification problems. In cases like this, we propose to merge the separate binary classification problems into 1 multi-class problem. To summarize, starting with a multi-label problem transformed as several separate binary problems we merge it to a multi-class problem. By doing this we can apply feature selection methods that select the best features in relation to all positive classes. More particularly, with the proposed transformation we obtained 8-class classification problem by using the following Eq. (3) to map each instance to a new artificial class. Here, AL_i denotes the artificial label of the multi-class problem instance i, where as L_i^1,L_i^2 and L_i^3 are the classes in the binary classifications of the same instance.

$$AL_i = 1 \times L_i^1 + 2 \times L_i^2 + 4 \times L_i^3 \tag{3}$$

In general, multi-label classification tasks where the number of positive labels is N, can be transformed to N binary classification problems [42]. Let the label of the i-th instance in the $j-th$ binary problem is L_i^j, where for L_i^j is 0 for negative instances and 1 for positive instances. With this transformation to multi-class problem the same instance will be labeled with AL_i as defined in Eq. (4):

$$AL_i = \sum_{j=1}^{N} 2^{j-1} \times L_i^j \tag{4}$$

After performing this transformation, the correlation-based feature subset selection can be applied. Depending on how many features are in the training dataset and how they are chosen (i.e. which aggregation was used), this method selects from 40 to 70 features. Considering the obtained attributes we have observed that one particular subset of 53 features was very common, henceforth the next phase was performed using that subset (shown on Table I).

C. Allocation of features into diverse subsets

After end of phase 2 we have a very concise dataset which, in this case, is described with 53 features. The correlation-based method for feature selection [35] does not rank the features, but we can rank them computed based on the information value calculated during phase 1. They can be ranked based on their maximum or average information value.

The goal of this phase is to optimally allocate the selected features into diverse subsets. For this task the number of subsets is set at 10, but in general, one can try various number of individual classifiers for the ensemble. Each subset should contain approximately equal number of features. The following subsections describe the schemes for allocation of features into subsets. Before we continue, let us define an *iteration* as allocating 1 feature to each subset (e.g. in this case choosing 10 features, 1 for each subset). The different schemes explained bellow, have different logic of choosing the next feature to allocate to a subset. If we consider the subsets as items that are ordered, we can decide which of them will get processed first. By being processed we mean allocating a feature to it. Likewise, the features are ordered by their maximum information value.

1) FIFO scheme: The First-In-First-Out (i.e. First-Come-First-Served) term has been widely used in data structures literature and queue theory. During 1 iteration the FIFO scheme would allocate the next best feature to the next subset. The following iteration will allocate features starting from the first subset and so on until there are no more features. Obviously, this scheme mostly favors the first subset and least favors the last. Using the leader-board dataset, we have obtained AUC of 0.9292. To summarize, this approach uses the maximum information values for ranking the features. When we used the average information value for ranking, the performance was slightly worse. The simple explanation for this is because the average information value is more consistent than the maximum, hence the FIFO scheme favors the first subsets more.

2) FIFO-independence scheme: In order to improve the FIFO scheme, we can dependent features in order to optimally allocate the features into subsets. The idea is to have independent features within 1 subset. The algorithm used during phase 2 is correlation based which ensures that the selected features have very low inter-correlation among them. However, if we use a more strict test for independence (p value = 0.01), then we can still find some pairs of dependent features. This improved FIFO scheme selects the next best feature that is independent to all features that are already in the subset. As it turns out, when using the leader-board dataset, this scheme slightly improved the performance to AUC of 0.9293.

3) Interchanging FIFO-FILO scheme: With this scheme in each iteration we change the logic from FIFO to FILO and vice versa. So, when assigning the first attribute to the first subset we choose the best feature, then for the second subset we choose the next best feature and so on, until the last subset has 1 feature. Then when assigning a second feature to all subsets, we start with the last subset and assign the best available feature to it. In like manner, we continue assigning the next best feature until the first subset has 2 features. In the next iteration the first subset will have priority, and so on until we run out of features. With this scheme the AUC ROC performance on the leader-board dataset was 0.9298, which was an additional improvement.

4) Monte Carlo scheme: This scheme randomly scatters the features to subsets. It is the simplest scheme and produced results ranging from 0.926 to 0.9321. We have analyzed the final distributions to subsets that produced better and the ones that produced worse results. When looking at the information values for all three target attributes of the features in each subset it was notable that the better performing arrangements of subsets had features that are medium or strong predictors in relation to 1 or 2 target attributes and weak predictors for the other target attribute.

We have concluded that this scheme might produce very good results, but in order to be consistent it needs to be improved. One way of doing this is to use this scheme as a starting point and later to make rearrangements by swapping some features between the subsets. Choosing which features to swap is based on the following logic:

- We first find a bad performing subset of features and determine which target attribute has least weak features in relation to it.
- Then find a subset where a lot of features are medium or strong predictors for the same target attribute.
- Swap the features from the 2 subsets.
- Repeat the process until no swaps can be made.

This algorithm generally helps both subsets. The first subset will get a stronger feature for the class that has bad performance. The second subset is also improved because the possibility of over-fitting because of too many strong predictors for it for particular target attribute is reduced. Using this technique we have finally arrived at the feature arrangement shown on Table I on the following page.

V. Conclusion

In this paper we have proposed a three-phase hybrid feature selection method that is able to extract features from datasets with thousands of features. This method is especially useful for datasets that originate from text processing areas of application. Additionally we have analyzed the different ways to aggregate information values of one feature in the case of multi-label classification. As a consequence we have pointed out the advantages and shortcomings of the aggregation types. Also we have proposed and analyzed different schemes of allocation of the selected features to diverse subsets that are suitable for training ensembles of classifiers. Equally important was the proposed method of transforming multi-label classification problems into multi-class in order to be able to apply some feature selection algorithms. We have tested the proposed methods on the AAIA'14 data mining competition dataset [41] and our solution has been recognized as one of the top 5.

References

[1] H. Liu and H. Motoda, *Feature Extraction, Construction and Selection a Data Mining Perspective*. Boston, MA: Springer US, 1998. ISBN 9781461557258 1461557259. [Online]. Available: http://dx.doi.org/10.1007/978-1-4615-5725-8

[2] C. J. Matheus and L. A. Rendell, "Constructive induction on decision trees," in *Proceedings of the 11th International Joint Conference on Artificial Intelligence - Volume 1*, ser. IJCAI'89. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1989, pp. 645–650. [Online]. Available: http://dl.acm.org/citation.cfm?id=1623755.1623857

[3] J. W. Osborne and A. Overbay, "The power of outliers (and why researchers should always check for them)," *Practical assessment, research & evaluation*, vol. 9, no. 6, pp. 1–12, 2004.

[4] P. Grassberger, R. Hegger, H. Kantz, C. Schaffrath, and T. Schreiber, "On noise reduction methods for chaotic data," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 3, no. 2, 1993.

[5] R. J. A. Little, *Statistical analysis with missing data*, 2nd ed., ser. Wiley series in probability and statistics. Hoboken, N.J: Wiley, 2002. ISBN 0471183865

[6] P. Royston, "Multiple imputation of missing values," *Stata Journal*, vol. 4, pp. 227–241, 2004.

[7] A. A. Hancock, E. N. Bush, D. Stanisic, J. J. Kyncl, and C. Lin, "Data normalization before statistical analysis: keeping the horse before the cart," *Trends in Pharmacological Sciences*, vol. 9, no. 1, pp. 29 – 32, 1988. doi: http://dx.doi.org/10.1016/0165-6147(88)90239-8. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0165614788902398

TABLE I
TABLE OF SELECTED FEATURES AND THEIR INFORMATION VALUES

Index	Subset	IV 1	IV 2	IV 3	Average IV	Max IV
3001	1	0.0700	0.3834	0.2928	0.2487	0.3834
258	1	0.0041	0.0612	0.2769	0.1141	0.2769
11463	1	0.2804	0.4048	0.4848	0.3900	0.4848
5041	1	0.0040	0.1496	0.1617	0.1051	0.1617
1880	1	0.4021	0.0992	0.1233	0.2082	0.4021
5996	1	0.0504	0.1541	0.2184	0.1410	0.2184
4415	2	0.1306	0.1852	0.0843	0.1334	0.1852
4698	2	0.0105	0.3537	0.3403	0.2348	0.3537
3027	2	0.1437	0.1611	0.1555	0.1534	0.1611
10428	2	0.0835	0.2628	0.4815	0.2759	0.4815
1772	2	0.1626	0.7043	0.5911	0.4860	0.7043
3955	2	0.0086	0.2998	0.2601	0.1895	0.2998
1509	3	0.0005	1.3586	1.0919	0.8170	1.3586
143	3	0.5076	1.4528	1.6283	1.1962	1.6283
8114	3	0.0443	0.6701	0.4605	0.3916	0.6701
1538	3	0.0236	0.3462	0.3900	0.2533	0.3900
7999	3	0.0372	0.1050	0.2726	0.1383	0.2726
7425	3	0.6674	0.2486	0.2163	0.3774	0.6674
5270	4	1.6636	0.0854	0.1233	0.6241	1.6636
460	4	0.0339	0.3748	0.4775	0.2954	0.4775
11701	4	0.3482	1.7748	1.9455	1.3562	1.9455
142	4	0.0051	1.8940	0.1132	0.6708	1.8940
11165	4	0.6436	0.3584	0.3611	0.4544	0.6436
139	5	0.0024	0.9596	0.6780	0.5467	0.9596
7055	5	0.4033	0.0169	0.0846	0.1683	0.4033
1335	5	0.0127	0.1994	0.0969	0.1030	0.1994
11825	5	0.3693	0.6592	0.5104	0.5130	0.6592
8635	5	0.3322	1.1180	1.0114	0.8205	1.1180
6660	6	0.1667	0.3141	0.3896	0.2902	0.3896
11459	6	0.0142	0.1378	0.1587	0.1036	0.1587
10771	6	0.3630	0.1023	0.1219	0.1957	0.3630
6779	6	0.0035	1.1367	0.8519	0.6640	1.1367
9657	6	0.2205	0.2340	0.1535	0.2027	0.2340
5306	7	0.4444	0.0590	0.0870	0.1968	0.4444
11100	7	0.0297	0.6239	0.5207	0.3914	0.6239
10638	7	0.5431	0.2341	0.1554	0.3109	0.5431
1244	7	0.1110	1.1247	0.5446	0.5934	1.1247
7187	7	0.0051	1.8906	0.1126	0.6694	1.8906
5909	8	0.0015	1.3041	1.0542	0.7866	1.3041
4210	8	0.0002	0.1624	0.1548	0.1058	0.1624
7007	8	0.1164	0.1948	0.1086	0.1400	0.1948
1767	8	0.0197	0.1962	0.1891	0.1350	0.1962
1152	8	0.2480	0.2208	0.1964	0.2217	0.2480
5925	9	0.0061	0.5386	0.4829	0.3425	0.5386
8039	9	0.5741	0.1034	0.0517	0.2431	0.5741
2182	9	1.0820	0.6352	0.8652	0.8608	1.0820
8073	9	0.5194	0.1280	0.2433	0.2969	0.5194
3257	9	0.6396	0.7863	0.5737	0.6665	0.7863
6162	10	0.0629	0.5856	0.6844	0.4443	0.6844
8487	10	0.1549	0.4232	0.3095	0.2959	0.4232
8914	10	0.0292	0.2425	0.0870	0.1196	0.2425
10968	10	0.1452	0.3279	0.3999	0.2910	0.3999
1038	10	0.2116	0.3394	0.3006	0.2839	0.3394

[8] J. Sola and J. Sevilla, "Importance of input data normalization for the application of neural networks to complex industrial problems," *Nuclear Science, IEEE Transactions on*, vol. 44, no. 3, pp. 1464–1468, Jun 1997. doi: 10.1109/23.589532

[9] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, *Advances in knowledge discovery and data mining*. Menlo Park, Calif.: AAAI Press : MIT Press, 1996. ISBN 0262560976 9780262560979

[10] E. Zdravevski, P. Lameski, and A. Kulakov, "Weight of evidence as a tool for attribute transformation in the preprocessing stage of supervised learning algorithms," in *Neural Networks (IJCNN), The 2011 International Joint Conference on*, July 2011. doi: 10.1109/IJCNN.2011.6033219. ISSN 2161-4393 pp. 181–188.

[11] T. M. Mitchell, *Machine Learning*, 1st ed. McGraw-Hill Science/Engineering/Math, 3 1997. ISBN 9780070428072. [Online]. Available: http://amazon.com/o/ASIN/0070428077/

[12] D. Mladenic and M. Grobelnik, "Feature selection for unbalanced class distribution and naive bayes," in *Proceedings of the Sixteenth International Conference on Machine Learning*, ser. ICML '99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999. ISBN 1-55860-612-2 pp. 258–267. [Online]. Available: http://dl.acm.org/citation.cfm?id=645528.657649

[13] R. O. Duda, *Pattern classification*, 2nd ed. New York: Wiley, 2001. ISBN 0471056693

[14] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. ISBN 1-55860-238-0

[15] H. Almuallim and T. G. Dietterich, "Learning with many irrelevant features," in *Proceedings of the Ninth National Conference on Artificial Intelligence - Volume 2*, ser. AAAI'91. AAAI Press, 1991. ISBN 0-262-51059-6 pp. 547–552. [Online]. Available: http://dl.acm.org/citation.cfm?id=1865756.1865761

[16] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1âĂŞ2, pp. 245 – 271, 1997. doi: http://dx.doi.org/10.1016/S0004-3702(97)00063-5 Relevance. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0004370297000635

[17] P. Langley, *Elements of machine learning*. San Francisco, Calif: Morgan Kaufmann, 1996. ISBN 1558603018

[18] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Machine Learning: Proceedings of the Eleventh International Conference*. Morgan Kaufmann, 1994, pp. 121–129.

[19] B. Raman and T. R. Ioerger, "Instance based filter for feature selection," *Journal of Machine Learning Research*, vol. 1, no. 3, pp. 1–23, 2002.

[20] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944968

[21] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1-2, pp. 273–324, Dec. 1997. doi: 10.1016/S0004-3702(97)00043-X. [Online]. Available: http://dx.doi.org/10.1016/S0004-3702(97)00043-X

[22] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter, "Distributional word clusters vs. words for text categorization," *J. Mach. Learn. Res.*, vol. 3, pp. 1183–1208, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944969

[23] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944974

[24] L. Hermes and J. Buhmann, "Feature selection for support vector machines," in *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, vol. 2, 2000. doi: 10.1109/ICPR.2000.906174. ISSN 1051-4651 pp. 712–715 vol.2.

[25] R. Anderson, *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation*. Oxford: Oxford University Press, 2007. ISBN 9780199226405

[26] S. Finlay, *Credit scoring, response modeling, and insurance rating: a practical guide to forecasting consumer behavior*, 2nd ed. Houndmills, Basingstoke, Hampshire ; New York: Palgrave Macmillan, 2012. ISBN 9780230347762

[27] N. E. Mays, Lynas, *Credit scoring for risk managers: the handbook for lenders*. S.l.: CreateSpace], 2010. ISBN 9781450578967 1450578969

[28] C. Lee and G. G. Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization," *Inf. Process. Manage.*, vol. 42, no. 1, pp. 155–165, Jan. 2006. doi: 10.1016/j.ipm.2004.08.006. [Online]. Available: http://dx.doi.org/10.1016/j.ipm.2004.08.006

[29] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, pp. 79–86, 1951.

[30] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proceedings of the Ninth International Workshop on Machine Learning*, ser. ML92. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1992. ISBN 1-5586-247-X pp. 249–256. [Online]. Available: http://dl.acm.org/citation.cfm?id=141975.142034

[31] I. Kononenko, "Estimating attributes: Analysis and extensions of relief," in *Machine Learning: ECML-94*, ser. Lecture Notes in Computer Science, F. Bergadano and L. De Raedt, Eds. Springer Berlin Heidelberg, 1994, vol. 784, pp. 171–182. ISBN 978-3-540-57868-0. [Online]. Available: http://dx.doi.org/10.1007/3-540-57868-4_57

[32] T. Jebara and T. Jaakkola, "Feature selection and dualities in maximum entropy discrimination," in *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000. ISBN 1-55860-709-9 pp. 291–300. [Online]. Available: http://dl.acm.org/citation.cfm?id=2073946.2073981

[33] A. Vehtari and J. Lampinen, "Bayesian input variable selection using posterior probabilities and expected utilities," *Report B31*, 2002.

[34] A. Y. Ng and M. I. Jordan, "Convergence rates of the voting gibbs classifier, with application to bayesian feature selection," in *In 18th International Conference on Machine Learning*. Morgan Kaufmann, 2001.

[35] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[36] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *ICML*, vol. 3, 2003, pp. 856–863.

[37] M. Dash, H. Liu, and H. Motoda, "Consistency based feature selection," in *Knowledge Discovery and Data Mining. Current Issues and New Applications*, ser. Lecture Notes in Computer Science, T. Terano, H. Liu, and A. Chen, Eds. Springer Berlin Heidelberg, 2000, vol. 1805, pp. 98–109. ISBN 978-3-540-67382-8. [Online]. Available: http://dx.doi.org/10.1007/3-540-45571-X_12

[38] A. Ben-Hur and I. Guyon, "Detecting stable clusters using principal component analysis," in *Functional Genomics*, ser. Methods in Molecular Biology, M. Brownstein and A. Khodursky, Eds. Humana Press, 2003, vol. 224, pp. 159–182. ISBN 978-1-58829-291-9. [Online]. Available: http://dx.doi.org/10.1385/1-59259-364-X%3A159

[39] P. Yang, W. Liu, B. Zhou, S. Chawla, and A. Zomaya, "Ensemble-based wrapper methods for feature selection and class imbalance learning," in *Advances in Knowledge Discovery and Data Mining*, ser. Lecture Notes in Computer Science, J. Pei, V. Tseng, L. Cao, H. Motoda, and G. Xu, Eds. Springer Berlin Heidelberg, 2013, vol. 7818, pp. 544–555. ISBN 978-3-642-37452-4. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-37453-1_45

[40] S. Das, "Filters, wrappers and a boosting-based hybrid for feature selection," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. ISBN 1-55860-778-1 pp. 74–81. [Online]. Available: http://dl.acm.org/citation.cfm?id=645530.658297

[41] "Aaia'14 data mining competition, howpublished = https://fedcsis.org/2014/dm_competition, note = Accessed: 2014-05-30."

[42] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning," *Pattern Recognition*, vol. 45, no. 9, pp. 3084 – 3104, 2012. doi: http://dx.doi.org/10.1016/j.patcog.2012.03.004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320312001203

[43] A. P. Bradley, "The use of the area under the {ROC} curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145 – 1159, 1997. doi: http://dx.doi.org/10.1016/S0031-3203(96)00142-2. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320396001422

[44] C. X. Ling, J. Huang, and H. Zhang, "Auc: A statistically consistent and more discriminating measure than accuracy," in *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, ser. IJCAI'03. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003, pp. 519–524. [Online]. Available: http://dl.acm.org/citation.cfm?id=1630659.1630736

[45] J. Huang and C. Ling, "Using auc and accuracy in evaluating learning algorithms," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, no. 3, pp. 299–310, March 2005. doi: 10.1109/TKDE.2005.50

[46] N. Siddiqi, *Credit risk scorecards: developing and implementing intelligent credit scoring*. Hoboken, N.J: Wiley, 2006. ISBN 9780471754510

[47] L. Bruce and D. Brotherton, "Information value statistic," in *Midwest SAS User Group 2013 Conference Proceedings*. Marketing Associates, LLC, 2013, pp. 1–18.

[48] A. Rrnyi, "On measures of entropy and information," in *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1961, pp. 547–561.

[49] P. J. Fleming and J. J. Wallace, "How not to lie with statistics: The correct way to summarize benchmark results," *Commun. ACM*, vol. 29, no. 3, pp. 218–221, Mar. 1986. doi: 10.1145/5666.5673. [Online]. Available: http://doi.acm.org/10.1145/5666.5673

[50] M. J. Schervish, "P values: what they are and what they are not," *The American Statistician*, vol. 50, no. 3, pp. 203–206, 1996.

[51] L. Talavera, "An evaluation of filter and wrapper methods for feature selection in categorical clustering," in *Advances in Intelligent Data Analysis VI*, ser. Lecture Notes in Computer Science, A. Famili, J. Kok, J. Pena, A. Siebes, and A. Feelders, Eds. Springer Berlin Heidelberg, 2005, vol. 3646, pp. 440–451. ISBN 978-3-540-28795-7. [Online]. Available: http://dx.doi.org/10.1007/11552253_40

7th Workshop on Computational Optimization

Many real world problems arising in engineering, economics, medicine and other domains can be formulated as optimization tasks. These problems are frequently characterized by non-convex, non-differentiable, discontinuous, noisy or dynamic objective functions and constraints which ask for adequate computational methods.

The aim of this workshop is to stimulate the communication between researchers working on different fields of optimization and practitioners who need reliable and efficient computational optimization methods.

We invite original contributions related to both theoretical and practical aspects of optimization methods.

Topics

The list of topics includes, but is not limited to:

- unconstrained and constrained optimization
- combinatorial optimization
- global optimization
- multiobjective optimization
- optimization in dynamic and/or noisy environments
- large scale optimization
- parallel and distributed approaches in optimization
- random search algorithms, simulated annealing, tabu search and other derivative free optimization methods
- nature inspired optimization methods (evolutionary algorithms, ant colony optimization, particle swarm optimization, immune artificial systems etc)
- hybrid optimization algorithms involving natural computing techniques and other global and local optimization methods
- optimization methods for learning processes and data mining
- computational optimization methods in statistics, econometrics, finance, physics, medicine, biology, engineering etc

Event Chairs

Fidanova, Stefka, Academy of Sciences, Bulgaria
Mucherino, Antonio, IRISA, France
Zaharie, Daniela, West University of Timisoara, Romania

Program Committee

Bartl, David, University of Ostrava, Czech Republic
Breaban, Mihaela
Cremonesi, Paolo
Gualandi, Stefano
Hoai Ann, Le Thi
Hosobe, Hiroshi, Hosei University, Japan
Iiduka, Hideaki, Kyushu Institute of Technology, Japan
Lavor, Carlile, IMECC-UNICAMP, Brazil
Marinov, Pencho, Bulgarian Academy of Science, Bulgaria
Michini, Carla
Miettinen, Kaisa, University of Jyvaskyla, Finland
Mihalas, Stelian, West University of Timisoara
Muscalagiu, Ionel, Politehnica University Timisoara
Nannicini , Giacomo
Parsopoulos, Konstantinos, University of Patras
Pop, Petrica
Roeva, Olympia, Institute of Biophysics and Biomedical Engineering, Bulgaria
Siarry, Patrick, Universite Paris XII Val de Marne, France
Slezak, Dominik, University of Warsaw & Infobright Inc., Poland
Stefanov, Stefan, South-West University ""Neofit Rilski, Bulgaria
Stuetzle, Thomas, Université Libre de Bruxelles (ULB), Belgium
Suganthan, Ponnuthurai Nagaratnam, Nanyang Technological University, Singapore
Tamir, Tami, The Interdisciplinary Center (IDC), Israel
Tvrdik, Josef, University of Ostrava, Czech Republic
Vrahatis, Michael, University of Patras, Greece
Wolfler Calvo, Roberto
Zilinskas, Antanas, Vilnius University

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 397–404
DOI: 10.15439/2014F145
ACSIS, Vol. 2

A Look-Forward Heuristic for Packing Spheres into a Three-Dimensional Bin

Hakim Akeb
ISC Paris Business School
22 boulevard du Fort de Vaux
75017 Paris, France
Email: hakeb@iscparis.com

***Abstract*—In this paper a look-forward heuristic is proposed in order to solve the problem of packing spheres into a three-dimensional bin of fixed height and depth but variable length. The objective is to pack all the spheres into the bin of minimum length. This problem is also known under the name of three-dimensional strip packing problem. The computational investigation, conducted on a set of benchmark instances taken from the literature, shows that the method is effective since it improves most of the best known results.**

I. Introduction

PACKING spheres can be used to model many solid state systems. Indeed, the association of different-sized spheres for example can approximate a given solid form. Packing (non-)identical spheres is for example used in the domain of stereotactic radio surgery radiation therapy (see for example the works of Gavriliouk [3], Sutou and Dai [13], and Wang [15]) where the target areas are delimited by spheres of different sizes.

The problem of packing spheres into a container was studied by several authors in the literature. The spheres can be of identical or different sizes (radii). The problem of packing non-identical spheres into a given 3D container was for example considered by Li and Ji [8] where a dynamics-based collective method for random sphere packing was proposed as well as an application to the problem of packing spheres into a cylinder container. The authors studied also the stability of the method and the convergence of their algorithm. Sutou and Dai [13] used a global optimization approach (including Linear Programming relaxation and branch-and-bound) in order to place unequal spheres inside a three-dimensional container. More precisely, the objective is to maximize the volume of the container (of fixed size) occupied by the placed spheres. This is also called the *Knapsack* version of the problem, i.e., the objective is not to place all the objects but those maximizing the obtained profit. The profit used often corresponds to the volume of the corresponding object placed. Stoyan, Yaskov, and Scheithauer [12] developed a mathematical model in order to place different-sized spheres inside a parallelepiped of fixed length and width but with variable height. The objective is then to minimize the height of the container. The proposed method uses different tools including extreme points and neighborhood search. Solutions are given for a set containing eight instances (designed by the authors) where the number of spheres varies from 20 to 60.

For the case of identical spheres, M'Hallah, Alkandari, and Mladenović [9] for example studied the problem of packing spheres of the same radius into the smallest containing sphere by using Variable Neighborhood Search (VNS) and Non-Linear Programming (NLP). VNS here consists to move some spheres situated in the neighborhood of a given placed sphere, then a NLP procedure is called in order to remove overlapping between spheres. M'Hallah and Alkandari [10] applied the same principle (VNS and NLP) as in [9] to solve the problem of packing unit spheres into the smallest cube. Soontrapa and Chen [11] considered the problem of packing identical spheres into a cube by using a random search technique based on the Monte Carlo method. The problem concerns actually the development of a fuel catalyst layer.

Finally Birgin and Sobral [2] studied the problem of packing identical and non-identical spheres into different three-dimensional containers. The objective is to minimize the dimension of the container. The method proposed by the authors is based on twice-differentiable models as well as non-linear programming.

The problem to solve in this paper is the Three-Dimensional Strip Packing Problem (3DSPP) which is known to be NP-Hard [6]. Given a set S containing n spheres $s_i, 1 \leq i \leq n$ where each sphere has radius r_i and is placed with its center at coordinates (x_i, y_i, z_i) in the Euclidean space. Let also $\mathbb{B}$ be a three-dimensional bin (rectangular cuboid or parallelepiped) of fixed height and depth (H, D) respectively but of unconstrained length L. The objective is then to place the n spheres inside the parallelepiped of minimum length such that no sphere overlaps another sphere and no sphere exceeds the container boundaries. The method presented is based on the use of several tools including the *Maximum Hole Degree* (MHD) heuristic, a modified look-forward strategy, and an interval search.

II. Problem formulation

The three-dimensional bin $\mathbb{B}$ has six faces $\mathbb{F} = \{\text{left}, \text{top}, \text{right}, \text{bottom}, \text{back}, \text{front}\}$ and is placed such that its bottom-left-back corner corresponds to the origin $O(0,0,0)$ of the axes in the Euclidean space as shown in Fig. 1. The length L, the height H, and the depth D of the container are associated with the $\overrightarrow{Ox}$, $\overrightarrow{Oy}$, and $\overrightarrow{Oz}$ axes respectively. Moreover, each

978-83-60810-58-3/$25.00

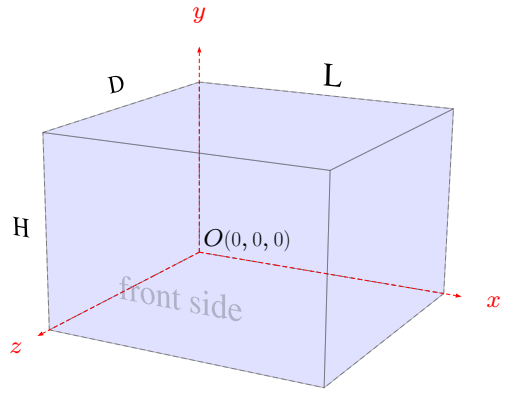


Fig. 1. The three-dimensional bin container placed with its bottom-left-back corner at the origin of the axes in the Euclidean space.

sphere $s_i \in S$ has radius r_i and its center's coordinates are (x_i, y_i, z_i). The 3DSPP can then be formulated as follows:

$$\min L \tag{1}$$

$$(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2 \geq (r_i + r_j)^2 \quad \text{for } 1 \leq i < j \leq n \tag{2}$$

$$x_i \geq r_i \qquad \forall i \in [1, \ldots, n] \tag{3}$$

$$x_i \leq L - r_i \qquad \forall i \in [1, \ldots, n] \tag{4}$$

$$y_i \geq r_i \qquad \forall i \in [1, \ldots, n] \tag{5}$$

$$y_i \leq H - r_i \qquad \forall i \in [1, \ldots, n] \tag{6}$$

$$z_i \geq r_i \qquad \forall i \in [1, \ldots, n] \tag{7}$$

$$z_i \leq D - r_i \qquad \forall i \in [1, \ldots, n] \tag{8}$$

Equation 1 indicates the objective (value) to minimize (the length L of the bin). Equation 2 is the non-overlapping constraint that verifies that any pair of distinct spheres $(s_i, s_j) \in S^2$ do not overlap each other. Equations 3–8 mean that each sphere must not exceed the boundaries of the container.

The distance between the edges of two distinct spheres s_i and s_j, denoted by $d_{i,j}$, is defined as follows:

$$d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} - r_i - r_j \quad \text{for } i \neq j \tag{9}$$

III. The 3DMHD heuristic for packing spheres into a three-dimensional bin

In this section, a greedy heuristic, denoted by 3DMHD (Three-Dimensional Maximum Hole Degree), for packing spheres into a three-dimensional bin is described. This is in fact the adaptation of the Maximum Hole Degree (MHD) heuristic [4], designed for packing circles, to the three-dimensional case.

Note that a simple way to pack the spheres inside the container consists for example to place the first sphere s_1 at the bottom-left-back corner, i.e., at coordinates (r_1, r_1, r_1). After that, at each step $i, (1 < i \leq n)$ a new sphere is chosen and is placed at the *best* position (that has the maximum hole degree). More precisely, let:

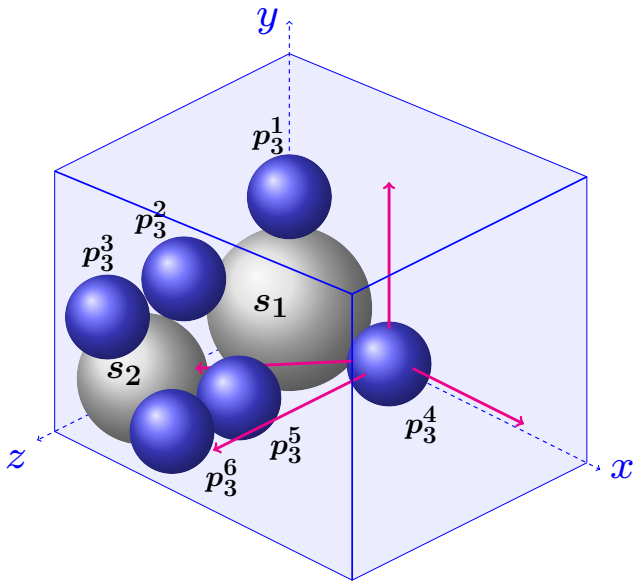


Fig. 2. The 3DMHD heuristic for packing spheres into a three-dimensional bin.

- S_{in} define the set of spheres already placed inside the container.
- S_{out} is the complementary set containing the spheres that are not yet placed (outside). Note that $S_{\text{in}} \cup S_{\text{out}} = S$.
- P denotes the set of possible positions (called *corner positions*) for the spheres of set S_{out}.

Fig. 2 shows an example where two spheres s_1 and s_2 (the two greatest ones) are already placed inside the container $\mathbb{B}$, i.e., $S_{\text{in}} = \{s_1, s_2\}$. The figure also indicates six possible corner positions for packing another sphere s_3. These positions are denoted by $\{p_3^1, \ldots, p_3^6\}$. Each position p_3^k is computed by using three objects, an object may be a sphere already placed or one of the six faces of the parallelepiped. These three objects denote set $\mathscr{T}(p_3^k)$ associated to this position. For example, position p_3^1 is computed by using sphere s_1, the left-edge and the back-face of the container, then $\mathscr{T}(p_3^1) = \{s_1, \text{left}, \text{back}\}$. Similarly, $\mathscr{T}(p_3^2) = \{s_1, s_2, \text{left}\}$.

Generally, let position $p_{i+1}^k \in P$, associated to a sphere of radius r_{i+1}^k, be one of the possible corner positions for the next sphere s_{i+1} to place. Then, the 3DMHD value for position p_{i+1}^k is defined as follows:

$$\lambda(p_{i+1}^k) = \max_{j \in S_{\text{in}} \cup \mathbb{F} \setminus \mathscr{T}(p_{i+1}^k)} 1 - \frac{d_{i+1,j}^k}{r_{i+1}^k} \tag{10}$$

Equation 10 means that the hole degree $\lambda(p_{i+1}^k)$ is computed for each position in the set of positions P (associated with set S_{out}) for the next sphere to place. This value uses the distance $d_{i+1,j}^k$ between the edge of position p_{i+1}^k and the nearest object j in the set $S_{\text{in}} \cup \mathbb{F} \setminus \mathscr{T}(p_{i+1}^k)$ that contains the spheres already placed, the six faces $(\mathbb{F})$ of the container but

Algorithm 1 The 3DMHD greedy heuristic

Require: Set S_{in} containing spheres already placed, S_{out} containing the remaining spheres to place, set P indicating the possible positions for spheres in S_{out} and the current length L of the container.

Ensure: TRUE if all the spheres are packed into the container, FALSE otherwise.

1: $i \leftarrow |S_{\text{in}}|$;
2: **while** $(P \neq \emptyset)$ **do**
3: Compute/update the 3DMHD value for each corner position $p \in P$;
4: Place the next sphere s_{i+1} at position p^* that has the maximum hole degree as shown in equation 11.
5: Move sphere s_{i+1} from S_{out} to S_{in};
6: Remove from set P the positions that overlap the new inserted sphere;
7: Compute new positions by using the new inserted sphere and the other objects already placed;
8: $i \leftarrow i+1$;
9: **end while**
10: **if** $(i = n)$ **then**
11: Set $L \leftarrow \max(x_i + r_i)$;
12: Update the best known length if L is smaller than this value;
13: **return** TRUE;
14: **else**
15: **return** FALSE;
16: **end if**

excluding set $\mathscr{T}(p^k_{i+1})$. The distance is divided by the radius r^k_{i+1} of the sphere corresponding to position p^k_{i+1}. Note that if a given position touches more than three objects, then $\lambda = 1$, meaning that this positions has a high probability to be chosen for placing the next sphere.

For example, Fig. 2 indicates the distance between position p^4_3 and four other objects: sphere s_2, the front face, the top face, and finally the right face of the container.

Then, the 3DMHD heuristic places the next sphere at position $p^* \in P$ that corresponds to the maximum value of $\lambda(p^k_{i+1})$ as indicated in equation 11.

$$p^* = \arg\max_{p^k_{i+1}} \lambda(p^k_{i+1}) \tag{11}$$

Algorithm 1 explains how the 3DMHD heuristic proceeds in order to place a set of spheres inside the container $\mathbb{B}$ of dimensions $(L \times H \times D)$. Procedure 3DMHD receives a partial solution $\{S_{\text{in}}, S_{\text{out}}, P\}$ indicating the spheres already packed into the container, the remaining spheres, and the set of corner positions for spheres in S_{out} respectively. The current length L of the container is also transmitted to the procedure. The heuristic's output is a boolean value indicating whether yes or no all the spheres were successfully packed into the container. So procedure 3DMHD is able to start with any partial solution were the number of spheres already packed is greater than or equal to zero.

At line 1 in algorithm 1 counter i indicating the number of spheres already packed is set to the number of spheres inside S_{in}. After that, in the **while** loop, the 3DMHD value is computed for each position $p \in P$ (line 3), this is done by using the formula of equation 10. At line 4, the best position p^* is chosen in order to place the next sphere s_{i+1}. After that, the new sphere moves from set S_{out} to set S_{in} (line 5) and the set of positions P is updated by removing those overlapping the new inserted sphere (line 6) and by computing new positions by using the new inserted sphere (line 7). Counter i is then incremented at line 8. The **while** loop ends when the set of positions P becomes empty meaning that no additional sphere can be packed. Then two cases can be distinguished: if $i = n$ then all the n spheres were successfully packed into the container. In this case the procedure computes at line 11 the exact value for L which is equal to $\max(x_i + r_i)$, i.e, using the most right placed sphere $s_i \in S_{\text{in}}$. If the obtained value L is smaller than the best known length then this value is updated (line 12) and the procedure returns TRUE (line 13). If $i < n$ then a feasible packing was not obtained and the procedure returns FALSE (line 15), this means that the current length L of the container has to be changed.

Note that one can test several values for the length L of the bin in order to try to compute a feasible solution with the 3DMHD heuristic (not necessarily a binary search but other more efficient strategies). This can be done for example by decreasing the length from an upper bound to a lower bound. Indeed, this strategy may escape from local optima (see Section IV below).

A. A Multi-Level Look-Forward strategy for the 3DSPP

This section describes a look-forward algorithm designed for the three-dimensional strip packing problem.

Look-forward (LF) strategies (see for example [7], [4], [1]) are often used in order to improve the results obtained by different algorithms. Its objective is to evaluate the future behavior of a decision (choice) made at a given step of the problem solving process. For example, in a greedy algorithm, the *best* decision among all the possible decisions is made at step i in order to move to the next step $i+1$. The look-ahead strategy tries several (or all) choices at step i and see what will be obtained when executing the greedy algorithm few steps ahead of until the end (this is often executed on a copy of the partial solution). After that, the decision actually made at step i is the one that had the best behavior or led to the best outcome.

In packing problems, the look-forward strategy often uses a parameter called *density* of a solution. The density of a solution S_{in}, denoted by $density(S_{\text{in}})$ is equal to the sum of the volumes of spheres in S_{in} divided by the volume of the container as indicated in Equation 12. The look-forward strategy selects then the decision that will obtain the highest density.

$$density(S_{\text{in}}) = \frac{4 \times \pi \times \sum_{i=1}^{|S_{\text{in}}|}(r_i^3)}{3 \times L \times H \times D} \tag{12}$$

Algorithm 2 LF-3DMHD

Require: Sets S_{in}, S_{out}, P, and the current length L of the container.
Ensure: TRUE if all the spheres are packed into the container, FALSE otherwise.

1: $i \leftarrow |S_{\text{in}}|$;
2: found $\leftarrow$ FALSE;
3: **while** ($P \neq \emptyset$ and found=FALSE) **do**
4: Sort the positions of set P in decreasing order of their hole degree (λ) value;
5: **for each** of the first $\psi_1 \times |P|$ positions $p \in P$ **do**
6: Let $S'_{\text{in}} \leftarrow S_{\text{in}}$, $S'_{\text{out}} \leftarrow S_{\text{out}}$ and $P' \leftarrow P$;
7: Insert the next sphere s'_{i+1} into S'_{in} at position p and update sets $S'_{\text{in}}, S'_{\text{out}}$, and P';
8: $density^* \leftarrow 0$;
9: Sort the positions of set P' in decreasing order of their hole degree (λ) value;
10: **for each** of the first $\psi_2 \times |P'|$ positions $p' \in P'$ **do**
11: Let $S''_{\text{in}} \leftarrow S'_{\text{in}}$, $S''_{\text{out}} \leftarrow S'_{\text{out}}$ and $P'' \leftarrow P'$;
12: Insert the next sphere s''_{i+2} into S''_{in} at position p' and update sets $S''_{\text{in}}, S''_{\text{out}}$, and P'';
13: found $\leftarrow$ 3DMHD($S''_{\text{in}}, S''_{\text{out}}, P'', L$);
14: **if** (found=TRUE) **then**
15: Set L equal to the length computed by 3DMHD;
16: **return** TRUE;
17: **else**
18: **if** ($density(S''_{\text{in}}) > density^*$) **then**
19: $density^* \leftarrow density(S''_{\text{in}})$;
20: **end if**
21: **end if**
22: **end for**
23: Assign to position $p \in P$ the density $density^*$ obtained after calling 3DMHD;
24: **end for**
25: Let $p^* \in P$ be the position that has obtained the highest density $density^*$;
26: Place the next sphere s_{i+1} at position p^* and move sphere s_{i+1} from S_{out} to S_{in};
27: Remove from set P the positions that overlap the new inserted sphere;
28: Compute new positions by using the new inserted sphere;
29: $i \leftarrow i+1$;
30: **end while**
31: **if** ($i = n$) **then**
32: Set $L \leftarrow \max(x_i + r_i)$ where x_i and r_i are the $x-$coordinate and the radius of sphere $s_i \in S_{\text{in}}$;
33: Update the best known length if L is smaller than this value;
34: **return** TRUE;
35: **else**
36: **return** FALSE;
37: **end if**

The algorithm that implements the look-forward strategy, denoted by LF-3DMHD, is described in algorithm 2. It receives as input parameters a partial solution $\{S_{\text{in}}, S_{\text{out}}, P\}$ where $|S_{\text{in}}|$ spheres are already packed, set S_{out} denotes the spheres that remain to pack and P contains the corner positions for spheres of set S_{out}. The algorithm receives also the current length (L) of the container. Algorithm LF-3DMHD returns TRUE if it succeeds to compute a feasible solution, FALSE otherwise.

Instruction at line 1 of algorithm 2 sets the counter i indicating the number of spheres already packed. At line 2, a boolean value (found) is set to FALSE (this indicator is set to TRUE if a feasible solution is obtained).

The difference between the look-forward strategy and the 3DMHD heuristic (described in algorithm 1) is that the look-forward tries (evaluates) several positions at each step of the packing process while the greedy heuristic 3DMHD selects, at each step, only one position (the best one) in order to pack the next sphere. Moreover, the look-forward used here contains two levels, i.e., it places the two next spheres and continues the placement of the remaining spheres by using the greedy heuristic 3DMHD (algorithm 1). This is implemented by using two nested **for** loops that begin at lines 5 and 10 respectively. In addition, the first **for** loop considers only the best $\psi_1 \times |P|$ positions with $0 < \psi_1 \leq 1$ and P is the set of corner positions in the first level. In the second **for** loop the algorithm considers only the best $\psi_2 \times |P'|$ with $0 < \psi_2 \leq 1$ and P' is the set of corner positions in the second level. So if for example $\psi_1 =$

0.5, then only the half best positions in the list of positions are considered in the first level of the look-forward strategy, and if $\psi_1 = 1$, then this means that all the positions will be considered. Using a value of ψ_1 and ψ_2 lower than 1 will of course decrease the computation time of the algorithm.

More precisely, the positions in set P are sorted in decreasing order of their hole degree value (λ). This is done at line 4. In the first **for** loop, the algorithm expends the current solution $\{S_{\text{in}}, S_{\text{out}}, P\}$ by choosing at each time a position $p \in P$ by creating a copy of the current solution denoted by $\{S'_{\text{in}}, S'_{\text{out}}, P'\}$ (line 6) and inserts the next sphere s'_{i+1} at that position (line 7). At line 8, a variable called $density^*$ is set to 0. This parameter is used in order to store the best density obtained in the second level of the look-forward. The corner positions of set P' are after that sorted in decreasing order of their λ value (line 9). The second **for** loop starts at line 10, after placing sphere s'_{i+1}. Like in the first level, only a proportion $\psi_2 \times |P'|$ of the best corner positions are taken into account in set P'. Then for each selected position $p' \in P'$, the procedure creates a copy, denoted by $\{S''_{\text{in}}, S''_{\text{out}}, P''\}$, for the current partial solution $\{S'_{\text{in}}, S'_{\text{out}}, P'\}$ (line 11). After that, the next sphere s''_{i+2} is placed at position p' (line 12). Then, the partial solution is evaluated by calling the 3DMHD heuristic (algorithm 1) at line 13 in order to try to pack the remaining $n-i-2$ spheres. If 3DMHD succeeded to pack all the remaining spheres, then it returns TRUE (line 14), the current length of the container is then set to the length computed by 3DMHD (line 15). The algorithm then exits at line 16 since it has succeeded to pack all the spheres (it returns TRUE). Otherwise (found=FALSE), this means that 3DMHD did not succeed to place all the remaining spheres, then the density of the obtained solution $density(S''_{\text{in}})$ is assigned to the best known density $density^*$ if a better value is obtained (line 19). The second **for** loop ends when all the selected positions $p' \in P'$ are evaluated and the best obtained density $(density^*)$ is assigned to position $p \in P$ that is currently considered in the first **for** loop.

At the output of the two **for** loops, the next sphere s_{i+1} is placed at position p^* (line 26) that has obtained the best density after calling 3DMHD. The set P of positions is then updated at line 27 by removing those that overlap the new inserted sphere and new positions are computed at line 28. The number of placed spheres (i) is incremented at line 29.

Instructions of the **while** loop (lines 3–30) are executed until a feasible solution is obtained (found=TRUE) or the set of positions P becomes empty. So if $i = n$ (line 31), this means that a feasible solution is reached, then the true length of the container is computed at line 32 and the best known length is updated if a better one is obtained (line 33). The algorithm returns TRUE (line 34). If $(i < n)$, then this means that algorithm LF-3DMHD did not succeed to compute a feasible solution and returns FALSE (line 36).

Finally, algorithm 2 can for example be called by an interval-search procedure that modifies the value of the length L of the container at each call as described in section III-B below.

Algorithm 3 (LF2)

Require: Instance S containing n spheres, the height H, and the depth D of the three-dimensional bin $\mathbb{B}$;

Ensure: The best length L^* obtained and the corresponding density $density^*$;

1: Set $L_{\min} \leftarrow \max\left(\frac{4\times\pi\times\sum_{i=1}^{n}(r_i^3)}{3\times H\times D}, 2\times r_{\max}\right)$ be the lower bound of the interval search;
2: Set $L_{\max} \leftarrow 3 \times L_{\min}$;
3: Set $\Delta L \leftarrow 0.01$;
4: $L \leftarrow L_{\max}$;
5: $L^* \leftarrow L$;
6: $density^* \leftarrow 0$;
7: **while** $(L \geq L_{\min})$ **do**
8: $S_{\text{in}} \leftarrow \emptyset$;
9: $S_{\text{out}} \leftarrow S$;
10: Create set P of positions corresponding to the placement of each sphere $s_i \in S$ of radius r_i at position (r_i, r_i, r_i) in the bin of dimensions $L \times H \times D$;
11: found $\leftarrow$ LF-3DMHD$(S_{\text{in}}, S_{\text{out}}, P, L)$;
12: **if** (found = TRUE) **then**
13: Update L if a lower value was obtained by LF-3DMHD;
14: $L^* \leftarrow L$;
15: Update the best density $density^*$;
16: **end if**
17: $L \leftarrow L - \Delta L$;
18: **end while**

B. Interval Search for Computing the Best Packing

This section describes the interval search, denoted by LF2 and described in algorithm 3, used in order to compute the best feasible packing. The search principle consists to decrease the value of the bin length L from an upper bound $L_{\max}$ by a given step ΔL until matching the lower bound $L_{\min}$. The search may also stop if the computation time limit is reached.

Algorithm 3 (LF2) explains how the heuristic proceeds in order to compute the best packing of the n spheres into the three-dimensional bin of minimum length. Procedure LF2 receives as input parameters the instance $S = \{s_1, ..., s_n\}$ containing n spheres of radii $r_1, ..., r_n$ respectively as well as the height H and the depth D of the three-dimensional bin $\mathbb{B}$. The output of the algorithm is the best length found L^* and the corresponding density $(density^*)$ that is equal to the sum of the volumes of the spheres divided by the volume of the bin $(L^* \times H \times D)$.

The continuous lower bound for the length of the container is used as the minimum value $(L_{\min})$ of the interval search (line 1). Note that if this value is lower than the diameter of the greatest sphere, then this diameter $(2 \times r_{\max})$ is used as the lower bound. The upper bound $L_{\max}$ of the interval search is set equal to $3 \times L_{\min}$. The step ΔL with which the length is decreased at each step is defined at line 3, this value is set to 0.01. The length of the container is then set equal to the upper bound $L \leftarrow L_{\max}$ (line 4) and the best length L^* is set equal to

L at line 5. The next instruction serves to initialize the value of the best known density ($density^*$) associated with the best length L^* (line 6).

After that, at each step in the **while** loop (lines 7–18) a starting configuration is created where the set S_{in} of spheres already packed is set equal to the empty set (line 8) and the set of the remaining spheres to pack (S_{out}) is set equal to the instance S. List P of positions for spheres in set S_{out} is then computed (line 10) so that each position is placed at (r_i, r_i, r_i). This is a novel method because most of the greedy heuristics start by placing one or several objects, here only the list of positions is computed and no object is placed.

Algorithm LF-3DMHD is then called at line 11 in order to try to compute a feasible solution (packing the n spheres into the bin of dimensions $L \times H \times D$.) If procedure LF-3DMHD succeeded to pack the n spheres (found=TRUE) then the value of L is updated if a lower value was computed by LF-3DMHD (line 13) and the best length L^* is set equal to L (line 14). The best density $density^*$, corresponding to L^* is then updated at line 15. The value of the length L is after that deceased (line 17), even if a feasible solution was not obtained by procedure LF-3DMHD. Indeed, this method is, to our opinion, preferable to a basic dichotomous search where the dimensions of the container are increased when a feasible solution was not obtained. This is not always a good strategy because, in our case for example, if a feasible solution is not obtained by using a given value of the length L, it may be obtained by using a lower value $L - \Delta L$. In fact, decreasing the value of the length L is a good strategy to escape from local optima in order to increase the solution quality.

Algorithm LF2 stops when the value of L becomes lower than the lower bound L_{min} or when the computation time limit is reached.

IV. Computational Results

In order to evaluate the performance of the proposed algorithm LF2 (algorithm 3), two sets of instances were considered:

- Six instances, denoted by SYS, proposed by Stoyan, Yaskov, and Scheithauer [12]. The number of spheres varies from 25 to 60. All the spheres have different radii in each instance (strongly heterogeneous instances).
- Twelve instances, denoted by KBG1,...,KBG12, proposed by Kubach, Bortfeldt, Tilli, and Gehring [5]. Here, the number of spheres is equal to 30 for the first six instances and 50 spheres for the six last ones. Moreover, instances KBG1–KBG3 and KBG7–KBG9 are strongly heterogeneous since all the radii are different. The other six instances KBG4–KBG6 and KBG10–KBG12 are weakly heterogeneous because there are only $n/10$ different radii in each instance, each radius is duplicated 10 times.

The different procedures and algorithms are coded in C++ language and executed under Linux environment on a computer with a 2.4 GHz processor. The results obtained are compared to those given the B1.6 algorithm [5] that is mainly based on a look-forward strategy and starting configurations, the results taken from [5] were also obtained on a 2.4 GHz processor. Algorithm B1.6 is in fact the adaptation of algorithm B1.5 [4] for placing circles inside a rectangular container to the three-dimensional case. Algorithm B16 however tries more starting configurations than B1.5 does. In addition, B1.6 uses a parameter denoted by τ $(0 < \tau \leq 1)$ that serves to indicate the proportion of corner positions evaluated at each step of the look-forward process. The authors in [5] tried two values: $\tau = 0.8$ and $\tau = 1$. The first case means that only 80% of positions are evaluated by the look-forward while the second case means that all positions are evaluated. So in fact, algorithm B1.6 is executed two times (60 minutes for each value of τ). It is to note that the proposed algorithm LF2 is executed only once during 60 minutes on each instance.

In algorithm LF2, the number of positions evaluated by the look-forward is set to 50% in the two levels ($\psi_1 = \psi_2 = 0.5$). So at each time the corner positions are sorted in decreasing order of their hole degree λ and only the first half ones are evaluated. The objective is of course to save computation time.

Table I shows the results obtained by the different algorithms. Column 1 indicates the instance's name and column 2 its size. The two next columns indicate the height H and the depth D of the container. Column 5 (SYS) indicates the results (best length) obtained by the SYS method [12] on instances SYS1–SYS6. Columns 6 and 7 contain the best results (the best length L and the corresponding density respectively) obtained by algorithm B1.6 on the 18 instances (SYS and KBG) when parameter τ is set equal to 0.8 (80% of positions are evaluated by the look-forward). The next two columns display the same results as the two previous columns but when parameter τ is set equal to 1 (all the positions are evaluated in the look-forward). Columns 10–14 contain the results obtained by the proposed algorithm LF2 on all the considered instances. Column 10 (L) gives the best length obtained and column 11 the corresponding density. Column 12 (t^*) indicates the time needed by algorithm LF2 for computing the best solution. The two last columns of table I indicate the percentage of improvement obtained by the proposed algorithm LF2 on algorithm B1.6. Column "Imp. 0.8" shows the improvement obtained when considering B1.6 with $\tau = 0.8$ and the last column "Imp. 1" is the percentage of improvement when B1.6 with $\tau = 1$ is considered. Note that the percentage of improvement is computed as follows: Imp. $= \frac{Density(\text{LF2}) - Density(\text{B1.6})}{Density(\text{LF2})}$. Finally, note that some solutions for KBG instances are optimal, this is the case for instances KBG2, KBG4, and KBG10. This is why there is an "*" before each value in the three columns that contain the corresponding density in table I.

The results of table I indicate that the proposed algorithm LF2 improves all the results obtained by the SYS method on the first six instances (the results of the SYS method are not known for instances KBG). Algorithm LF2 improves B1.6 with $\tau = 0.8$ in 11 cases out of 18 and the two algorithms reach the optimal value of the container length for instances KBG2, KBG4, and KBG10 since the computed length is equal to the greatest diameter in the instance. Algorithm B1.6 with $\tau = 0.8$ remain better than LF2 on instances KBTG5, KBTG6,

TABLE I
RESULTS OBTAINED BY THE PROPOSED METHOD LF2 ON INSTANCES SYS AND KBTG

				SYS	**B1.6 $\tau = 0.8$ (3600 s)**		**B1.6 $\tau = 1$ (3600 s)**		**Algorithm LF2 (3600 s)**				
Instance	n	H	D	L	L	*Density*	L	*Density*	L	*Density*	t^*	Imp. 0.8	Imp. 1
SYS1	25	5.5	6.9	9.8668	9.5397	53.160	9.2874	54.604	9.2234	**54.983**	1911	3.32	0.69
SYS2	35	6.5	7.9	9.6221	9.2608	55.077	9.1280	55.878	9.1138	**55.965**	2680	1.59	0.16
SYS3	40	5.5	6.9	9.4729	9.0540	53.554	8.9850	53.965	8.9316	**54.288**	2900	1.35	0.59
SYS4	45	8.5	9.9	11.0862	10.8932	53.771	10.8760	53.856	10.7653	**54.410**	3600	1.17	1.02
SYS5	50	8.5	9.9	11.6453	11.2170	54.975	11.3494	54.334	11.1948	**55.084**	2030	0.20	1.36
SYS6	60	8.5	9.9	12.8416	12.5339	54.346	12.3745	55.046	12.2519	**55.597**	3330	2.25	0.99
KBG1	30	10	10		–	53.772	–	54.096	11.2063	**54.494**	2400	1.32	0.73
KBG2	30	10	10		–	* **30.071**	–	* **30.071**	1.9900	* **30.071**	2	0.00	0.00
KBG3	30	10	10		–	50.614	–	51.387	18.9231	**51.693**	3300	2.09	0.59
KBG4	30	10	10		–	* **37.765**	–	* **37.765**	1.9960	* **37.765**	1	0.00	0.00
KBG5	30	10	10		–	**48.278**	–	**48.278**	1.9279	48.181	1930	-0.20	-0.20
KBG6	30	10	10		–	**48.966**	–	47.792	18.8807	48.847	3400	-0.24	2.16
KBG7	50	10	10		–	54.623	–	55.372	13.5075	**55.824**	2030	2.15	0.81
KBG8	50	10	10		–	44.924	–	45.060	2.6027	**46.639**	326	3.68	3.39
KBG9	50	10	10		–	52.210	–	**52.732**	29.7023	51.783	3420	-0.82	-1.83
KBG10	50	10	10		–	* **51.866**	–	* **51.866**	1.8100	* **51.866**	9	0.00	0.00
KBG11	50	10	10		–	51.629	–	**52.708**	5.2640	52.658	420	1.95	-0.09
KBG12	50	10	10		–	**52.120**	–	51.757	22.2060	52.063	1000	-0.11	0.59
Average						50.096		50.365		**50.678**		1.09	0.61

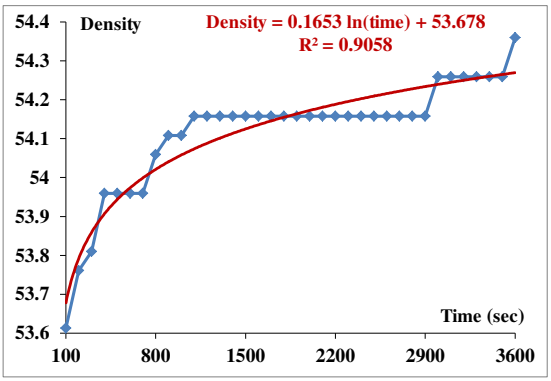


Fig. 3. Estimation of the evolution of the best density obtained with the time by algorithm LF2 on instance SYS4 ($n = 45$) spheres. The regression is based on a logarithmic function, $R^2 = 90.58\%$.

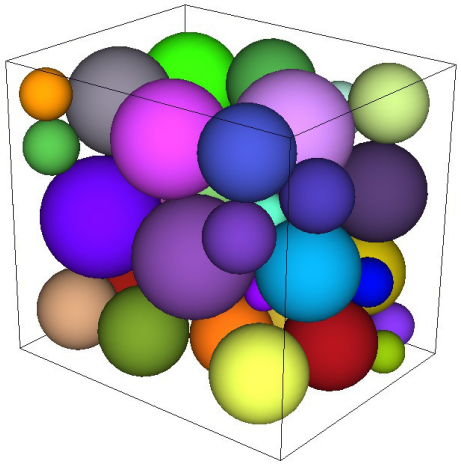


Fig. 4. Solution obtained by algorithm LF2 on instance SYS4 ($n = 45$, $L = 10.7653$ ($Density = 54.410\%$)).

KBTG9, and KBG12). Finally, the last row of the table (column "Imp. 0.8") indicates that algorithm LF2 improves B1.6 with 1.09% in average.

Table I indicates also that algorithm LF2 improves B1.6 with $\tau = 1$ in 12 cases. The two algorithm reach the optimal solution on instances KBG2, KBG4, and KBG10. And algorithm B1.6 with $\tau = 1$ is better than LF2 on instances KBTG5, KBTG9, and KBG11) but the percentage of improvement has decreased to 0.61%.

Fig. 3 indicates the evolution of the density of the solution according to the computation time on instance SYS4 (n=45 spheres). The evolution follows a logarithmic function with a coefficient of determination $R^2 > 90\%$. This means that the density of the obtained solution begins by increasing quickly since the length is near to the upper bound $L_{\max}$ and it is then easier to compute a feasible solution. The improvement of the density slows down when the length approaches the lower bound $L_{\min}$.

Fig. 4 gives the solution obtained by the proposed algorithm LF2 on instance SYS4 that has 35 spheres. The best length obtained is equal to 10.7653 (the best value obtained by B1.6 was 10.8760), this corresponds to an improvement of 1.02%.

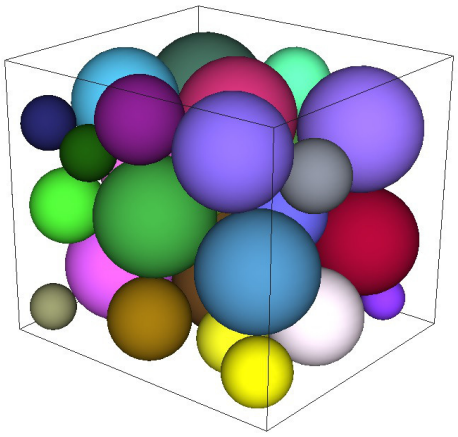


Fig. 5. Solution obtained by algorithm LF2 on instance KBG1 ($n = 30$, $Density = 54.494\%$).

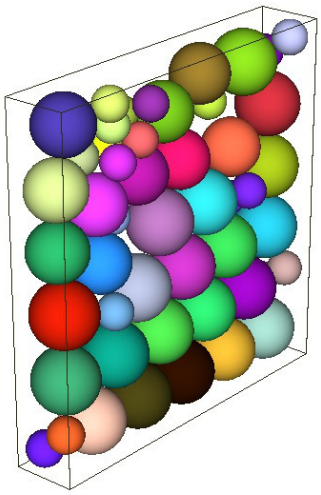


Fig. 6. Optimal solution obtained by algorithm LF2 on instance KBG10 ($n = 50$, $Density = 51.866\%$).

Fig. 5 displays the solution obtained by algorithm LF2 on instance KBG1 that contains 30 spheres. The density obtained is equal to 54.494% (the best value obtained by B1.6 was 54.096%), so the improvement obtained is equal to 0.73%.

Finally, Fig. 6 displays the optimal solution obtained by algorithm LF2 on instance KBG10 that contains 50 spheres but where the number of different radii is only 5. The optimal density is in this case equal to 51.866% and the corresponding optimal length is equal to $2 \times r_{\max}$, i.e., $L = 1.810$, where $r_{\max} = 0.905$ is the greatest radius in the instance.

V. Conclusion

In this paper, a look-forward heuristic was proposed in order to solve the problem of packing spheres into a three-dimensional bin. The first novelty is that method starts with an empty configuration instead of placing one or several pieces inside the container. The second difference is that the interval search proceeds by decreasing the value of the length of the bin instead of using a dichotomous search, the objective is to escape from local optima. Finally the look-forward procedure uses a double search (two levels) instead of one level.

The obtained results on the tested instances showed that the proposed method is effective since it has succeeded to improve or reach almost all the best known results published in the literature. As a future work, it will be interesting to design a new heuristic for packing weakly heterogeneous spheres because it is well-known that the MHD heuristic was designed for packing strongly heterogeneous circles and spheres.

References

[1] H. Akeb, M. Hifi, and D. Lazure, "An Improved Algorithm for the Strip Packing Problem," Proceedings of the Federated Conference on Computer Science and Information Systems, FedCSIS 2012, Wroclaw, Poland, pp. 357–364. ISBN 978-83-60810-51-4. https://fedcsis.org/proceedings/2012/pliks/93.pdf

[2] E.G. Birgin and F.N.C. Sobral, "Minimizing the object dimensions in circle and sphere packing problems," *Compt. Oper. Res.*, vol. 35, 2008, pp. 2357–2375. http://dx.doi.org/10.1016/j.cor.2006.11.002

[3] E. O. Gavriliouk, "Unequal sphere packing problem in the context of stereotactic radiosurgery," Shaker Verlag, 2007, 96 pages.

[4] W. Q. Huang, Y. Li, H. Akeb, and C. M. Li, Y., "Greedy algorithms for packing unequal circles into a rectangular container," *J. Oper. Res. Soc.,* vol. 56, 2005, pp. 539–548. http://dx.doi.org/10.1057/palgrave.jors.2601836

[5] T. Kubach, A. Bortfeldt, T. Tilli, and H. Gehring, "Greedy algorithms for packing unequal sphere into a cuboidal strip or a cuboid," *Asia Pac. J Oper. Res.*, vol. 28, 2011, pp. 739–753. http://dx.doi.org/10.1142/S0217595911003326

[6] L.K. Lenstra and A.H.G Rinnooy Kan, "Complexity of packing, covering, and partitioning problems," In Schrijver A (ed.), Packing and Covering in Combinatorics. Amsterdam: Mathematisch Centrum, 1979, pp. 275–291.

[7] M. Lin, R. Chen, and J. S. Liu, "Lookahead Strategies for Sequential Monte Carlo," *Statist. Sci.*, vol 28, 2013, pp. 69–94. http://dx.doi.org/10.1214/12-STS401

[8] Y. Li and W. Ji, "Stability and convergence analysis of a dynamics-based collective method for random sphere packing," *J. Comput. Phys.*, vol. 250, 2013, pp. 373–387. http://dx.doi.org/10.1016/j.jcp.2013.05.023

[9] R. M'Hallah, A. Alkandari, and N. Mladenović, "Packing unit spheres into the smallest sphere using VNS and NLP," *Compt. Oper. Res.*, vol. 40, 2013, pp. 603–615. http://dx.doi.org/10.1016/j.cor.2012.08.019

[10] R. M'Hallah and A. Alkandari, "Packing unit spheres into a cube using VNS," *Electron. Notes Discrete Math.*, vol. 39, 2012, pp. 201–208. http://dx.doi.org/10.1016/j.endm.2012.10.027

[11] K. Soontrapa and Y. Chen, "Mono-sized sphere packing algorithm development using optimized Monte Carlo technique," *Adv. Powder Technol.*, vol. 24(6), 2013, pp. 955–961. http://dx.doi.org/10.1016/j.apt.2013.01.007

[12] Y. Stoyan, G. Yaskow, and G. Scheithauer, "Packing of various radii solid spheres into a parallelepiped," *Cent. Europ. J. Oper. Re.*, vol. 11, 2003, pp. 389–407.

[13] A. Sutou and Y. Dai, "Global optimization approach to unequal sphere packing problems in 3D," *J. Optimiz. Theory App.*, vol. 114, 2002, pp. 671–694. http://dx.doi.org/10.1023/A:1016083231326

[14] W. Visscher and M. Bolsterli, "Random packing of equal and unequal spheres in two and three dimensions," *Nature*, vol. 239, 1972, pp. 504–507. http://dx.doi.org/10.1038/239504a0

[15] J. Wang, "Packing of unequal spheres and automated radiosurgical treatment planning," *J. Comb. Optim.*, vol. 3, 1999, pp. 453–463. http://dx.doi.org/10.1023/A:1009831621621

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 405–411
DOI: 10.15439/2014F31
ACSIS, Vol. 2

3-D filter SQP method for optimal control of the multistage differential-algebraic systems with inconsistent initial values

Paweł Drąg
Institute of Computer Engineering, Control and Robotics
Wrocław University of Technology
Janiszewskiego 11-17, 50-372 Wrocław, Poland
Email: pawel.drag@pwr.edu.pl

Krystyn Styczeń
Institute of Computer Engineering, Control and Robotics,
Wrocław University of Technology
Janiszewskiego 11-17, 50-372 Wrocław, Poland
Email: krystyn.styczen@pwr.edu.pl

***Abstract*—In the article a 3-dimensional filter method for solving optimal control problems of differential-algebraic equations (DAEs) was presented. Direct multiple shooting method, which is appropriate for the control problems of the multistage DAE systems, leads to the large-scale nonlinear programming problems. In the proposed approach the extended Fletcher's filter with three inputs was used. The filter method promotes global convergence without the need to use a penalty function. The first input of the filter denotes the value of the cost function. The second and third inputs come from two types of equality constraints - consistent initial conditions of the DAE system and continuity constraints on the state trajectories. The new algorithm was tested on the optimal control problem of a fed-batch fermentor for penicillin production. The numerical simulations were executed in MATLAB environment using Wroclaw Center for Networking and Supercomputing.**

***Keywords*—optimal control, DAE systems, inconsistent initial conditions, filter algorithm, nonlinear programming.**

I. Introduction

CONTROL and optimization of the complex and multistage differential-algebraic systems (DAEs) play a key role in a lot of technological systems. The dynamical behavior of the processes can be described by the differential equations. But conservation laws, balance equations, boundary conditions as well as interface with the environmental signals are modeled using the algebraic equations. Today, the differential-algebraic equations are one of the most elegant and simple ways to model a physical system, because they allow the creation of separate models for subcomponents that can then be pasted together [3], [4].

After some elimination processes, DAE systems can be rewritten in the form of ordinary differential equations, which can not present the nature of the process in the same manner like DAEs. A few advantages of a DAE formulation are the following: (1) it may be difficult to reformulate the problem as an ODE when nonlinearities are present, (2) the algebraic equations typically describe conservation laws or explicit equality constraints and they should be kept invariant, (3) it is easier to vary design parameters in an implicit model, (4) the implicit model does not require the modeling simplifications often necessary to get an ODE, (5) the variable keep their original physical interpretation, (6) the system structure can be exploited by problem-specific solvers, (7) less specialized mathematical expertise is required on the part of the designer [5].

One group of the approaches to the optimal control of complex dynamical processes are direct methods, which reformulate the original infinite dimensional optimization problem as a finite dimensional nonlinear programming (NLP) problem. In direct methods, the control and both differential and algebraic states are parametrized. Direct multiple shooting is one of the most popular direct methods. It enables using the efficient DAE solvers to calculate the function values and derivatives accurately. Since the integrations are decoupled on different multiple shooting intervals, this method is well suited for parallel computing. In this manner the control and optimization of the unstable dynamical modes can be considered. The approach allows an effective treatment of control and state path constraints [7]. Using the multiple shooting methods results in a large-scale NLP problem.

In this paper a tri-dimensional filter method based on the line search technique is considered. The filter method to solve nonlinear programming problem can be seen as an alternative to the traditional merit function approach. In this method, compared to the traditional penalty function methods, in which adjustment of the penalty parameter can be problematic, may make the trial steps accepted more easily.

The first idea of the filter method was to interpret the NLP problem as a bi-objective optimization problem with two conflicting purposes. The objective function had to be minimized, but the constraint violation should be minimized too. In the method presented in [10], all the constraints violations were added together and only one constraint violation was defined.

However, each constraint may have its own behavior. Some constraints can be highly nonlinear, while some others are linear or nearly linear [17]. There is the other situation, when the constraints can be grouped depending on the role in the mathematical model.

In the article the constraints were splited into equality constraints for consistent initial conditions for DAE model and equality constraints, which measure the discontinuity of the differential state trajectories. Thus the filter consists of three values: value of the objective function, equality constraints,

978-83-60810-58-3/$25.00

which measure inconsistency of the initial conditions of DAE model, and equality constraints for continuity of the differential state trajectories.

The article is constructed as follows. In the 2nd section the multistage optimal control problem of differential-algebraic systems was presented. Then, in 3rd section, the 3-D filter algorithm was presented. The results of the numerical simulations were discussed in the 4th section.

In the article was used the same notation as in [7] and [20].

II. Problem statement

Let us consider the problem of optimal control of the process with the performance cost function

$$\min_{(u(t),x(t),z(t),p)} \int_{t_0}^{t_f} L\big(x(t),z(t),u(t),p\big)dt + E\big(x(t_f)\big), \quad (1)$$

subject to a system of the index-one differential-algebraic equations (DAE) [6]

$$\begin{aligned} B(\cdot)\dot{x}(t) = &\quad f\big(x(t),z(t),u(t),p\big) \\ 0 = &\quad g\big(x(t),z(t),u(t),p\big), \end{aligned} \quad (2)$$

where x and z denote the differential and algebraic state variables, respectively, u is the vector valued control function, whereas p is a vector of system parameters, which does not depend on the time. Matrix $B\big(x(t),z(t),u(t),p\big)$ is assumed to be invertible. Then the DAE is in a semi-explicit form.

The initial values for the differential and algebraic states and values for the system parameters are prescribed

$$x(t_0) = x_0, \quad (3)$$

$$p(t_0) = p_0. \quad (4)$$

In addition, the terminal constraints

$$r_1\big(x(t_f),p\big) = 0, \quad r_2\big(x(t_f),p\big) \geq 0, \quad (5)$$

as well as the state and control inequality constraints

$$h(x(t),z(t),u(t),p) \geq 0 \quad (6)$$

have to be satisfied.

There is a quite other situation, when the multistage DAE system is considered, because each stage can be described by other set of the differential-algebraic equations.

Let us assume, that there are N stages in the complex industrial process and there is an independent variable t, for example time or length of the chemical reactor.

For a suitable partition of the time horizon $[t_0,t_f]$ into N subintervals $[t_i,t_{i+1}]$ with

$$t_0 < t_1 < \cdots < t_N = t_f, \quad (7)$$

the control function $u(t)$ is discretized. It could be represented by a piecewise constant, piecewise linear or a polynomial approximation [19]. If the control function is parametrized as a piecewise constant vector function, then

$$u(t) = u^l \quad (8)$$

for $t \in [t_{l-1},t_l], \qquad l = 1,\cdots,N.$

By the multiple shooting method, the DAE is parametrized in some sense too. The solution of the DAE system is decoupled on the N intervals $[t_l,t_{l+1}]$. In this manner it introduces the initial values s_x^l and s_z^l of the differential and algebraic states at times t_i as the additional optimization variables.

The trajectories $x(t)$ and $z(t)$ are obtained as a sum of trajectories $x^l(t)$ and $z^l(t)$ on each interval $[t_{l-1},t_l]$. The trajectories $x^l(t)$ and $z^l(t)$ are the solutions of an initial value problem

$$\begin{aligned} B^l(\cdot)\dot{x}(t) = &\quad f^l\big(x^l(t),z^l(t),u^l(t),p\big) \\ 0 = &\quad g^l\big(x^l(t),z^l(t),u^l(t),p\big) + \alpha^l(t_l)g^l\big(s_x^l,s_z^l,u^l,p\big) \\ t \in &\quad [t_{l-1},t_l], \qquad l = 1,\cdots,N. \end{aligned} \quad (9)$$

The relaxation parameter $\alpha^l(t_l)$ was introduced to allow an efficient DAE solution for the initial values and controls s_x^l, s_z^l, u^l, that may temporarily violate the consistency conditions. In this manner, the trajectories $x^l(t)$ and $z^l(t)$ on the interval $[t_{l-1},t_l]$ are functions of the initial values, controls and parameters s_x^l, s_z^l, u^l, p.

The integral part of the cost function is evaluated on each interval independently

$$\begin{aligned} &\min_{\substack{s_x^l,s_z^l,u^l,p \\ l=1,\cdots,N}} \int_{t_0}^{t_1} L^1\big(x^1(t),z^1(t),u^1(t),p\big)dt + \cdots + \\ &+ \int_{t_{N-1}}^{t_N} L^N\big(x^N(t),z^N(t),u^N(t),p\big)dt + E\big(x(t_N)\big) = \\ &= \min_{\substack{s_x^l,s_z^l,u^l,p \\ l=1,\cdots,N}} \sum_{l=1}^{N} \int_{t_{l-1}}^{t_l} L^l\big(x^l(t),z^l(t),u^l(t),p\big)dt + \\ &+ E\big(x(t_N)\big). \end{aligned} \quad (10)$$

The parametrization of the optimal control problem of the multistage DAE systems using the multiple shooting approach and a piecewise constant control representation leads to the following nonlinear programming problem

$$\begin{aligned} &\min_{\substack{s_x^l,s_z^l,u^l,p \\ l=1,\cdots,N}} \sum_{l=1}^{N} \int_{t_{l-1}}^{t_l} L^l\big(x^l(t),z^l(t),u^l(t),p\big)dt + \\ &+ E\big(x(t_N)\big) = \min_{\chi} \Phi(\chi), \end{aligned} \quad (11)$$

subject to the continuity conditions

$$s_x^l = x^{l-1}(t_{l-1}), \qquad l = 2,\cdots,N, \quad (12)$$

the consistency conditions

$$0 = g^l(s_x^l,s_z^l,u^l,p), \qquad l = 1,\cdots,N, \quad (13)$$

control and path constraints imposed pointwise at the multiple shooting nodes

$$h^l(s_x^l,s_z^l,u^l,p) \geq 0, \qquad l = 1,\cdots,N, \quad (14)$$

the terminal constraints

$$r_1(s_x^l,s_z^l,p) = 0, \quad r_2(s_x^l,s_z^l,p) \geq 0, \quad (15)$$

lower and upper bounds on the decision variables

$$\chi_L \leq \chi \leq \chi_U, \tag{16}$$

$$\chi = [s_x^1, \cdots, s_x^N, s_z^1, \cdots, s_z^N, u^1, \cdots, u^N, p]^T, \tag{17}$$

$$\chi_L = [s_{x,L}^1, \cdots, s_{x,L}^N, s_{z,L}^1, \cdots, s_{z,L}^N, u_L^1, \cdots, u_L^N, p_L]^T, \tag{18}$$

$$\chi_U = [s_{x,U}^1, \cdots, s_{x,U}^N, s_{z,U}^1, \cdots, s_{z,U}^N, u_U^1, \cdots, u_U^N, p_U]^T, \tag{19}$$

and with the DAE system in each interval

$$\begin{aligned} B^l(\cdot)\dot{x}(t) = &\quad f^l\big(x^l(t), z^l(t), u^l(t), p\big) \\ 0 = &\quad g^l\big(x^l(t), z^l(t), u^l(t), p\big) + \alpha^l(t_l) g^l(s_x^l, s_z^l, u^l, p), \\ t \in &\quad [t_{l-1}, t_l], \qquad l = 1, \cdots, N. \end{aligned} \tag{20}$$

III. The filter method

About 10 years ago, Fletcher and Leyffer proposed the filter methods to solve nonlinear programming (NLP) as an alternative to the traditional merit function approach. The underlying concept of filter is quite simple, being based on the multi-objective optimization, that is, the trial point is accepted provided there is a sufficient decrease of the objective function or the constraint violation function. In addition, the computational results presented in [10] were also very encouraging. The trust region filter sequential quadratic programming (SQP) methods have been studied in [9], [11]. On the other hand, the filters approach has been used also in conjunction with the line search strategy [20], [21], with interior point methods [18] and with the pattern search method [1]. Finally, the multidimensional filters have been employed to solve least squares problems, nonlinear equations and unconstrained optimization problems [12], [13].

In this paper, a tri-dimensional filter method based on the line search strategy, was proposed. The main idea of a filter is to interpret the NLP problem as a bi-objective optimization problem with two conflicting purposes: minimizing the objective function and the constraints violation. So, the formal filter in [10] consisted of two parts: the value of the objective function and the constraint violation. It means, that all the constraints are considered together and only one constraint violation is defined. However, each constraint may have its own behavior. For example, some constraints may be highly nonlinear, while some others are linear or nearly linear. In this work only equality constraints are considered. But they are two kinds of constraints, which have definitely a different meaning and applications. Thus, the new filter consists of three inputs: objective function value, inconsistency of the initial conditions for differential-algebraic equations and continuity of the differential state trajectories.

A. A line search filter approach

As it was assumed, the considered problem is stated as

$$\min_{\chi \in \mathcal{R}^n} f(\chi) \tag{21}$$

subject to

$$c(\chi) = 0, \tag{22}$$

where the objective function $f : \mathcal{R}^n \to \mathcal{R}$ and the equality constraints $c : \mathcal{R}^n \to \mathcal{R}^m$ with $m < n$ are sufficiently smooth.

The Karush-Kuhn-Tucker (KKT) for the nonlinear programming problem (21)-(22) are

$$g(\chi) + A(\chi)\lambda = 0, \tag{23}$$

$$c(\chi) = 0, \tag{24}$$

where $A(\chi) = \nabla c^T(\chi)$ denotes the transpose of the Jacobian of the constraints $c(\chi)$ and $g(\chi) = \nabla f(\chi)$ denotes the gradient of the objective function. The vector λ corresponds to the Lagrange multipliers for the equality constraints. Under constraint qualifications assumption, the KKT conditions are the first order optimality conditions for (21)-(22) [15].

Given an initial estimate χ_0, the line search algorithm generates a sequence of improved estimates χ_k of the solution for the NLP. For this purposes in each iteration k a search direction d_k is computed from the linearization at χ_k of the KKT conditions

$$\begin{bmatrix} H_k & A_k \\ A_k^T & 0 \end{bmatrix} \begin{pmatrix} d_k \\ \delta\lambda_k \end{pmatrix} = - \begin{pmatrix} g_k \\ c_k \end{pmatrix}, \tag{25}$$

where $A_k = A(\chi_k)$, $g_k = g(\chi_k)$ and $c_k = c(\chi_k)$.

The symmetric matrix H_k denotes the Hessian $\nabla^2_{\chi\chi}\mathcal{L}(\chi_k, \lambda_k)$ of the Lagrangian

$$\mathcal{L}(\chi, \lambda) = f(\chi) + c^T(\chi)\lambda \tag{26}$$

of the nonlinear programming problem or an approximation of the Hessian.

The vector λ_k is some estimate of the optimal multipliers corresponding to the equality constraints, and $\delta\lambda_k$ in (25) can be used to determine a new estimate λ_{k+1} for the next iteration.

After a search direction d_k has been computed, a step size $\alpha_k \in (0, 1]$ is determined in order to obtain the next iterate

$$\chi_{k+1} = \chi_k + \alpha_k d_k. \tag{27}$$

It would be ideally to guarantee, that the sequence χ_k of iterates converges to a solution of the NLP. So, for this purposes, a backtracking procedure was proposed.

In the backtracking line search procedure a decreasing sequence of step size $\alpha_{k,l} \in (0, 1] \quad (l = 0, 1, 2, \cdots)$ is tried until some acceptance criterion is satisfied. Traditionally, a trial step size $\alpha_{k,l}$ is accepted if the corresponding trial point

$$\chi_k(\alpha_{k,l}) = \chi_k + \alpha_{k,l} d_k \tag{28}$$

provides sufficient reduction of a merit function, such as the exact penalty function

$$\phi_\rho(\chi) = f(\chi) + \rho\theta(\chi), \tag{29}$$

where the infeasibility measure $\theta(\chi)$ was defined as

$$\theta(\chi) = \|c(\chi)\|. \tag{30}$$

Under certain regularity assumptions it can be shown that a feasible strict local minimum of the exact penalty function coincides with a local solution of the NLP if the value of the penalty parameter $\rho > 0$ is chosen sufficiently large.

The overall algorithm for solving the equality constrained NLP problem is as follows.

ALGORITHM 1. The filter line search SQP algorithm [20]

Given: Starting point χ_0; constants $\theta_{\max} \in (\theta(\chi_0), \infty)$; γ_θ, $\gamma_f \in (0,1)$; $\delta > 0$; $\gamma_\alpha \in (0,1]$; $s_\theta > 1$; $s_f \geq 1$; $\eta_f \in (0, \frac{1}{2})$; $0 < \tau_1 < \tau_2 < 1$.

1. Initialize.
   Initialize the filter $\mathcal{F}_0 = \{(\theta, f) \in \mathcal{R}^2 : \theta \geq \ \theta_{max}\}$ and the iteration counter $k \leftarrow 0$.
2. Check convergence.
   Stop if χ_k is a stationary point of the NLP (21)-(22), i.e. if it satisfies the KKT conditions (23)-(24) for some $\lambda \in \mathcal{R}^m$.
3. Compute search direction
   Compute the search direction d_k from the linear system (25). If this system is detected to be ill-conditioned, go to the feasibility restoration phase in step 8.
4. Backtracking line search.
   4.1 Initialize line search.
   Set $\alpha_{k,0} = 1$ and $l \leftarrow 0$.
   4.2 Compute new trial point.
   If the trail step size becomes too small, i.e., $\alpha_{k,l} < \alpha_k^{\min}$, go to the feasibility restoration phase in step 8. Otherwise, compute the new trail point $\chi_k(\alpha_{k,l}) = \chi_k + \alpha_{k,l} d_k$.
   4.3 Check acceptability to the filter.
   If $\chi_k(\alpha_{k,l}) \in \mathcal{F}_k$, reject the trial step size and go to step 4.5.
   4.4 Check sufficient decrease with respect to current iterate.
   4.5 Choose new trial step size
   Choose $\alpha_{k,l+1} \in [\tau_1 \alpha_{k,l}, \tau_2 \alpha_{k,l}]$, set $l \leftarrow l + 1$, and go back to step 4.2.
5. Accept trial point.
   Set $\alpha_k = \alpha_{k,l}$ and $\chi_{k+1} = \chi_k(\alpha_k)$.
6. Augment filter if necessary.
7. Continue with next iteration.
   Increase the iteration counter $k \leftarrow k + 1$ and go back to step 2.
8. Feasibility restoration phase.
   Compute a new iterate χ_{k+1} by decreasing the infeasibility measure θ so that χ_{k+1} satisfies the sufficient decrease conditions (31)-(32) and is acceptable to the filter, i.e., $(\theta(\chi_{k+1}), f(\chi_{k+1})) \notin \mathcal{F}_k$.
   Augment the filter and continue with the regular iteration in step 7.

Line search methods that use a merit function ensure sufficient progress toward the solution. Hence, here it is required that the next iterate provides at least as much progress in one of the measures θ or f that corresponds to a small fraction of the current constraint violation, $\theta(\chi_k)$. It means, that for fixed constants $\gamma_\theta, \gamma_f \in (0,1)$ a trial step size $\alpha_{k,l}$ provides sufficient reduction with respect to the current iterate χ_k if

$$\theta(\chi_k(\alpha_{k,l})) \leq (1 - \gamma_\theta)\theta(\chi_k) \tag{31}$$

or

$$f(\chi_k(\alpha_{k,l})) \leq f(\chi_k) - \gamma_f \theta(\chi_k). \tag{32}$$

In a practical implementation, the constants γ_θ, γ_f typically are chosen to be small.

B. A multidimensional filter

The multidimensional filter algorithm was stated by Gould, Leyffer and Toint in the article [12]. It was used for solving nonlinear equations and nonlinear least-squares.

The following system of nonlinear equations is considered

$$c(\chi) = 0, \tag{33}$$

where c is twice continously differentiable function from $\mathcal{R}^n$ into $\mathcal{R}^m$. In the next step, the equation (33) is partitioned into p sets $\{c_i(\chi)\}_{i \in \mathcal{I}_j}$ for $j = 1, \cdots, p$, with $\{1, \cdots, n\} = \mathcal{I}_1 \bigcup \mathcal{I}_2 \bigcup \cdots \mathcal{I}_p$ and

$$\theta_j(\chi) = \|c_{\mathcal{I}_j}\| \tag{34}$$

for $j = 1, \cdots, p$, where $\|\cdot\|$ is the Euclidean norm and $c_{\mathcal{I}_j}$ is the vector whose components are the components of c indexed by $\mathcal{I}_j$.

The point is therefore a solution of (33) if and only if

$$\theta_j(\chi) = 0 \tag{35}$$

for $j = 1, \cdots, p$. The quantity $\theta_j(\chi)$ may be interpreted as the size of the residual of the jth set of equations at the point χ.

The classical approach for solving (33) is to minimize a merit function involving some norm of the residual

$$\min_{\chi \in \mathcal{R}^n} f(\chi) = \frac{1}{2}\|\theta(\chi)\|^2. \tag{36}$$

The main idea of filter algorithms for constrained optimization is that new iterates of the underlying iterative algorithm can be accepted if they do not perform, compared to past iterates kept in the filter, worse on both important and typically conflicting accounts for this type of problem: feasibility and low objective function value.

In the context of nonlinear equations, one may consider driving each of the $\{\theta_i(\chi)\}_{i=1}^p$ to zero as an independent task.

A point χ_1 dominates a point χ_2 whenever

$$\forall j = 1, \cdots, p \qquad \theta_j(\chi_1) \leq \theta_j(\chi_2). \tag{37}$$

Thus, if iterate χ_{k_1} dominates iterate χ_{k_2}, the latter is of no real interest, since χ_{k_1} is at least as good as χ_{k_2} for each of the equation sets. All, what is needed to do now, is to remember iterates that are not dominated by other iterates using a structure called a filter.

A filter is a list $\mathcal{F}$ of p-tuples of the form $(\theta_{1,k}, \cdots, \theta_{p,k})$ such that

$$\theta_{j,k} < \theta_{j,l} \tag{38}$$

for at least one $j \in \{1, \cdots, p\}$ and $k \neq l$.

Filter methods propose to accept a new trial iterate χ_k^+ if it is not dominated by any other iterate in the filter and χ_k.

Additionally, it is inappropriate to accept a new point χ_k^+ if $\theta(\chi_k)$ is arbitraily close to being dominated by another point

already in the filter [16]. To avoid this situation, the acceptability condition should be more strength. More formally, one can say that a new trial point χ_k^+ is acceptable for the filter $\mathcal{F}$ if and only if

$$\forall \theta_l \in \mathcal{F} \quad \exists j \in \{1, \cdots, p\} \quad \theta_j(\chi_k^+) < \theta_{j,l} - \gamma_\theta \delta(\|\theta_l\|, \|\theta_k^+\|), \tag{39}$$

where $\gamma_\theta \in (0, 1/\sqrt{p})$ is a small positive constant and where $\delta(\cdot,\cdot)$ is one of the following

$$\delta(\|\theta_l\|, \|\theta_k^+\|) = \|\theta_l\| \tag{40}$$

$$\delta(\|\theta_l\|, \|\theta_k^+\|) = \|\theta_k^+\| \tag{41}$$

or

$$\delta(\|\theta_l\|, \|\theta_k^+\|) = \min(\|\theta_l\|, \|\theta_k^+\|). \tag{42}$$

C. A tri-dimensional filter SQP algorithm

The groups of constraints violations for NLP problem (21)-(22) are defined as follows

$$\mathcal{S}_I = \frac{1}{2} \sum_{i=1}^{p} c_i^2(\chi) \tag{43}$$

for continuity of the differential state trajectories and

$$\mathcal{S}_{II} = \frac{1}{2} \sum_{i=p+1}^{m} c_i^2(\chi), \tag{44}$$

which represents inconsistency of the initial conditions.

These two groups of constraints are defined similarly, but they have definitely different meanings and play another role in the optimization process.

In a filter $\mathcal{F}$, triples of values $(\mathcal{S}_I(\chi), \mathcal{S}_{II}(\chi), f(\chi))$ are considered.

Definition 1. The iterate χ_k dominates the iterate χ_l if and only if $\mathcal{S}_I(\chi_k) \leq \mathcal{S}_I(\chi_l)$, $\mathcal{S}_{II}(\chi_k) \leq \mathcal{S}_{II}(\chi_l)$ and $f(\chi_k) \leq f(\chi_l)$. It is denoted by $\chi_k \preceq \chi_l$.

Thus, if $\chi_k \preceq \chi_l$, the latter is of no real interest, since χ_k is at least as good as χ_l with respect to three violations. Furthermore, if $\chi_k \preceq \chi_l$, one can say that the triple $(\mathcal{S}_I(\chi_k), \mathcal{S}_{II}(\chi_k), f(\chi_k))$ dominates the triple $(\mathcal{S}_I(\chi_l), \mathcal{S}_{II}(\chi_l), f(\chi_l))$.

Definition 2. The kth filter is a list of triples $\{\mathcal{S}_I(\chi_l), \mathcal{S}_{II}(\chi_l), f(\chi_l))\}_{l<k}$, such that no triple dominates any other.

Let $\mathcal{F}_k$ denote the indices in the kth filter

$$\mathcal{F}_k = \{l < k : \chi_j \not\preceq \chi_l \quad \forall j \in \{0, 1, 2, \cdots, k-1\} \backslash \{l\}\} \tag{45}$$

Filter methods accept a trial point $\chi_{k+1} = \chi_k + \alpha d_k$ if its corresponding triple $(\mathcal{S}_I(\chi_{k+1}), \mathcal{S}_{II}(\chi_{k+1}), f(\chi_{k+1}))$ is not dominated by any other triple in the kth filter, neither the triple corresponding to χ_k, i.e. $(\mathcal{S}_I(\chi_k), \mathcal{S}_{II}(\chi_k), f(\chi_k))$

Definition 3. A new trial point χ_{k+1} is said to be "acceptable to the kth filter" if χ_{k+1} is acceptable to χ_l for all $l \in \mathcal{F}_k$.

In this manner defined 3-dimensional line search-SQP filter was tested on optimal control problem of nonlinear differential-algebraic system with inconsistent initial conditions.

IV. Case study: optimal control of a fed-batch fermentor for penicillin production

This problem considers a fed-batch reactor for the production of penicillin [2]. We consider here the free terminal time version where the objective is to maximize the amount of penicillin using the feed rate as the control variable. The mathematical statement of the free terminal time problem is as follows.

Find $u(t)$ and t_f over $t \in [t_0, t_f]$ to maximize

$$J = x_2(t_f) \cdot x_4(t_f) \tag{46}$$

subject to differential-algebraic system

$$\frac{dx_1}{dt} = h_1 x_1 - u\left(\frac{x_1}{500 x_4}\right), \tag{47}$$

$$\frac{dx_2}{dt} = h_2 x_1 - 0.01 x_2 - u\left(\frac{x_2}{500 x_4}\right), \tag{48}$$

$$\frac{dx_3}{dt} = -h_1 \frac{x_1}{0.47} - h_2 \frac{x_1}{1.2} - x_1 \frac{0.029 x_3}{0.0001 + x_3} + \frac{u}{x_4}\left(1 - \frac{x_3}{500}\right), \tag{49}$$

$$\frac{dx_4}{dt} = \frac{u}{500}, \tag{50}$$

$$h_1 = 0.11\left(\frac{x_3}{0.006 x_1 + x_3}\right), \tag{51}$$

$$h_2 = 0.0055\left(\frac{x_3}{0.0001 + x_3(1 + 10 x_3)}\right), \tag{52}$$

where x_1, x_2 and x_3 are the biomass, penicillin and substrate concentration (g/L), and x_4 is the volume (L). The initial conditions are

$$x(t_0) = [1.5 \quad 0 \quad 0 \quad 7]^T. \tag{53}$$

There are several path constraints for state variables

$$0 \leq x_1 \leq 40, \tag{54}$$

$$0 \leq x_2 \leq 25, \tag{55}$$

$$0 \leq x_3 \leq 10. \tag{56}$$

The upper and lower bounds on the only control variable (feed rate of substrate) are

$$0 \leq u \leq 50. \tag{57}$$

The control problem of a fed-batch fermentor for penicillin production was solved with the proposed 3-D SQP filter algorithm combined with multiple shooting method.

At first, the overall time domain was divided into 20 equidistant intervals. It results in 20 differential-algebraic submodels, each of them consists of 4 differential equations and 2 algebraic equations. Initial conditions only for the first stage are known. So, there are 76 decision variables connected with initial values for differential variables and 40 variables, which represent pointwise values of algebraic states. The last decision variable was the duration time of the process.

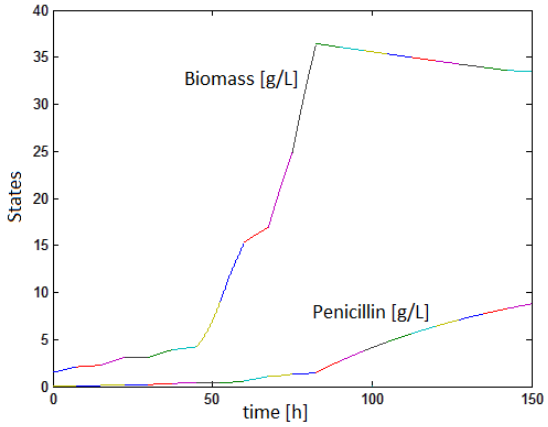


Fig. 1. The optimal trajectories of the both biomass and penicillin concentrations

Backward Differentiation Formula of order 1 was used to solve DAE systems at each stage.

The initial values for decision variables were as follows

$$\chi_{1,x_{1,2}}, \cdots, \chi_{19,x_{1,20}} = 1.5, \tag{58}$$

$$\chi_{20,x_{2,2}}, \cdots, \chi_{38,x_{2,20}} = 0.0, \tag{59}$$

$$\chi_{39,x_{3,2}}, \cdots, \chi_{57,x_{3,20}} = 0.0, \tag{60}$$

$$\chi_{58,x_{4,2}}, \cdots, \chi_{76,x_{4,20}} = 7.0, \tag{61}$$

$$\chi_{77,h_{1,1}}, \cdots, \chi_{96,h_{1,20}} = 10.0, \tag{62}$$

$$\chi_{97,h_{2,1}}, \cdots, \chi_{116,h_{2,20}} = 10.0, \tag{63}$$

$$\chi_{117,u_1}, \cdots, \chi_{136,u_{20}} = 10.0, \tag{64}$$

$$\chi_{137,t_f} = 110.0[h]. \tag{65}$$

The solution, with the accuracy 10^{-6} for each input of the filter, was obtained after 6 hours. The final value of the objective function is $89.5473[g]$. The duration of the whole process is 150 hours. There are the optimal trajectories of the both biomass and penicillin concentrations in Fig. 1.

V. Conclusion

In the article optimal control problem of complex systems with differential-algebraic constraints was considered. One of the most important question in optimization and control of DAE systems concerns on consistent initial conditions. For these purposes the 3-dimensional line search-SQP filter was designed. It enables simultaneous optimization of objective function and treatment of constraints.

The filter algorithm, which consists of three inputs, was tested on the optimal control problem of a fed-batch fermentor for penicillin production. The encouraging results were obtained.

In real-life applications, very often multistage technological processes are under considerations. The multiple shooting method, which enables efficient treatment of such processes, incorporates additional equality constraints and decision variables to the NLP problem, connected with continuity of the differential state trajectories.

The most important advantage of the multiple shooting method is a possibility of control and optimization of highly nonlinear differential-algebraic systems in an open-loop. Application in this field is currently also an important challenge for the presented 3-dimensional SQP filter algorithm [8], [14].

Acknowledgment

The project was supported by the grant of National Science Centre Poland DEC- 2012/07/B/ST7/01216.

References

[1] C. Audet, J.E. Dennis Jr. 2004. A pattern search filter method for nonlinear programming without derivatives. SIAM Journal on Optimization. 14:980-1010, http://dx.doi.org/10.1137/S105262340138983X.

[2] J.R. Banga, E. Balsa-Canto, C.G. Moles, A.A. Alonso. 2005. Dynamic optimization of bioprocesses: Efficient and robust numerical strategies. Journal of Biotechnology. 117:407-419, http://dx.doi.org/10.1016/j.jbiotec.2005.02.013.

[3] J.T. Betts. 2010. Practical Methods for Optimal Control and Estimation Using Nonlinear Programming, Second Edition. SIAM, Philadelphia, http://dx.doi.org/10.1137/1.9780898718577.

[4] L.T. Biegler. 2010. Nonlinear Programming. Concepts, Algorithms and Applications to Chemical Processes. SIAM, Philadelphia, http://dx.doi.org/10.1137/1.9780898719383.

[5] L.T. Bielger, S. Campbell, V. Mehrmann. 2012. DAEs, Control, and Optimization. Control and Optimization with Differential-Algebraic Constraints. SIAM, Philadelphia, http://dx.doi.org/10.1137/9781611972252.ch1.

[6] K.E. Brenan, S.L. Campbell, L.R. Petzold. 1996. Numerical Solution of Initial- Value Problems in Differential-Algebraic Equations. SIAM, Philadelphia, http://dx.doi.org/10.1137/1.9781611971224.

[7] M. Diehl, H.G. Bock, J.P. Schlöder, R. Findeisen, Z. Nagy, F. Allgower. 2002. Real-time optimization and nonlinear model predictive control of processes governed by differential-algebraic equations. Journal of Process Control. 12:577-585, http://dx.doi.org/10.1016/S0959-1524(01)00023-3

[8] P. Drąg, K. Styczeń. 2012. A Two-Step Approach for Optimal Control of Kinetic Batch Reactor with electroneutrality condition. Przegląd Elektrotechniczny. 6:176-180.

[9] R. Fletcher, N.I.M. Gould, S. Leyffer, P.L. Toint, A. Watcher. 2002. Global convergence of a trust region SQP-filter algorithms for general nonlinear programming. SIAM Journal on Optimization. 13:635-659, http://dx.doi.org/10.1137/S1052623499357258.

[10] R. Fletcher, S. Leyffer. 2002. Nonlinear programming without a penalty function. Mathematical Programming. 91:239-269, http://dx.doi.org/10.1007/s101070100244.

[11] R. Fletcher, S. Leyffer, P.L. Toint. 2002. On the global convergence of a Filter-SQP algorithm. SIAM Journal on Optimization. 13:44-59, http://dx.doi.org/10.1137/S105262340038081X.

[12] N.I.M. Gould, S. Leyffer, P.L. Toint. 2004. A multidimensional filter algorithm for nonlinear equations and nonlinear least squares. SIAM Journal on Optimization. 15:17-38, http://dx.doi.org/10.1137/S1052623403422637.

[13] N.I.M. Gould, C. Sainvitu, P.L. Toint. 2005. A filter trust region method for unconstrained optimization. SIAM Journal on Optimization. 16:341-357, http://dx.doi.org/10.1137/040603851.

[14] M. Kwiatkowska. 2012. Antimicrobial PVC composites. Processing technologies and functional properties of polymer nanomaterials for food packaging : International COST Workshop, Wroclaw, Poland, September 11-12, 2012, pp. 40-41.

[15] J. Nocedal, S.J. Wright. 2006. Numerical Optimization. Second Edition. Springer, New York, http://dx.doi.org/10.1007/978-0-387-40065-5

[16] E. Rafajłowicz, K. Styczeń, W. Rafajłowicz. 2012. A modified filter SQP method as a tool for optimal control of nonlinear systems with spatio-temporal dynamics. Int. J. Appl. Math. Comput. Sci. 22:313-326, http://dx.doi.org/10.2478/v10006-012-0023-8.

[17] C. Shen, W. Xue, D. Pu. 2009. Global convergence of a tri-dimensional filter SQP algorithm based on the line search method. Applied Numerical Mathematics. ‘59:235-250, http://dx.doi.org/10.1016/j.apnum.2008.01.005.

[18] M. Ulbrich, S. Ulbrich, L.N. Vicente. 2004. A globally convergent primal-dual interior filter method for nonconvex nonlinear programming. Mathematical Programming. 100:379-410, http://dx.doi.org/10.1007/s10107-003-0477-4.

[19] V.S Vassiliadis, R.W.H. Sargent, C.C. Pantelides. 1994. Solution of a Class of Multistage Dynamic Optimization Problems. 1. Problems without Path Constraints. Ind. Eng. Chem. Res. 33:2111-2122, http://dx.doi.org/10.1021/ie00033a014.

[20] A. Wätcher, L.T. Biegler. 2005. Line search filter methods for nonlinear programming: Motivation and global convergence. SIAM Journal on Optimization. 16: 1-31, http://dx.doi.org/10.1137/S1052623403426556.

[21] A. Wätcher, L.T. Biegler. 2005. Line search filter methods for nonlinear programming: Local convergence. SIAM Journal on Optimization. 16:32-48, http://dx.doi.org/10.1137/S1052623403426544.

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 413–420

DOI: 10.15439/2014F373
ACSIS, Vol. 2

Hybrid GA-ACO Algorithm for a Model Parameters Identification Problem

Stefka Fidanova
Institute of Information and
Communication Technology
Bulgarian Academy of Science,
Acad. G. Bonchev Str., bl. 25A,
1113 Sofia, Bulgaria
E-mail: stefka@parallel.bas.bg

Marcin Paprzycki
System Research Institute
Polish Academy of Sciences
Warsaw and Warsaw Management Academy
Warsaw, Poland
E-mail: marcin.paprzycki@ibspan.waw.pl

Olympia Roeva
Institute of Biophysics and
Biomedical Engineering
Bulgarian Academy of Science,
Acad. G. Bonchev Str., bl. 105,
1113 Sofia, Bulgaria
E-mail: olympia@biomed.bas.bg

***Abstract*—In this paper, a hybrid scheme, to solve optimization problems, using a Genetic Algorithm (GA) and an Ant Colony Optimization (ACO) is introduced. In the hybrid GA-ACO approach, the GA is used to find a feasible solutions to the considered optimization problem. Next, the ACO exploits the information gathered by the GA. This process obtains a solution, which is at least as good as—but usually better than—the best solution devised by the GA. To demonstrate the usefulness of the presented approach, the hybrid scheme is applied to the parameter identification problem in the *E. coli* MC4110 fed-batch fermentation process model. Moreover, a comparison with both the conventional GA and the stand-alone ACO is presented. The results show that the hybrid GA-ACO takes the advantages of both the GA and the ACO, thus enhancing the overall search ability and computational efficiency of the solution method.**

***Index Terms*—Genetic Algorithm; Genetic algorithms; Ant Colony Optimization; hybrid; model parameter identification; *E. coli*; fed-batch fermentation process**

I. Introduction

TO SOLVE different optimization problems we can apply various techniques and approaches, namely exact algorithms (Branch-and-Bound, Dynamic Programming, local search techniques) [1], [2], [3], heuristics [5], [6], and meta-heuristics (Genetic Algorithms, Ant Colony Optimization, Particle Swarm Optimization, Simulated Annealin, Tabu Search, etc.) [4], [7], [8]. Today, the use of meta-heuristics has received more and more attention. These methods offer good solutions, even global optima, within reasonable computing time [9]. An even more efficient behavior, and a higher flexibility when dealing with real-world and large-scale problems, can be achieved through a combination of a meta-heuristic with other optimization techniques, the so-called hybrid metaheuristic [7], [13], [12], [14], [20], [21], [19].

The main goal of all hybrid algorithms is to exploit the advantages of different optimization strategies, while avoiding their disadvantages. Choosing an adequate combination of metaheuristic techniques one can achieve a better algorithm performance in solving hard optimization problems. Developing such effective hybrid algorithm requires expertise from different areas of optimization. There are many hybridization techniques that have shown to be successful for different applications [10], [11].

In this paper, we investigate a hybrid metaheuristic method that combines the Genetic Algorithms (GA) and the Ant Colony Optimization (ACO), named GA-ACO. There already exist some applications of the ACO-GA hybrid for several optimization problems. In [15], [16] a hybrid metaheuristic ACO-GA, for the problem of sports competition scheduling is presented. In the proposed algorithm first, the GA generates activity lists, thus providing the initial population for the ACO. Next, the ACO is executed. In the next step, the GA, based on the crossover and mutation operations, generates a new population. Authors of [17] presented a hybrid algorithm in which the ACO and the GA search alternately and cooperatively in the solution space. Test examples show that the hybrid algorithm can be more efficient and robust than the traditional population based heuristic methods. In [18], the problem of medical data classification is discussed. Authors propose a hybrid GA-ACO and show the usefulness of the proposed approach on a number of benchmark real-world medical datasets. For solving NP-hard combinatorial optimization problems, in [22], a novel hybrid algorithm combining the search capabilities of the ACO and the GA is introduced. As a result a faster and better search algorithm capabilities is achieved.

Provoked by the promising results obtained from the use of hybrid GA-ACO algorithms, we propose a hybrid algorithm, i.e. collaborative combination of the GA and the ACO methods for the model parameters optimization of the *E. coli* fermentation process. The effectiveness of the GA and the ACO have already been demonstrated for model parameter optimization considering fed-batch fermentation processes (see, [24]). Moreover, parameter identification of cellular dynamics models has especially become a research field of great interest. Robust and efficient methods for parameter identification are thus of key importance.

The paper is organized as follows. The problem formulation is given in Section 2. The proposed hybrid GA-ACO technique is described in Section 3. The numerical results and a discussion are presented in Section 4. Conclusion remarks are done in Section 5.

978-83-60810-58-3/$25.00

II. Problem Formulation

A. *E. coli* Fed-batch Fermentation Model

The mathematical model of the fed-batch fermentation process of the *E. coli* is presented by the following non-linear differential equation system [28]:

$$\frac{dX}{dt} = \mu X - \frac{F_{in}}{V} X \tag{1}$$

$$\frac{dS}{dt} = -q_S X + \frac{F_{in}}{V}(S_{in} - S) \tag{2}$$

$$\frac{dV}{dt} = F_{in} \tag{3}$$

where

$$\mu = \mu_{max} \frac{S}{k_S + S} \tag{4}$$

$$q_S = \frac{1}{Y_{S/X}} \mu \tag{5}$$

- X is the biomass concentration, [g/l];
- S is the substrate concentration, [g/l];
- F_{in} is the feeding rate, [l/h];
- V is the bioreactor volume, [l];
- S_{in} is the substrate concentration in the feeding solution, [g/l];
- μ and q_S are the specific rate functions, [1/h];
- μ_{max} is the maximum value of the specific growth rate, [1/h];
- k_S is the saturation constant, [g/l];
- $Y_{S/X}$ is the yield coefficient, [-].

For the model parameters identification, experimental data of an *E. coli* MC4110 fed-batch fermentation process can be used. The experiments providing the real-world data were performed in the Institute of Technical Chemistry, University of Hannover, Germany. The detailed description of the fermentation condition and experimental data can be found in [23], [27].

The fed-batch process starts at time t = 6.68 h, after batch phase. The initial liquid volume is 1350 ml. Before inoculation a glucose concentration of 2.5 g/l was established in the medium. Glucose concentration, in the feeding solution is 100 g/l. The temperature was controlled at 35 °C, the pH at 6.9. The stirrer speed was initially set to 900 rpm and later was increased to 1800 rpm, so that the dissolved oxygen concentration was never below 30%. The aeration rate was kept at 275 l/h and the carbon dioxide was measured in the exhaust gas. The process was stopped at time t = 11.54 h.

The bioreactor, as well as the FIA measurement system is shown in Figure 1. The feed rate profile and the dynamics of the measured substrate concentration are presented, respectively in Fgure 2 and Figure 3.

For the considered non-linear mathematical model of the *E. coli* fed-batch fermentation process (Eq. (1) - Eq. (5)) the parameters that should be identified are:

Fig. 1. *E. coli* MC4110 fed-batch fermentation process: bioreactor and FIA measurement system

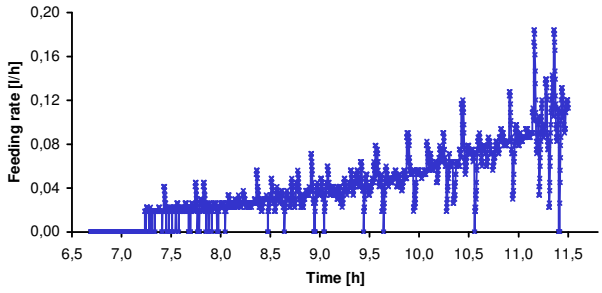


Fig. 2. *E. coli* MC4110 fed-batch fermentation process: feed rate profile

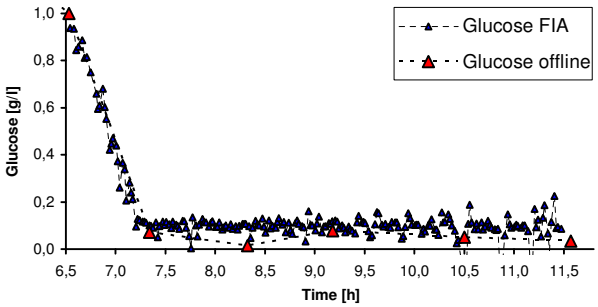


Fig. 3. *E. coli* MC4110 fed-batch fermentation process: measured substrate concentration

- maximum specific growth rate (μ_{max}),
- saturation constant (k_S),
- yield coefficient ($Y_{S/X}$).

The following upper and lower bounds of the model parameters are considered [24]:

$$0 < \mu_{max} < 0.7,$$

$$0 < k_S < 1,$$

$$0 < 1/Y_{S/X} < 30.$$

In the model identification procedures, measurements of main process variables (biomass and glucose concentration) are used. For the on-line glucose determination the FIA system has been employed. For the biomass, an off-line analysis was performed [27].

B. Optimization Criterion

The objective of the optimization consists of adjusting the parameters (μ_{max}, k_S and $Y_{S/X}$) of the non-linear mathematical model function (Eq. (1) - Eq. (5)) to best fit a data set. The objective function is presented as a minimization of a distance measure J between the experimental and the model predicted values of the main state variables (biomass X and substrate S):

$$J = \sum_{i=1}^{m} \left(X_{\exp}(i) - X_{\,\mathrm{mod}}\,(i)\right)^2 + \\ + \sum_{i=1}^{m} \left(S_{\exp}(i) - S_{\,\mathrm{mod}}\,(i)\right)^2 \rightarrow \min \tag{6}$$

where m is the number of experimental data; $X_{\exp}$ and $S_{\exp}$ is the known experimental data for biomass and substrate; $X_{\,\mathrm{mod}}$ and $S_{\,\mathrm{mod}}$ are the model predictions for the biomass and the substrate with a given set of parameters (μ_{max}, k_S and $Y_{S/X}$).

III. Methodology

A. Genetic Algorithm

Genetic Algorithm is a metaheuristic technique based on an analogy with the genetic structure and behaviour of chromosomes within a population of individuals using the following foundations [33]:

- chromosomes in a population compete for resources and mates;
- those chromosomes most successful in each "competition" will produce more off-spring than those chromosomes that perform poorly;
- genes from "good" chromosomes propagate throughout the population so that the two good parents will sometimes produce offspring that are better than either parent;
- thus, each successive generation will become more suited to their environment (will move closer to an optimal solution).

The structure of the GA, shown by the pseudocode, is presented in Figure 4).

Genetic Algorithm
$i = 0$
Initial population $Pop(0)$
Evaluate $Pop(0)$
while (not done) **do** (test for termination criterion)
  $i = i + 1$
  Select $Pop(i)$ from $Pop(i-1)$
  Recombine $Pop(i)$
  Mutate $Pop(i)$
  Evaluate $Pop(i)$
end while
Final solution

Fig. 4. Pseudocode for GA

The GA, mainly operates on binary strings and using a recombination operator with mutation. It is based on a population of chromosomes, $Pop(t) = x_1^t, ..., x_n^t$ for generation t. Each chromosome introduces a potential solution to the problem and is implemented as some data structure S. Each solution is evaluated according its "fitness." Fitness of a chromosome is assigned proportionally to the value of the objective function of the chromosomes. Then, a new population (generation $t + 1$) is formed by selecting better chromosomes (the selection step).

A roulette wheel, developed by Holland [30] is the most often used selection method. The probability, P_i, for each chromosome to be selected is defined by:

$$P[\text{Individual } i \text{ is chosen}] = \frac{F_i}{\sum_{j=1}^{PopSize} F_j}, \tag{7}$$

where F_i equals the fitness of the chromosome i and $PopSize$ is the population size.

Selected members of the new population have been subjected to transformations by means of "genetic" operators to form a new solution. There are unary transformations m_i (mutation type), which create new chromosomes by a small change in a single chromosome ($m_i : S \rightarrow S$), and higher order transformations c_j (crossover type), which create new chromosomes by combining parts from several chromosomes ($c_j : S \times \ldots \times S \rightarrow S$). The combined effect of selection, crossover and mutation gives so-called reproductive scheme growth equation (the schema theorem) [29]:

$$\xi\,(S, t+1) \geq \\ \xi\,(S,t) \cdot eval\,(S,t)\,/\bar{F}\,(t) \left[1 - p_c \cdot \frac{\delta\,(S)}{m-1} - o\,(S) \cdot p_m\right].$$

A good schemata receives an exponentially increasing number of reproductive trials in successive generations.

B. Ant Colony Optimization

The ACO is a stochastic optimization method that mimics the social behavior of real ants colonies, which try to find shortest rout to feeding sources and back. Real ants lay down quantities of pheromone (chemical substance) marking the path that they follow. An isolated ant moves essentially at random but an ant encountering a previously laid pheromone will detect it and decide to follow it with high probability and reinforce it with a further quantity of pheromone. The repetition of the above mechanism represents the auto-catalytic behavior of a real ant colony, where the more ants follow a given trail, the more attractive that trail becomes. Hence, the overall idea of the optimization approach comes from observing such behavior, in which ants are collectively able to find the shortest path to the food.

The ACO is implemented by instantiating a team of software agents, which simulate the ants behavior, walking around the graph representing the problem to solve. The requirements of the ACO algorithm are as follows [25], [26]:

- The problem needs to be represented appropriately, to allow the ants to incrementally update the solutions through the use of a probabilistic transition rules, based on the amount of pheromone on the trail and other problem specific knowledge.
- Existence of a problem-dependent heuristic function that measures the quality of components that can be added to the current partial solution.
- Explication of a set of rules for pheromone updates, which specify how to modify the pheromone value in specific situations.
- A probabilistic transition rule, based on the value of the heuristic function and the pheromone value, that is used to iteratively construct a solution needs to be provided.

The structure of the ACO algorithm, represented as a pseudocode, is depicted in Figure 5. The transition probability $p_{i,j}$, to choose the node j, when the current node is i, is based on the heuristic information $\eta_{i,j}$ and the pheromone trail level $\tau_{i,j}$ of the move, where $i, j = 1, \ldots, n$.

$$p_{i,j} = \frac{\tau_{i,j}^a \eta_{i,j}^b}{\sum\limits_{k \in Unused} \tau_{i,k}^a \eta_{i,k}^b}, \tag{8}$$

where $Unused$ is the set of unused nodes of the graph.

The higher the value of the pheromone and the heuristic information, the more profitable it is to select this move and to continue the search. In the beginning, the initial pheromone level (across the graph) is set to a small positive constant value τ_0; later, the ants update this value after completing the solution construction stage. Different ACO algorithms adopt different criteria to update the pheromone level.

The pheromone trail update rule is given by:

$$\tau_{i,j} \leftarrow \rho \tau_{i,j} + \Delta\tau_{i,j}, \tag{9}$$

where ρ models pheromone evaporation (a process that takes place in the nature) and $\Delta\tau_{i,j}$ is a new added pheromone, which is proportional to the quality of the solution. Thus better solutions will receive more pheromone than others and will be more desirable in a next iteration.

Ant Colony Optimization
Initialize number of ants;
Initialize the ACO parameters;
while not end-condition **do**
  for $k = 0$ **to** number of ants
    ant k choses start node;
    while solution is not constructed **do**
      ant k selects higher probability node;
    end while
  end for
  Update-pheromone-trails;
end while

Fig. 5. Pseudocode for ACO

IV. Hybrid GA-ACO Algorithm

We propose to combine two metaheuristics, namely the GA [29], [30] and the ACO [31]. The GA is a population-based method where initial population is randomly generated. Thus the randomly generated initial solutions are further genetically evaluated. As seen above, the ACO algorithm is a population-based as well. The difference, as compared with the GA, is that the ACO does not need initial population. The ACO is a constructive method, in which the ants look for good solutions guided by the parameter called the pheromone. At the beginning the initial pheromone is the same for the all arcs of the graph representing the problem. After every iteration, the pheromone levels are updated (in all arcs; in arcs traveled by the ant the pheromone level is increasing, while in abandoned arcs it it evaporating). As the result, the elements representing better solutions receive more pheromone than others and become more desirable in a next iteration. In our hybrid algorithm the solutions constructed (proposed) by the GA are treated as solutions achieved by the ACO in some previous iteration, and we use them to specify the initial pheromone level in the solution graph. After that we search for the solution using the ACO algorithm. The structure of the proposed hybrid GA-ACO algorithm is shown by the pseudo-code in Figure 6.

V. Numerical Results and Discussion

The theoretical background of the GA and the ACO is presented in details[24]. For the considered here model problem of parameter identification, we used real-value coded GA instead binary encoding. Therefore the basic operators in the applied GA are as follows:

- encoding – real-value,
- fitness function – linear ranking,
- selection function – roulette wheel selection,
- crossover function – extended intermediate recombination,
- mutation function – real-value mutation,

GA-ACO hybrid algorithm
$i = 0$
Initial population $Pop(0)$
Evaluate $Pop(0)$
while not end-condition **do**
    $i = i + 1$
    Select $Pop(i)$ from $Pop(i-1)$
    Recombine $Pop(i)$
    Mutate $Pop(i)$
    Evaluate $Pop(i)$
end while
Best GA solution for ACO
Initialize number of ants;
Initialize the ACO parameters;
Initialize the pheromone
while not end-condition **do**
    for $k = 0$ **to** number of ants
        ant k choses start node;
        while solution is not constructed **do**
            ant k selects higher probability node;
        end while
    end for
    Update-pheromone-trails;
end while
Final solution

Fig. 6. Pseudocode for Hybrid GA-ACO

- reinsertion – fitness-based.

In the applied ACO algorithm, the problem is represented by graph and the artificial ants try to construct the shortest path (under specified conditions). In our case the graph of the problem is represented by three partity graph. There are not arcs inside a level and there are arcs between (three) levels. Every level corresponds to one of the model parameters we identify (μ_{max}, k_S and $Y_{S/X}$). Every level consists of 1000 vertexes, which corresponds to 1000 uniformly distributed points in the domain (interval) of every one of the considered model parameters. The pheromone is positioned on the arcs. The ants create a solution starting from random node from the first level. They chose nodes from other levels applying the probabilistic rule. In this application the probabilistic rule uses only the pheromone value. We can think that the heuristic information is constant. Thus the ants will prefer the nodes with maximal quantity of the pheromone.

To set the optimal settings of the GA and the ACO algorithms parameters, we performed several runs of the algorithms with varying parameters, according to the considered here optimization problem. The resulting optimal settings of the GA and the ACO parameters are summarized in Table I and in Table II.

The computer, used to run all identification procedures, was an Intel Core i5-2329 3.0 GHz, with 8 GB Memory, Windows 7 (64bit) operating system and Matlab 7.5 environment.

TABLE I
PARAMETERS OF GA

Parameter	Value
ggap	0.97
xovr	0.7
mutr	0.05
maxgen	200
individuals	100
nvar	3
inserted rate	100 %

TABLE II
PARAMETERS OF ACO ALGORITHM

Parameter	Value
number of ants	20
initial pheromone	0.5
evaporation	0.1
generations	200

We performed 30 independent runs of the hybrid GA-ACO. The hybrid algorithm started with population of 20 chromosomes. We used 40 generations to find the initial solution. Next, we took the achieved best GA solution to specify the ACO initial pheromones. Next, the ACO was used to obtain the best model parameters vector using 20 ants for 100 generations (see, Table III).

For comparison of performance of the hybrid algorithm we used the pure GA and the pure ACO. They were run (30 times) with (optimized) parameters shown in Table I and in Table II.

The main numerical results, obtained when solving the parameter identification problem, are summarized in Table IV. In this table we show the best, worst and average values of the objective function achieved by the pure ACO, the pure GA and the hybrid GA-ACO algorithms after 30 run of every one of them, as well as their running times. The obtained average values of the model parameters (μ_{max}, k_S and $Y_{S/X}$) are summarized in Table V.

As it can be seen, from Table IV, the hybrid GA-ACO achieves values of the objective function that are similar to these obtained by the pure GA and the pure ACO algorithms. In the same time, the running time of the proposed hybrid algorithm is about two times shorter. The pure ACO algorithm starts with an equal initial pheromone distribution for all problem elements. In the case of the hybrid GA-ACO we use the best solution found by the GA to specify the initial distribution of the pheromone (used by the ACO). Thus our ACO algorithm uses the GA "experience" and starts from a "better" pheromone distribution. This strategy helps the ants

TABLE III
PARAMETERS OF GA-ACO ALGORITHM

Parameter	Value
ggap	0.97
xovr	0.7
mutr	0.05
GA maxgen	40
individuals	20
nvar	3
inserted rate	100 %
number of ants	20
initial pheromone	0.5
evaporation	0.1
ACO generations	100

TABLE IV
RESULTS FROM MODEL PARAMETERS IDENTIFICATION PROCEDURES

Value	Algorithm	Algorithm performance	
		T, [s]	J
best	GA	67.5172	4.4396
	ACO	67.3456	4.9190
	GA-ACO	38.7812	4.3803
worst	GA	66.5968	4.6920
	ACO	66.6280	6.6774
	GA-ACO	41.4495	4.6949
average	GA	67.1370	4.5341
	ACO	69.5379	5.5903
	GA-ACO	39.4620	4.5706

to find "good solutions" using less computational resources (e.g. like computer time and memory). As a matter of fact, our hybrid algorithm uses more than three times less memory than the pure ACO and the pure GA algorithms.

In Table VI we compare results achieved in current work with results obtained in our earlier work [32]. There, we had run the ACO algorithm for several iterations and used it to generate an initial populations for the GA algorithm. Thus the GA started from a population that was closer to the good (optimal) solution than a randomly generated population. We observe that the ACO-GA and the GA-ACO algorithms achieve very similar results, and in a similar running time. We run the ANOVA test to measure the relative difference between the two algorithms. The two hybrid algorithms achieves statistically equivalent results, but the GA-ACO algorithm uses 30% less memory. Thus we can conclude that hybrid GA-ACO algorithm performs better than the ACO-GA hybrid algorithm.

TABLE V
PARAMETERS' ESTIMATIONS OF THE *E. coli* FED-BATCH FERMENTATION PROCESS MODEL

Value	Algorithm	Model parameters		
		μ_{max}	k_S	$1/Y_{S/X}$
average	GA	0.4857	0.0115	2.0215
	ACO	0.5154	0.0151	2.0220
	GA-ACO	0.4946	0.0123	2.0204

TABLE VI
RESULTS FROM MODEL PARAMETERS IDENTIFICATION PROCEDURES: ACO-GA

Value	ACO-GA performance	
	T, [s]	J
best	35.5212	4.4903
worst	41.4495	4.6865
average	36.1313	4.5765

In Figure 7, the comparison of the dynamics of measured and modeled biomass concentration is shown. With a solid line we show the modeled biomass during the fermentation process, while with stars we show the measured biomass concentration. We put only several stars because the two line are almost overlapped. In Figure 8 the comparison between the time profiles of measured and modeled substrate concentration, during the fermentation process, is shown. On both figures we observe how close are the modeled and the measured data. Thus we illustrate the quality of our hybrid GA-ACO algorithm.

VI. CONCLUSION

In this paper we propose a hybrid GA-ACO algorithm for parameter identification of the *E. coli* fed-batch fermentation process. In the proposed approach, first, we start the GA for several generations with a small population. Next, we use the best solution found by the GA, to instantiate the initial pheromone distribution for the ACO algorithm. We observe that our hybrid GA-ACO algorithm achieves results similar to the pure GA and the pure ACO algorithms, but it is using less computational resources (time and memory). The used time is two times smaller while the used memory is three times smaller. With this algorithm we understand how important is the pheromone distribution for good performance of the ACO algorithm. We compare our hybrid GA-ACO approach, with a hybrid ACO-GA algorithm. Both hybrid algorithms

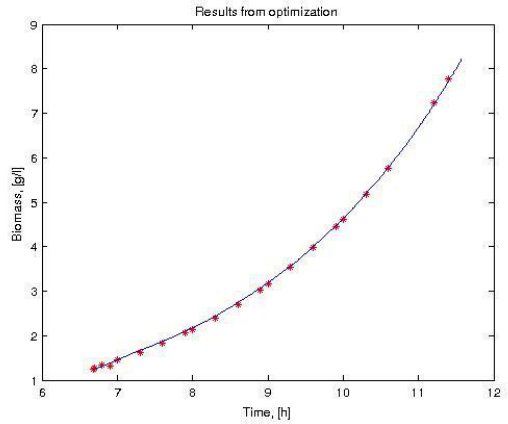


Fig. 7. *E. coli* fed-batch fermentation process: comparison between measured and modeled biomass concentration

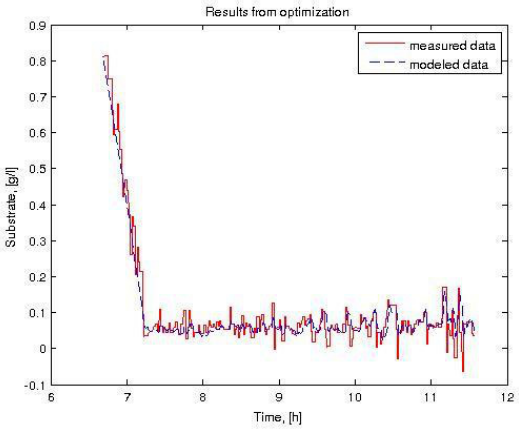


Fig. 8. *E. coli* fed-batch fermentation process: comparison between measured and modeled substrate concentration

achieve statistically similar results for a similar running time, but GA-ACO algorithm uses about 30% less memory, which is important when one is to solve large problems.

ACKNOWLEDGMENT

Work presented here is a part of the Poland-Bulgarian collaborative Grant "Parallel and distributed computing practices" and the project AComIn "Advanced Computing for Innovation," Grant 316087, funded by FP7 Capacity Programme (Research Potential of Convergence Regions).

REFERENCES

[1] G. J. Woeginger, "Exact Algorithms for NP-Hard Problems: A Survey", Lecture Notes in Computer Science, Volume 2570, 2003, pp. 185–207.
[2] M. Battarra , A. A. Pessoa, A. Subramanian and E. Uchoa, "Exact Algorithms for the Traveling Salesman Problem with Draft Limits", European Journal of Operational Research, Volume 235, Issue 1, 2014, pp. 115–128.
[3] I. Dumitrescu and T. Sttzle, "Combinations of Local Search and Exact Algorithms", G.R. Raidl (Ed.) et al., Applications of Evolutionary Computation, Lecture Notes in Computer Science, Vol. 2611, 2003, pp. 211-223.
[4] F. Glover and G. Kochenberger (Eds.), "Handbook of Metaheuristics", International Series in Operations Research and Management Science, Kluwer Academic Publishers, Vol. 57, 2003.
[5] N. Harvey, "Use of Heuristics: Insights from Forecasting Research", Thinking & Reasoning, Vol. 13 Issue 1, 2007, pp. 5–24.
[6] H. Smith, "Use of the Anchoring and Adjustment Heuristic by Children", Current Psychology: A Journal For Diverse Perspectives On Diverse Psychological Issues, Vol. 18 Issue 3,1999, pp. 294–300.
[7] C. Blum and A. Roli, "Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison", ACM Computing Surveys, Vol. 35(3), 2003, pp. 268-308.
[8] I. Boussaid, J. Lepagnot and P. Siarry, "A Survey on Optimization Metaheuristics", Information Sciences, Vol. 237, 2013, pp. 82-117.
[9] J. Toutouh, "Metaheuristics for Optimal Transfer of P2P Information in VANETs", MSc Thesis, University of Luxembourg, 2010.
[10] P. Tangpattanakul, N. Jozefiwiez and P. Lopez, "Biased Random Key Genetic Algorithm with Hybride Decoding for Multi-objective Optimization", In Proc. of FedCSIS conference, Poland, 2013, pp. 393 – 400.
[11] E. Deniz Ulker and A. Haydar, "A Hybrid Algorithm Based on Differential Evolution, Particle Swarm Optimization and Harmony Search Algorithms", n Proc. of FedCSIS conference, Poland, 2013, pp.417 – 420.
[12] E. G. Talbi and El-ghazali (Ed.), "Hybrid Metaheuristics", Studies in Computational Intelligence, Vol. 434, 2013, XXVI, 458 p. 109 illus.
[13] E. G. Talbi, "A Taxonomy of Hybrid Metaheuristics", Journal of Heuristics, 8, 2002, pp. 541-564.
[14] A. Georgieva and I. Jordanov, "Hybrid Metaheuristics for Global Optimization using Low-discrepancy Sequences of Points", Computers and Operation Research, Vol. 37(3), 2010, pp. 456–469.
[15] H. Guangdong, P. Ling and Q. Wang, "A Hybrid Metaheuristic ACO-GA with an Application in Sports Competition Scheduling", Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, Vol. 3, 2007, pp. 611–616.
[16] H. Guangdong and Q. Wang, "A Hybrid ACO-GA on Sports Competition Schedulingby Ant Colony Optimization - Methods and Applications", Edited by Avi Ostfeld, 2011, pp. 89–100.
[17] A. Csebfalv, "A Hybrid Meta-heuristic Method for Continuous Engineering Optimization", Civil Engineering, Vol. 53/2, 2009, pp. 93-100.
[18] S. AlMuhaideb and M. El B. Menai, "A New Hybrid Metaheuristic for Medical Data Classification", Int. J. of Metaheuristics, Vol. 3(1), 2014, pp. 59–80.
[19] Yi H., Q. Duan and T. Warren Liao, "Three Improved Hybrid Metaheuristic Algorithms for Engineering Design Optimization", Applied Soft Computing, Vol. 13(5), 2013, pp. 2433-2444.
[20] M. Lukasiewycz, M. Gla, F. Reimann, and J. Teich, "Opt4J - A Modular Framework for Meta-heuristic Optimization", In Proc. of the Genetic and Evolutionary Computing Conference (GECCO 2011), Dublin, Ireland, 2011, pp. 1723-1730.
[21] S. Masrom, S. Z. Z. Abidin, P. N. Hashimah, and A. S. Abd. Rahman, "Towards Rapid Development of User Defined Metaheuristics Hybridisation", International Journal of Software Engineering and Its Applicatons, Vol. 5, 2011.
[22] A. Acan, "A GA + ACO Hybrid for Faster and Better Search Capability", In: Ant Algorithms: Proc. of the Third International Workshop, ANTS 2002, Lecture Notes in Computer Science, 2002.
[23] O. Roeva, T. Pencheva, B. Hitzmann, St. Tzonkov, "A Genetic Algorithms Based Approach for Identification of Escherichia coli Fed-batch Fermentation", Int. J. Bioautomation, Vol. 1, 2004, pp. 30–41.
[24] O. Roeva and S. Fidanova, "Chapter 13. A Comparison of Genetic Algorithms and Ant Colony Optimization for Modeling of E. coli Cultivation Process", In book "Real-World Application of Genetic Algorithms", O. Roeva (Ed.), InTech, 2012, pp. 261–282.
[25] E. Bonabeau, M. Dorigo and G. Theraulaz, *Swarm Intelligence: From Natural to Artificial Systems*, New York,Oxford University Press, 1999.
[26] M. Dorigo and T. Stutzle, *Ant Colony Optimization*, MIT Press, 2004.
[27] M. Arndt and B. Hitzmann, "Feed Forward/feedback Control of Glucose Concentration during Cultivation of *Escherichia coli*", 8th IFAC Int. Conf. on Comp. Appl. in Biotechn, Canada, 2001, pp. 425–429.
[28] O. Roeva, "Improvement of Genetic Algorithm Performance for Identification of Cultivation Process Models", Advanced Topics on Evolutionary Computing, Book Series: Artificial Intelligence Series-WSEAS, 2008, pp. 34–39.

[29] D. E. Goldberg, “Genetic Algorithms in Search, Optimization and Machine Learning”, Addison Wesley Longman, London, 2006.
[30] J. H. Holland, “Adaptation in Natural and Artificial Systems”, 2nd Edn. Cambridge, MIT Press, 1992.
[31] M. Dorigo and T. Stutzle, “Ant Colony Optimization”, MIT Press, 2004.
[32] O. Roeva, S. Fidanova, V. Atanassova, ”Hybrid ACO-GA for Parameter Identification of an E. coli Cultivation Process Model”, Large-Scale Scientific Computing, Lecture Notes in Computer Science 8353, Springer, Germany, ISSN 0302-9743, 2014, 288 – 295.
[33] http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol1/hmw/article1.html (last accessed April 14, 2014)

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 421–428
DOI: 10.15439/2014F284
ACSIS, Vol. 2

Width Beam and Hill-Climbing Strategies for the Three-Dimensional Sphere Packing Problem

Mhand Hifi* and Labib Yousef
EPRAOD EA 4669, Université de Picardie Jules Verne
7 rue du Moulin Neuf, 80000 Amiens, France.
Emails: {mhand.hifi, labib.yousef}@u-picardie.fr
*Corresponding author.

***Abstract*—In this paper we propose to enhance a width-beam search in order to solve the three-dimensional sphere packing problem. The goal of the problem is to determine the minimum length of the container having fixed width and height, that packs *n* predefined unequal spheres. The width-beam search uses a greedy selection phase which determines a subset of eligible positions for packing the predefined items in the target object and selects a subset of nodes for exploring some promising paths. We propose to handle lower bounds in the tree and apply a hill-climbing strategy in order to diversify the search process. The performance of the proposed method is evaluated on benchmark instances taken from the literature. The obtained results are compared to those reached by some recent methods available in the literature. Encouraging results have been obtained.**

***Index Terms*—Beam; Heuristic; Hill-Climbing; Optimization; Packing.**

I. Introduction

THIS paper deals with the *Three-Dimensional Sphere Packing Problem* (noted 3DSPP), where an instance of such a problem is defined by a set I of n unequal items and an object $\mathcal{P}$ having fixed width W and height H and, unlimited length (another representation for similar problems can be found in Wascher *et al*. [20]). In this case, each item $i \in I = \{1, \ldots, n\}$ is characterized by its radius r_i and the goal of the problem is to minimize the length, denoted L, of the object $\mathcal{P}$ such that all items of I are packed in the target object, without overlapping. The 3DSPP may be formulated as follows:

$$\min \quad L \tag{1}$$
$$(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2 \geq (r_i + r_j)^2 \tag{2}$$
$$\forall (i,j) \in I \times I,\ i < j$$
$$r_i \leq x_i \leq L - r_i, \forall i \in I \tag{3}$$
$$r_i \leq y_i \leq H - r_i, \forall i \in I \tag{4}$$
$$r_i \leq z_i \leq W - r_i, \forall i \in I \tag{5}$$
$$\underline{L} \leq L \leq \overline{L} \tag{6}$$
$$(x_i, y_i, z_i) \in \mathbb{R}^3_+, \forall i \in I \tag{7}$$

where the objective function (1) minimizes the length of the object $\mathcal{P}$ containing all the items of I, Eq. (2) ensures the non-overlap constraint of any pair of distinct items *(i, j)* of $I \times I$; that is, the distance between the centers of both items which must be greater than or equal to the sum of their radii and Eqs (3)-(5) ensure that all items of I belong to the target object $\mathcal{P}$ of dimensions *(L, W, H)* and Eq. (7) ensures that all items are placed in the object $\mathcal{P}$. Also, it is easy to start any approach by a trivial value representing the sum of the spheres' area affected to $\underline{L}$ (Eq. (6)) and a feasible solution value, that can be obtained by applying a simple greedy procedure, affected to $\overline{L}$.

In this paper, we propose new strategies in order to enhance a width-beam search based algorithm already proposed in Hifi and Yousef [8]. First, we introduce a new greedy strategy in order to generate some eligible nodes. Second, we propose to curtail the search process by estimating a global lower bound. Third ans last, a hill-climbing strategy is used in order to correct the global lower bound and to select some nodes realizing highest potentials able to reach better solutions.

The remainder of the paper is organized as follows. Section II gives a literature review for the 3DSPP and some of its variants. The problem representation is discussed in Section III-A. The greedy selection phase, which serves to determining a subset of eligible positions for predefined items to pack, is detailed in Section III-B. A modified version of the algorithm is described in Section III-D, where a lower bound is used for exploring better paths that diversify the search. Moreover, because of the huge number of feasible positions that a predefined item may generate, both beam width and hill-climbing strategies cooperate for selecting the best promising nodes at the same level of the developed tree. Section IV evaluates the performance of the proposed algorithm and compares its produced results to those reached by the original width-beam search and recent methods available in the literature. Finally, Section V concludes by summarizing the contribution of this paper.

II. Related works

The 3DSPP belongs to the well-known family of Cutting and Packing (CP) that represents a natural combinatorial optimization problems. Problems of CP family admit numerous real-world applications in the domain of industrial engineering, logistics, manufacturing, production process, automated planning, etc. One of the more recent paper addressing an optimization with a packing problem is due to Sutou and Dai [18], where the unequal sphere problem has been used for tackling an application of the automated radio-surgical treatment planning. Wang [19] has also considered sphere packing problems as an optimization tool for the radio-surgical

treatment planning. Other problems of the CP family have been described and redefined in Wascher *et al.* [20] where an instance of these problems can be defined by a set of predetermined items to be packed in one or many larger containers (objects) so as to minimize the unused area / space or in some cases to maximize a utility function. Furthermore, the items are bounded by their dimensions (rectangular, circular, or irregular) and the objects can be bounded (rectangular, circular, ...) or unbounded (strips / parallelepipeds, ...).

Among existing papers addressing sphere packing problems, Lochmann *et al*. [12] proposed a statistical analysis for packing random spheres with variable radius distribution. Li and Ji [11] discussed a dynamics-based collective method for random sphere packing and they tried to apply it to the problem of packing sphere into a cylinder container. In this paper, the stability of the method and its convergence were tackled. Farr [3] studied the problem of random close packing fractions of log-normal distributions of hard spheres. The author tailored a one-directional approach in order to predict a close packing of spheres of log-normal distributions of sphere sizes.

Packing spheres into a container has been addressed by Sutou and Dai [18] who proposed a global optimization approach. Stoyan *et al.* [17] designed a mathematical model in order to pack spheres into an open container, where both height and widths are fixed whereas the length is unfixed. They use a neighboring search based upon extremum points in order to construct and improve a series of solutions. M'Hallah *et al.* [13] proposed a heuristic based on combining VNS with nonlinear programming solver. The method iterates some moves of the current configuration and complete the partial configuration with a solver dedicated for nonlinear programmes. M'Hallah and Alkandari [14] considered the principle used in [13] to solve the problem of packing identical spheres into the smallest containing sphere. Soontrapa and Chen [16] tackled the problem of packing identical spheres into a smallest containing sphere by using a random search according to Monte Carlo's method. Birgin and Sobral [1] proposed twice-differentiable non-linear programming models for the problem of packing both circles and spheres into different containers where the containers may be circular, rectangular, etc. In order to find a global solution for their proposed models, ALGENCAN solver was used for generating a multiple starts solutions. Finally, Hifi and Yousef [8] investigated the use of a dichotomous search for solving the three-dimensional sphere packing problem (an extensive efficient models and methods for packing both circular and sphere problems were reviewed in Hifi and M'Hallah [4]).

In this paper, we propose to enhance the algorithm proposed in Hifi and Yousef [8] by considering three modifications: (i) considering a modified greedy strategy in order to generate eligible nodes, (ii) an estimation of the global lower bound for curtailing the search process and, (iii) the hill-climbing used for correcting the global lower bound and selecting only some nodes with highest potentials.

III. A WIDTH-BEAM SEARCH FOR 3DSPP

In this part, the problem representation and strategies used are first described in Secsion III-A. Second, Section III-B describes the greedy procedure in order to build feasible packings containing all items of I. Third, Section III-C discusses the principle of the proposed algorithm and its main steps. Fourth and last, Section III-D presents the modifications used for enhancing the algorithm.

A. Representation of the problem

The local strategy is based on the simple Greedy Principle (called GP) where the minimum distance position is favored for packing a series of predefined items. GP is then used as the first evaluation operator for finding a subset of possible positions of the next item to pack. Such a procedure uses the following notations:

- The bottom-left-depth corner of $\mathcal{P}$ is positioned at $(0, 0, 0)$ and $\mathcal{P}$ is characterized by a set formed with six labels (namely faces): $\mathbb{F} = \{\text{left}, \text{top}, \text{right}, \text{bottom}, \text{depth}, \text{front}\}$. Then, $\mathcal{P}$ is represented in the Euclidean space, as illustrated in Figure 1.
- The center of the i-th item belonging to I is positioned at (x_i, y_i, z_i).
- The distance $\delta_{i,j}$ between two items i and j is computed as follows: $\forall (i, j) \in I^2$,

$$\delta_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} - (r_i + r_j).$$

Note that assigning an item $i \in I$ to a possible position of $\mathcal{P}$ while respecting non-overlapping between this item and the left-face of $\mathcal{P}$ requires to satisfy the following distance: $\delta_{i,\text{left}} = x_i - r_i$. For more general cases, Table I reports the distance to be satisfied whenever an item i is assigned to a selected position of $\mathcal{P}$.

TABLE I
THE DISTANCE BETWEEN AN ITEM i AND A FACE f

f	$\delta_{i,f} \mid i \in I,\ f \in \mathbb{F}$
left	$x_i - r_i$
bottom	$y_i - r_i$
depth	$z_i - r_i$
right	$L - x_i - r_i$
top	$H - y_i - r_i$
front	$W - z_i - r_i$

B. Defining eligible positions

It is well-known that tailored heuristics are mainly based on the strategies which are able to guide well the search process. These strategies may be use some selection criteria in order to provide either partial or final solutions for the problem to solve. Herein, we consider a simple greedy principle (GP) which is based on searching the position realizing the minimum distance position between items and feces. In fact, GP is used as a selection criterion for defining a set of eligible positions to assign to the predefined item i (not

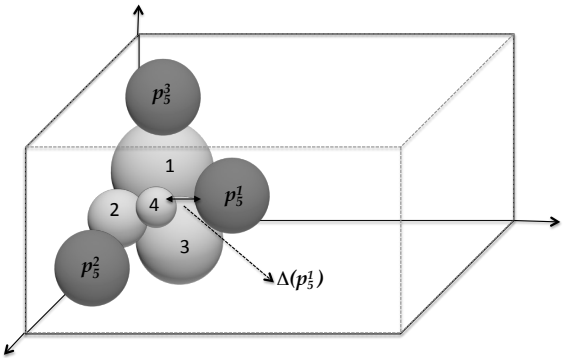


Fig. 1. Illustration of the mechanism used for computing eligible positions considered by GP.

already positioned) among all eligible positions representing the set P_{I_i}.

In what follows, we assume that the (center of the) first item $i = 1$ of I is positioned at the position (r_1, r_1, r_1) and $\forall i \in I,\ i \geq 2$, the following notations are considered:

- I_i denotes the set of items of I already positioned in the current object $\mathcal{P}$.
- $\overline{I_i}$ contains the items of I which are not yet assigned to $\mathcal{P}$.
- P_{I_i} denotes the set of distinct eligible positions for the next item i to pack given the set of packed items I_i.
- An eligible position $p_{i+1} \in P_{I_i}$ (for the item i) is determined by using three elements $e_1,\ e_2$ and e_3 where an element is either an item of I already positioned (representing the set I_i) or one of the six faces belonging to $\mathbb{F}$.
- $\mathcal{T}_{p_{i+1}}$ represents the set composed of the three elements $e_1,\ e_2$ and e_3.

Figure 1 illustrates GP's mechanism on a small example. For this example, assume that the first four items are already positioned in the object $\mathcal{P}$; then, there are three eligible positions that emerge for the next item 5 to pack. Following the above notations, $I_4 = \{1, 2, 3, 4\}$ and $P_{I_4} = \{p_5^j,\ j = 1, \ldots, 3\}$. First, the position p_5^1 touches both items 1 and 3 and, the "bottom" face of $\mathcal{P}$. Second, the position p_5^2 is obtained by using the item 2 and both faces "left" and "bottom" of $\mathcal{P}$. Third and last, the position p_5^3 is computed by using the item 1 and both faces "left" and "depth" of $\mathcal{P}$. Finally, it follows that $\mathcal{T}_{p_5^1} = \{1, 2, \text{bottom}\}$, $\mathcal{T}_{p_5^2} = \{2, \text{left}, \text{bottom}\}$ and $\mathcal{T}_{p_5^3} = \{1, \text{left}, \text{depth}\}$. Further, item 4 is positioned around the three items 1, 2 and 3 that means $\mathcal{T}_4 = \{1, 2, 3\}$. For the next item 5 to pack, its coordinates are computed by using both items 1 and 3 and one of the faces (the "bottom" in this case) of $\mathcal{P}$ that gives $\mathcal{T}_5 = \{1, 3, \text{bottom}\}$. We recall that the objective of the problem is to minimize the length of the target object $\mathcal{P}$. It means that the right face of $\mathcal{P}$ can be omitted and only the five faces can be considered when optimizing the length of the target $\mathcal{P}$. Hence, all eligible positions may be obtained by using the fifth faces, the positioned items and the current item to pack.

Then, the value corresponding to the $(i+1)$-th item to pack, when positioned at the eligible position $p_{i+1}^k \in P_{I_i}$ by GP, is computed as follows

$$\Delta(p_{i+1}^k) = \min_{j \in I_i \cup \mathbb{F}''} \delta_{(i+1,j)} \tag{1}$$

where $\mathbb{F}' = \mathbb{F} \setminus \{\text{right}\}$ and $\mathbb{F}'' = \mathbb{F}' \setminus \mathcal{T}_{p_{i+1}^k}$.

Finally, when GP is used, it starts by positioning the first item $i = 1$ at the bottom-left-depth position, i.e., at the position (r_1, r_1, r_1), while the remaining $n - 1$ items are successively positioned according to the minimum distance rule (cf., Eq. (1)). As illustrated in Figure 1, the item 5 will be placed at position p_5^1 because its corresponding distance realizes the minimum value.

C. A width-beam search heuristic for the 3DSPP

Beam Search (BS) has been first proposed in [15] for tackling the scheduling problem and it has since been successfully applied to many other combinatorial optimization problems (some adaptations can be found in Della Croce *et al.* [2], Hifi *et al.* [5], [6], [7] and, Yavuza [21]). Such an approach can be viewed as a truncated tree search procedure where its objective is to avoid exhaustive search by performing a partial enumeration of the solution space.

1) Packing all items on the target object: At each level of the developed tree, only a subset of nodes (called the *set of elite nodes*) are selected for further branching and the other nodes are discarded, where no backtracking is performed. For each level, the cardinality of the elite nodes to be investigated is fixed to ω that is called the *beam width*. Generally, these selected ω nodes represent those having a high potential to lead the best solutions for the treated problem. Furthermore, each node is assessed via an *evaluation function* whose role is to provide a promising separation mechanism of the nodes of each level of the developed tree.

As observed in Hifi and Yousef [8], applying BS to 3DSPP required to define the nodes of the tree and the branching mechanism out of the nodes of B. Herein, a node η_i is represented by the pair of subsets:

1) The first subset I_i containing all items assigned to the target object $\mathcal{P}$ and,
2) The complementary subset $\overline{I}_i$ containing the unassigned items.

Moreover, branching out of a node η_i is equivalent to create at most $|P_{I_i}|$ branches emanating out of the current node (related to the eligible positions as described in Section III-B). Each resulting node corresponds to packing the subset of items I_i and assigning to the current item i a favorite eligible position. Moreover, each of these created nodes will be represented by a pair of two complementary subsets of items of I. Further, in order to explore a reasonable number of nodes, a width-beam search almost of the standard beam search has been

used in Hifi and Yousef [8]. Therefore, all nodes emanating from the same level are simultaneously evaluated following an estimator operator and only the best ones are selected for the rest of the search.

Such a process is described by the main steps of Algorithm 1, where it works according to a given node, namely η^ℓ. This node is the one taken at the level ℓ of the developed tree. Thereafter, the initialization step is applied for starting the set B containing the best provided nodes regarding the starting node η^ℓ (lines 1 to 3), the initialization of the variable `feasible` to false (line 4) and, fixing the runtime limit $t_{\max}$ (line 5) for which the algorithm stops when that time is performed (in this case, the control parameter t_{iter}, for the limit $t_{\max}$, is initialized to zero). Note that the variable `feasible` is used for controlling the (un)feasibility of the series of the solutions builded. The main loop (line 7) starts by choosing the best eligible positions for each node belonging to B. These positions are computed by using GP's selection (cf., Section III-B). Second, all created nodes are stored in a provisional set B_ω where the potential of each of these nodes are evaluated according to the final solution provided by iteratively applying GP as a heuristic (cf. as discussed in the last paragraph of Section III-B). Thereafter, for each final solution (either feasible or unfeasible for the target object $\mathcal{P}$), the potential of a node $\eta \in B_\omega$ is represented by the *density of the positioned items* in $\mathcal{P}$. Whenever one of these constructed solutions provides a feasible solution (line 11), i.e., all items are positioned in the target object $\mathcal{P}$, then the algorithm stops with a feasible solution (i.e., setting the variable `feasible` to true). Otherwise, the set B of the best nodes is updated (line 12) by the ω nodes which realizing the highest densities and the current level of the developed tree is incremented. The internal runtime t_{iter} is then incremented and the process is iterated until either B is reduced to an empty set or when $t_{\max}$, the limited runtime, is performed, i.e., $t_{\text{iter}} \geq t_{\max}$.

Algorithm 1 . Beam Search for the 3DSPP: BS

Input. A node η^ℓ.
Output.`feasible` // setting equal to `true` whenever a feasible packing is reached, `false` otherwise

1: **Initialization Step.**
2: Let ω be a predefined beam width.
3: Set $B = \{\eta^\ell\}$, where η^ℓ denotes the input node associated to the ℓ-th level.
4: Set the variable `feasible` to `false` /* no feasible solution at hand */
5: Let $t_{\max}$ be a maximum fixed runtime and t_{iter} (initialized to 0) be a counter which serves to control the time spent for exploration the space search.
6: **Iterative Step.**
7: **while** $\Big((B \neq \emptyset)$ and $(t_{\text{iter}} < t_{\max})\Big)$ **do**
8: Branch from the current level ℓ by selecting the ω eligible positions for each node $\eta_{\ell_i} \in B$;
9: Insert all obtained nodes into B_ω;
10: Evaluate the potential of each node belonging to B_ω using GP for completing the path.
11: If a feasible solution is given by GP, then set `feasible` to `true` and **exit**;
12: Replace B by the best ω nodes of B_ω realizing highest densities and, increment the level ℓ.
13: Update the current runtime t_{iter}.
14: **end while**

Algorithm 2 . A Dichotomous Search Based Heuristic: DSBH

Input. An instance of 3DSPP.
Output.An object $\mathcal{P}$ of dimensions (L_{best}, W, H) and the coordinates of all items of I.

1: **Initialization step**
2: Call an iterative GP on the open strip (∞, W, H) and let $\overline{L}$ be the starting length reached.
3: Set $\underline{L} \leftarrow \frac{4\pi}{3\times W\times H}\sum_{i\in N}(r_i^3)$ and $L^\star = \overline{L}$.
4: Set ω to a predefined minimum value.
5: **Iterative step**
6: **while** $\Big($the runtime limit is not performed$\Big)$ **do**
7: **repeat**
8: $L^\star = (\overline{L} + \underline{L})/2$
9: Generate the starting node η^1 with its three sets I_i, $\overline{I}_i$ and P_I^1.
10: Set `feasible` $\longleftarrow$ BS(η^1), where BS is called with $(L^\star, W, H)$
11: **If** $\Big($`feasible=true`$\Big)$ **then** set $\overline{L} = L^\star$; $\underline{L} = L^\star$ otherwise
12: **until** $(\overline{L} - \underline{L} \geq \alpha)$
13: Set $\underline{L} \leftarrow \frac{4\pi}{3\times W\times H}\sum_{i\in N}(r_i^3)$ and increment ω.
14: **end while**

2) Using a dichotomous search: Because Algorithm 1 is applied on the target container $\mathcal{P}$, then one can repeat the same principle on a series of target containers $\mathcal{P}_1, \ldots, \mathcal{P}_r$, where $r \geq 1$. Indeed, one can starts the search with the initial interval $[\underline{L}, \overline{L}]$, where $\underline{L}$ denotes a lower bound for the 3DSPP and $\overline{L}$ its upper bound (in the case of a feasible solution exists, its objective value is assigned to $\overline{L}$). Then, for each fixed interval, Algorithm 1 tries to construct the best feasible solution by packing all the items into the current target object; that is, $(L^\star, W, H)$, where $L^\star \in [\underline{L}, \overline{L}]$.

The main steps of the dichotomous search are summarized in Algorithm 2. First, it starts by defining the initial interval $[\underline{L}, \overline{L}]$ where the upper bound $\overline{L}$ is obtained by applying GP as a heuristic on the open object, i.e., (∞, W, H). The main loop "repeat ... until" (cf., lines 7 - 12) of the dichotomous procedure serves to explore a series of neighborhoods depending on the values of ω. At line 8, a new target upper bound is computed, namely $L^\star = (\overline{L}+\underline{L})/2$. Line 9 generates the initial node positioned at the bottom-left-depth corner (in the position (r_1, r_1, r_1)) and creates its corresponding sets I_i, $\overline{I}_i$ and P_I^1 (as discussed in Section III-B). At line 10, BS is called with the target value of the object $(L^\star, W, H)$ and the created sets reached at the next step. Line 11 serves to update the interval search where its upper bound is updated whenever a feasible solution is obtained, the lower bound is updated otherwise. Thereafter, the process is iterated untill the gap between both lower and upper bounds becomes closest to a certain tolerance, namely α. Finally, the aforementioned process is iterated a certain number of times following the values of ω (line 13) and according to the runtime limit fixed.

D. A modified version of the width-beam search for 3DSPP

We first describe the modified GP that tries to generate some interesting eligible positions. Second and last, we introduce the lower bound in order to curtail the search; this upper bound cooperates with both hill-climbing strategy and beam width strategies in order to select future nodes for branching.

1) A modified GP: The Modified GP (MGP) provides other eligible positions able to homogeneously concentrate a subset of items on the target object. Return now to Figure 1 and observe the candidate positions of item p_5: any eligible position induces a packing concentrated on the bottom-left-depth position. By applying this principle to the instance SYS1 (one of the instances tested in Section IV), one can observe that all packed items are focused on the starting position that leaves the other parts of the object sufficiently empty.

Herein, we first propose to modify such a placement by adding three corner positions whenever all eligible positions for a selected item to pack; that are $\mathcal{T}_{p_2^4} = \{\text{top}, \text{depth}, \text{left}\}$, $\mathcal{T}_{p_2^5} = \{\text{top}, \text{left}, \text{front}\}$ and $\mathcal{T}_{p_2^6} = \{\text{bottom}, \text{left}, \text{front}\}$.

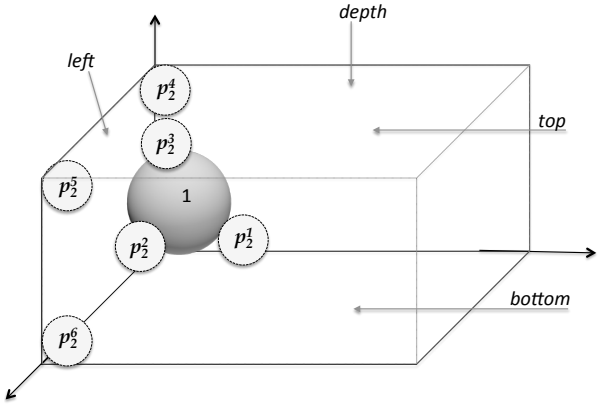


Fig. 2. Illustration of MGP's eligible positions.

Figure 2, where we assume that the first item is already positioned in the object $\mathcal{P}$, shows six eligible positions that emerge for the next item 2 to pack. According to the representation described above, $I_1 = \{1\}$ and $P_{I_2} = \{p_2^j,\ j = 1, \ldots, 6\}$. First, the position p_2^1 touches the item 1 and both faces "depth" and "bottom" of $\mathcal{P}$. Second, the position p_1^2 is obtained by using the item 1 and both faces "left" and "bottom" of $\mathcal{P}$. Third, the position p_1^3 is computed by using the item 1 and both faces "left" and "depth" of $\mathcal{P}$. Finally, all other positions touch three faces of $\mathcal{P}$. It follows that $\mathcal{T}_{p_2^1} = \{1, \text{depth}, \text{bottom}\}$, $\mathcal{T}_{p_2^2} = \{1, \text{left}, \text{bottom}\}$, $\mathcal{T}_{p_2^3} = \{1, \text{left}, \text{depth}\}$, $\mathcal{T}_{p_2^4} = \{\text{top}, \text{left}, \text{depth}\}$, $\mathcal{T}_{p_2^5} = \{\text{top}, \text{left}, \text{front}\}$ and $\mathcal{T}_{p_2^6} = \{\text{bottom}, \text{left}, \text{front}\}$.

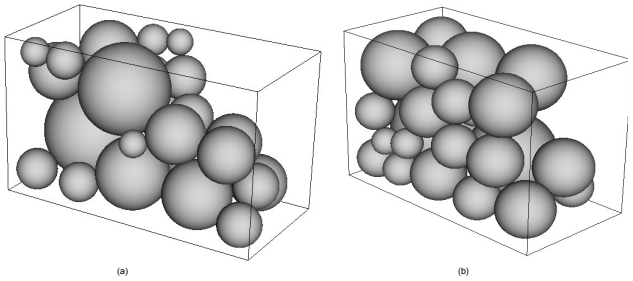


Fig. 3. Illustration of (a) GP's behavior and (b) its modified version MGP's on SYS1 instance.

On the one hand, as illustrated in Figure 2, the item 2 can be packed at one of the six positions $p_2^j,\ j = 1, \ldots, 6$, and if the smallest distance to the left side of $\mathcal{P}$ is favored, then five positions remains favorable (cf., Figure 3.(a)). In this case GP provides an object of length equal to 11.51. On the other hand, If such a principle is applied, from Figure 3.(b) and for the same instance, one can observe the improvement made on the packing which at the same time produces a better length (10.49 in this case) for the target object.

E. Using the hill-climbing strategy

Hill-Climbing (HC) can be used as a curtailing strategy for avoiding exhaustive search. In this case, the search process may perform a partial enumeration of the solution space, where in term of tree, only a subset of paths are taken for further branchings and the other nodes are discarded. Further, each selected node is assessed via its evaluation function whose role is to provide a promising separation mechanism of the nodes. In our study, we introduce the HC strategy that is used for avoiding exhaustive search that is equivalent to an augmented beam search (Hifi *et al.* [7] and Yavuza [21])), where a subset of paths are taken for further branchings and the other nodes are discarded. At each step of the search procedure, a node η is selected and after evaluating all its successors, only the best ω nodes are chosen for further branchings. Each selected node is assessed via its evaluation function whose role is to provide a promising separation mechanism of the nodes.

Algorithm 3 describes a modified BS (noted MBS) where MBS replaces BS in the dichotomous search DSBH (cf., Algorithm 2, which also noted MDSBH for the modified DSBH). We recall that a node corresponds to a partial solution at level $\ell \leq n-1$ and the set B of current nodes contains initially the starting nodes of the root node B_0, whereas B_ω containing the offspring nodes is initialized to the empty set.

Algorithm 3 . A Modified BS (MBS)

Input. A set of items I and a predefined length l_{best}.
Output. `feasible`..

1: **Initialization Step.**
2: Let ω and ϵ be two predefined values.
3: Set $B = B_0$, where B_0 denotes the starting eligible nodes according to the first packed item $i = 1$.
4: Set the level $\ell = 1$ and $B\ell = \emptyset$.
5: Set the variable `feasible` to `false` /* no feasible solution at hand */
6: **Iterative Step.**
7: **while** $\Big((B \neq \emptyset)$ and (the runtime limit is not performed) and $(\ell < n)\Big)$ **do**
8: **for** each $\eta \in B$ **do**
9: Let $B_\eta = \{\gamma_1, \ldots, \gamma_{|P_{I_\eta}|}\}$ be the successors of η.
10: Evaluate the potential of each node γ belonging to B_η by computing $g(\gamma)$ and $h'(\gamma)$.
11: For each $\gamma \in B_\eta$ apply ISBH$(\gamma, L^\star)$ and update $L^\star$ if necessary with the incumbent solution.
12: Set $B_\ell = B_\ell \cup B_\eta$;
13: **end for**
14: Filter B_ℓ by keeping the ω best nodes realizing the smallest values of $L^\star / \big(g(\gamma) + h'(\gamma)\big)$.
15: Replace all the nodes of B by those of B_ℓ, increment ℓ and set $B_\ell = \emptyset$.
16: **end while**

On the one hand, a selected node η taken from B (step 7), whose evaluation is z_η, creates a subset of nodes $B_\eta =$

$\{\gamma_1, \ldots, \gamma_{|P_{I_\eta}|}\}$, where each resulting node is evaluated according to its *cost operator*; that is,

$$z_\eta = g(\eta) + h(\eta).$$

On the other hand, because $|P_{I_\eta}|$ is large, only some nodes are chosen for further branchings. Indeed (line 9), if a node γ of B_η packs at most $n-1$ items, then it remains in B_η when $z'(\gamma) < z^\star$, such that

$$z'(\gamma) = g(\eta) + h'(\eta) \quad (2)$$

where $h'(\eta) = (1+\epsilon)h(\eta)$ and ϵ is considered as a small predefined value that is used for making a correction on the complementary lower bound $h(\eta)$. Whenever Eq. (2) is not satisfied, then γ is removed from B_η.

Further, since we try to diversify the search that allows for exploring new solutions, we apply BGP on all selected nodes (line 10). Then, $L^\star$ is updated whenever BGP produces a better length; in this case, its corresponding incumbent solution is also updated. The rest of the nodes belonging to B_η (line 11) are reordered in nondecreasing order of their estimated lower bound $z'(\gamma)$ and only the best ω nodes are selected and becomes the new nodes of B for further branchings. This process is iterated until no further branching is possible, i.e., until $B = \emptyset$, or the last level is equal to n, or when the fixed runtime limit is performed. Note also that, at lines 9 and 10, if a node γ of B_η is a leaf (i.e, no further branching is possible out of γ), then its objective function value z_γ is computed and compared to $z^\star$. If $z_\gamma < z^\star$, then the incumbent solution is set to a leaf node; $z^\star$ is then updated: $z^\star = z_\gamma$; and γ is removed from B_η.

IV. Computational results

In this section we investigate the effectiveness of the modified width beam search-based heuristic (noted MDSBH) on two sets of benchmark instances: Set1 and Set2. The proposed algorithm was coded in C++ and tested on an Intel Core 2 Duo (2.53 Ghz and with 4 Gb of RAM) and the runtime limit was fixed to one hour.

The first set "Set1" contains six instances (SYS1, ..., SYS6) extracted from Stoyan *et al.* [17], where the number of the predefined items varies from 25 to 60. These instances have been already tested using Stoyan *et al.*'s [17], Birgin and Sobral's [1] and Kubach *et al.*'s [10] approaches.

The second set "Set2" contains six instances (KBTG1, KBTG2, KBTG3, KBTG7, KBTG8, and KBTG9) taken from Kubach *et al.* [10]. For each instance, both dimensions W and H of the object are fixed to 10 whereas the number of the predefined items is fixed to 30 (resp. 50) for the first (resp. last) three instances. Moreover, these six instances have been already tested in Kubach *et al.* [10] where they represent the six instances with unequal spheres.

A. Performance of MDSBH vs five heuristics: Set1

Generally, when using approximate algorithms to solve optimization problems, it is well-known that different parameter settings for the approach lead to results of variable quality. As discussed in Section III-D, MDSBH considers three parameters: the beam width ω, the value of ϵ used for correcting the value of the global lower bound and the maximum runtime limit to fix. Our computational study was conducted by varying ω in the discrete interval $\{5, 6, 7, \ldots\}$, the maximum runtime limit was fixed to 3600 seconds (which can be considered as a standard runtime limit considered by algorithms of the literature) and ϵ which takes one of the following values: $0.1,\ 0.2$ and 0.3. Of course, the upper value of ω depends on the limited runtime and the size of the instance.

In order to show the effect of these parameters, we first discuss the quality of the solutions obtained by MDSBH when varying the value of ϵ. Table II shows MDSBH's objective values when varying ϵ from 0.1 to 0.3. From Table II, we observe that MDSBH with $\epsilon = 0.2$ provides better average results since it realizes a value of 9.939 compared to both values 9.957 and 9.961, which corresponds to $\epsilon = 0.1$ and $\epsilon = 0.3$, respectively.

	MDSBSs' solutions when varying ϵ		
Label	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$
SYS1	9.1946	9.1796	10.9001
SYS2	8.910122	8.8922	8.8922
SYS3	8.6862	8.6702	8.6862
SYS4	10.2154	10.2012	10.2300
SYS5	10.9237	10.8954	10.9222
SYS6	11.8105	11.7943	11.8105
Av.	9.957	9.939	9.961

TABLE II
Behavior of MDSBH, when varying ϵ, on the instances of Set1.

Figure 4 illustrates the configurations realized by MDSBH for instance SYS1. Hence, for the rest of the paper, $\epsilon = 0.2$ is chosen for evaluating the performance of MDSBH on all benchmark instances of the literature.

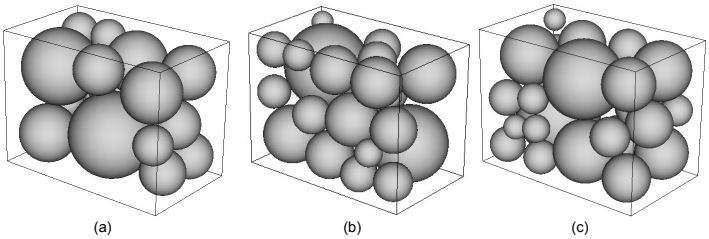


Fig. 4. SYS1 solutions (with MDSBH) when varying ϵ: (a) $\epsilon = 0.1$ with length $L^\star = 9.1946$, (b) $\epsilon = 0.2$ with length $L^\star = 9.1796$, and, (c) $\epsilon = 0.3$ with length $L^\star = 10.9001$.

Second, for the instances of Set1, Table III compares the results of MDSBH to those reached by five algorithms: SYS (Stoyan *et al.* [17]), BSA (Birgin and Sobral [1]), KBTG$_s$ (Kubach *et al.* [10]), its parallel version noted KBTG$_p$ (proposed in Kubach *et al.* [9]), where the known solutions of the literature are taken from Kubach *et al.* [9], [10] and HY (Hifi and Yousef [8]).
Column 1 of Table III shows the instance label, Column 2 displays the objective value $L^\star_{\rm SYS}$ realized by STS whereas column 3 displays BSA's objective value (noted $L^\star_{\rm BSA}$).

#Inst. Label	SYS $L^\star_{\mathrm{SYS}}$	BSA $L^\star_{\mathrm{BSA}}$	$L^\star_{\mathrm{KBTG}_s}$	$L^\star_{\mathrm{KBTG}_p}$	HY $L^\star_{\mathrm{HY}}$	MDSBH $L^\star_{\mathrm{MDSBH}}$	ω^*
SYS1	9.912	9.7942	9.2874	9.2656	9.2431	9.1796$^\diamond$	26
SYS2	9.623	-	9.1280	8.9301	8.9164	8.8922$^\diamond$	29
SYS3	9.473	9.3090	8.9850	8.7178	8.7055	8.6702$^\diamond$	31
SYS4	11.086	11.0962	10.8760	10.4042	10.2357	10.2012$^\diamond$	36
SYS5	11.646	11.6211	11.3494	10.9865	10.9359	10.8954$^\diamond$	34
SYS6	12.842	12.7215	12.3745	11.8399	11.8178	11.7943$^\diamond$	16
Av.	10.764	10.908	10.333	10.024	9.976	9.939	

TABLE III
PERFORMANCE OF MDSBH VERSUS THE FIVE HEURISTICS OF THE LITERATURE ON INSTANCES OF SET1. THE SYMBOL " − " (RESP. "$\diamond$") MEANS THAT THE VALUE FOR THIS INSTANCE IS NOT AVAILABLE (RESP. CORRESPONDS TO THE BEST SOLUTION).

Columns 4 and 5 report the solutions (noted $L^\star_{\mathrm{KBTG}}$) provided by the sequential KBTG$_s$ algorithm, column 5 reports the best solutions reached by the parallel version of KBTG (noted $L^\star_{\mathrm{KBTG}_p}$) without fixing the runtime limit and column 6 displays the results reached by HY. Column 7 displays the solution realized by MDSBH (noted $L^\star_{\mathrm{MDSBH}}$). Finally, column 8 reports the best value of ω for which its best solution is performed. All results of Table III are summarized in Table IV, where it tallies the percentage improvement (when it happens) yielded by MDSBH when compared to the results reached by the five other algorithms (noted %SYS, %BSA, %KBTG$_s$, %KBTG$_p$ and %HY according to the heuristics SYS, BSA, KBTG$_s$, KBTG$_p$ and HY, respectively).

#Inst. Label	MDSBH vs all heuristics (% Improvement) %SYS	%BSA	%KBTG$_s$	%KBTG$_p$	%HY
SYS1	7.39	6.28	1.16	0.93	0.69
SYS2	7.59	-	2.58	0.42	0.27
SYS3	8.47	6.86	3.50	0.55	0.41
SYS4	7.98	8.07	6.20	1.95	0.34
SYS5	6.45	6.24	4.00	0.83	0.37
SYS6	8.16	7.29	4.69	0.39	0.20
Av.	7.67	6.95	3.69	0.84	0.38

TABLE IV
PERCENTAGE IMPROVEMENTS BETWEEN ALL TESTED HEURISTICS: MDSBH, HY, SYS, BSA AND BOTH KBTG$_s$ AND KBTG$_p$ ON INSTANCES OF SET1.

The analysis of the results of both Tables III and IV follows:

1) First, MDSBH outperforms the five algorithms SYS, BSA, KBTG$_s$, KBTG$_p$ and HY. Indeed, it is able to reach the best solutions for all instances of Set1.
2) Second, when comparing MDSBHs' results to those reached by SYS, one can observe that the percentage of the improvement varies from 6.45% (instance SYS5) to 8.47% (instance SYS3). This percentage improvement remains interesting when comparing MDSBHs' results to those reached by BSA: in this case, such improvement varies from 6.24% (instance SYS5) to 8.07% (instance SYS4).
3) Third, the improvement remains positive when comparing MDSBH's results to those provided by the sequential (resp. parallel) KBTG algorithm. Indeed, the improvement when compared to the sequential version varies from 1.16% (instance SYS1) to 6.20% (SYS4) whereas it varies from 0.39 (instance SYS6) to 1.95% (instance SYS4) when compared to the parallel version.
4) Fourth and last, MDSBH realizes better results than those reached by HY; in this case, the percentage improvement varies from 0.20% (instance SYS6) to 0.69% (instance SYS1).

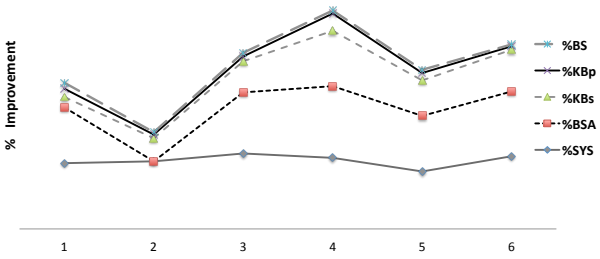


Fig. 5. Variation of the percentage improvement realized by MDSBH when compared to the results of the five heuristics (SYS, BSA, KBTG$_s$, KBTG$_p$ and HY) on the instances of Set1.

Figure 5 shows the behavior of MDSBH on the instances of Set1 where each curve represents the improvement variation realized according to the algorithm SYS, BSA and both KBTG$_s$, KBTG$_p$ and HY, respectively.

B. Performance of MDSBH versus KBTG and HY heuristics: Set2

In this section, we compare the results reached by MDSBH to those reached by KBTG$_s$ and HY (note that, for this type of instances, both KBTG$_s$ and HY realize the best objective values of the literature). This comparison is performed on the instances of the second group Set2 taken from Kubach *et al.* [10]. Herein, instead of determining the minimum length $L^\star$ of the target container $\mathcal{P}$, Kubach *et al.* [10] computed the density of all packed items in the final object $\mathcal{P}$. Therefore, we also report the best length $L^\star$ of the final object $\mathcal{P}$ as in Hifi and Yousef [8], since it corresponds to the dual problem that maximizes the density of the occupied area (or equivalently to minimizing the unused area).

#Inst. Label	$d^\star_{KBTG_s}$	HY $L^\star$	HY $d^\star$	MDSBH	MDSBH %HY	MDSBH
KBTG1	54.096	10.9031	56.0092	10.8076	23	0.884
KBTG2	30.071	1.9900	30.071	1.9900	23	0.000
KBTG3	51.387	18.2415	53.6243	18.1936	24	0.263
KBTG7	55.372	13.0997	57.5662	12.9653	14	1.037
KBTG8	45.060	2.5825	47.004	2.5820	13	0.019
KBTG9	52.732	27.8033	55.3203	27.7152	26	0.318
Av.	48.120		49.932			0.420

TABLE V
PERFORMANCE OF BSBH VERSUS KBTG$_s$ ON INSTANCES OF SET2.

The results realized by the three tested methods (MDSBH, KBTG$_s$ and HY, respectively) are reported in Table V. Column 1 displays the instance label, column 2 reports the solution value (expressed in term of density) realized by KBTG$_s$'s algorithm (extracted from Kubach *et al.* [9], [10]), columns 3 and 4 display both HY's length and its density whereas columns 5 and 6 report the best length realized by

MDSBH and the value of ω for which the best solution is reached. Finally, the last column displays the percentage of improvement realized by MDSBH according to the solution values reached by both $KBTG_s$ and HY, respectively. The analysis of the results of Table V follows.

1) HY outperforms KBTGs since it provides an average density of 49.932% whereas KBTGs realizes a percentage value of 48.120%.
2) MDSBH remains competitive since it improves most solutions reached by both KBTGs and HY. Indeed, it is able to improve five out of six best solutions while it matches the other solution (instance KBTG2) when compared to the results reached by HY.
3) For the improved solutions (except for the instance KBTG2 where all algorithms reach the optimal solution), MDSBH realizes an improvement varying from 0.019% (instance KBTG8) to 1.037% (instance KBTG7).
4) Globally, the average improvement over all instances is equal to 0.420%, as displayed by the last line of Table V.

V. Conclusion

In this paper the three-dimensional sphere packing problem is solved by using a modified dichotomous search-based heuristic. The proposed method is based upon three complementary phases: (i) a modified greedy selection phase which tries to select more eligible positions to iteratively pack all predefined items into the target object, (ii) a width beam search combined with hill-climbing strategies for exploring promising paths and (iii) a dichotomous search for providing a best target object, that is able to pack all items without overlapping. The performance of the modified algorithm was evaluated on benchmark instances available in the literature. The provided results were compared to those reached by the original version of the algorithm, as well as to the results given by some recently proposed heuristics. The new version of the method remains competitive and succeeded in yielding new solutions for many instances.

References

[1] E. G. Birgin and F.N.C. Sobral. Minimizing the object dimensions in circle and sphere packing problems. Computers & Operations Research, 35, 2357–2375, 2008 (DOI: 10.1016/j.cor.2006.11.002).

[2] F. Della Croce, M. Ghirardi and R. Tadei. Recovering beam search approach for combinatorial optimization problems. Journal of Heuristics, 10, 89–104, 2004 (DOI: 10.1023/B:HEUR.0000019987.10818.e0).

[3] R. S. Farr. Random close packing fractions of log-normal distributions of hard spheres. Powder Technology, 245, 28–34, 2013 (DOI: 10.1016/j.powtec.2013.04.009).

[4] M. Hifi and R. M'Hallah. A literature review on circle and sphere packing problems: models and methodologies. Advances in Operations Research, Article ID 150624, 22 p, 2009 (doi.org/10.1155/2009/150624).

[5] M. Hifi and R. M'Hallah. Beam search and non-linear programming tools for the circular packing problem, International Journal of Mathematics in Operational Research, 1, 476–503, 2009 (DOI: 10.1504/IJ-MOR.2009.026278).

[6] M. Hifi, R. M'Hallah and T. Saadi. Algorithms for the constrained two-staged two-dimensional cutting problem. INFORMS, Journal on Computing, 20 212–221, 2008.

[7] M. Hifi and T. Saadi. A cooperative algorithm for constrained two-staged two-dimensional cutting problems. International Journal of Mathematics in Operational Research, 9, 104–124, 2010 (DOI: 10.1504/IJOR.2010.034363).

[8] M. Hifi and L. Yousef. A dichotomous search-based heuristic for the three-dimensional sphere packing problem. Working paper, Exposed in the Seminary of ROAD Team, Laboratory EPROAD, Universi´ e de Picardie Jules Verne, october 2013.

[9] T. Kubach, A. Bortfeldt, T. Tilli, and H. Gehring. Parallel greedy algorithms for packing unequal spheres into a cuboidal strip or a cuboid. Working Paper, Department of Management Science, University of Magdeburg, (Diskussionsbeitrag der Fakult at f ur Wirtschaftswissenschaft der FernUniversit at in Hagen). No 440, Hagen 2009.

[10] T. Kubach, A. Bortfeldt, T. Tilli, and H. Gehring. Greedy algorithms for packing unequal sphere into a cuboidal strip or a cuboid. Asia-Pacific Journal of Operational Research, 28(06), 739–753, 2011 (DOI: 10.1142/S0217595911003326).

[11] Y. Li and W. Ji. Stability and convergence analysis of a dynamics-based collective method for random sphere packing. Journal of Computational Physics, 250, 373–387, 2013 (DOI: 10.1016/j.jcp.2013.05.023).

[12] K. Lochmann, L. Oger, and D. Stoyan. Statistical analysis of random sphere packings with variable radius distribution. Solid State Sciences. 8(12), 1397–1413, 2006 (DOI: 10.1016/j.solidstatesciences.2006.07.01).

[13] R. M'Hallah, A. Alkandari, and N. Mladenovi´ c. Packing unit spheres into the smallest sphere using VNS and NLP. Computers & Operations Research, 40(2), 603–615, 2013 (DOI: 10.1016/j.cor.2012.08.019).

[14] R. M'Hallah and A. Alkandari. Packing unit spheres into a cube using VNS. Electronic Notes in Discrete Mathematics, 39(1), 201–208, 2012.

[15] P. S. Ow and T.E. Morton. Filtered beam search in scheduling, International Journal of Production Research, 26, 297–307, 1988 (DOI:10.1080/00207548808947840).

[16] K. Soontrapa and Y. Chen. Mono-sized sphere packing algorithm development using optimized Monte Carlo technique. Advanced Powder Technology, 24(6), 955–961, 2013 (DOI: 10.1016/j.apt.2013.01.007).

[17] Y. Stoyan, G. Yaskow, and G. Scheithauer. Packing of various radii solid spheres into a parallelepiped. Central European Journal of Operational Research, 11, 389–407, 2003.

[18] A. Sutou and Y. Dai. Global optimization approach to unequal sphere packing problems in 3D. Journal of Optimization Theory and Applications, 114, 671–694, 2002 (DOI: 10.1023/A:1016083231326).

[19] J. Wang. Packing of unequal spheres and automated radio-surgical treatment planning. Journal of Combinatorial Optimization, 3, 453–463, 1999 (DOI: 10.1023/A:1009831621621).

[20] G. Wascher, H. Haussner and H. Schumann. An improved typology of cutting and packing problems. European Journal of Operational Research, 183, 1109–1130, 2007 (DOI: 10.1016/j.ejor.2005.12.047).

[21] M. Yavuza. Iterated beam search for the combined car sequencing and level scheduling problem. International Journal of Production Research, 51, 3698–3718, 2013 (DOI:10.1080/00207543.2013.765068).

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 429–438
DOI: 10.15439/2014F268
ACSIS, Vol. 2

Meta-optimization method for wavelet-based damage identification in composite structures

Andrzej Katunin, Piotr Przystałka
Silesian University of Technology,
Institute of Fundamentals of Machniery Design
ul. Konarskiego 18A, 44-100 Gliwice, Poland
Email: {andrzej.katunin,piotr.przystalka}@polsl.pl

***Abstract*—The damage identification problem is one of crucial problems during operation of machines' elements made of polymeric composites. Therefore the appropriate non-destructive techniques should be developed in order to detect and identify the damages with the best possible accuracy. Moreover, such methods should be applicable in various testing conditions. One of the intensively developed directions in non-destructive damage assessment is a class of methods based on wavelet analysis of modal shapes of vibration applied for a tested structure. The effectiveness of an algorithm is strongly dependent on the type of applied wavelet and its parameters. The proposed approach uses a combination of the wavelet-based damage identification algorithm with multi-objective meta-optimization in order to select optimal parameters of applied wavelets and determine a front of optimal non-dominated solutions. Based on these solutions the operator can choose the desired accuracy of damage identification with respect to the suitable computation time.**

I. Introduction

SINCE the polymeric composites are more and more applicable as constructional materials in various industrial branches (e.g. automotive, aircraft and aerospace industries) and the manufactured elements are often subjected to critical loads during their operation, the development of appropriate damage identification methods, which will be able to detect and identify the damages specific for these materials, seems to be a necessity. From the majority of recently developed and applied non-destructive methods and techniques one could select a group of methods, which are based on analysis of vibration data of a tested structure. These methods have several advantages with respect to others, e.g. the possibility of carrying out on-field diagnostics, simplicity of measurements and concluding about the damage presence, a possibility of performing the measurements without unmounting the tested element from the machine, etc. However, for increasing the accuracy of detection and identification of damages the advanced signal processing methods are usually used.

One of the intensively developed approaches in the damage assessment problems is a wavelet-based analysis. Since the most of these problems are referred to the structural diagnostics of spatial domains the wavelet-based algorithms were extended for application on two-dimensional data. Numerous studies, both theoretical and experimental, were based on various wavelet transforms and various wavelets in order to obtain relevant information about damage state of a tested structure. A number of researchers developed their algorithms based on continuous wavelet transform (CWT) [1], [2], [3], [4] or stationary wavelet transform (SWT) [5]. The analysis of the applied wavelet-based algorithms was presented in [6]. The previous comparison studies of various wavelet transforms and various wavelets applied in the damage identification algorithm [7] show that the most accurate and computationally efficient algorithm is provided by application of the discrete wavelet transform (DWT) together with B-spline wavelets.

Following the recent advances in the field of improvement of wavelet-based damage identification methods one can notice that they are generally based on hybridization of the mentioned algorithm with various soft computing methods. There are numerous hybridizations of wavelet-based algorithm with artificial neural networks [8], [9], [10] as well as optimization algorithms: Krawczuk et al. [11] use genetic algorithms for improvement of cracks detection in beams, while the authors of [12] applied particle swarm optimization for improvement of detectibility of damages.

Further studies [14] of the first author allow developing a more efficient algorithm, which was based on fractional discrete wavelet transform (FrDWT) introduced in [15] with application of fractional B-spline wavelets. Based on the application of extended (two-dimensional) version of the fractional B-spline wavelets proposed in [16] it was possible to select the wavelet parameters suitable for the investigated problem in order to achieve the most accurate results of damage identification, which constitutes an improvement of accuracy with respect to integer-valued order of B-spline wavelets used previously [7], [17]. In order to select optimal parameters of the applied 2D fractional B-spline wavelets the authors hybridized a wavelet-based damage identification algorithm with various optimization algorithms (evolutionary algorithm, direct search algorithm, simulated annealing algorithm and particle swarm optimization) [18], which allowed for the further improvement of the damage identification effectiveness. The method was validated on numerical models and on experimental data achieved from vibrometric measurements of artificially damaged composite structures. Moreover, in the same paper, the authors surveyed the literature regarding the optimization problem of the wavelets' parameters mainly for

The research project was financed by the National Science Centre (Poland) granted according the decision no. DEC-2011/03/N/ST8/06205.

the damage identification purposes. It was noted that, the optimization approaches might be included into two groups: classic methods and soft computing and heuristic ones. One of the main conclusions from this review was that there was the lack of sufficient methods to tune behavioural parameters of optimization algorithms in order to have a much more practical algorithm for damage identification, which could be implemented in the embedded system of the end-user device. In order to improve application abilities of the wavelet-based damage identification algorithm and evaluate the effectiveness of optimized wavelet parameters (and thus the effectiveness of damage detection and identification) with respect to the computational time the problem could be formulated as multi-objective meta-optimization one.

The meta-optimization is a quite novel approach, which found numerous applications in the engineering problems. One of the earliest attempts to meta-optimization can be found in [19], where the genetic algorithm was used in order to find best mutation and crossover rates for another lower-level genetic algorithm. In the next years, there were similar trials to this problem by many authors, e.g. see [20], [21], [22]. Also in this subject, the authors of the paper [23] discussed the most important issues related to tuning evolutionary algorithm parameters by means of various meta-optimization methods. Their main conclusion was that it was no matter what kind of tuner algorithms to be used in this task, because for each case, it was possible to get a much better result from evolutionary computations with meta-optimization than relying on own intuition and the usual parameter setting conventions. The similar strategy as in the case of evolutionary algorithms can be observed for other soft computing method. For example in [24], the authors proposed the concept in which a superordinate swarm ('superswarm') can be used to optimize the parameters of subordinate swarms ('subswarms'). Subordinate swarms were used for neural network training. Another point of the view is given in [25]. Branke and Elomari in their work proposed the method that could be used, in a single run, to identify the best parameter settings for all possible computational budgets. Their approach allows to save a lot of time. In the best of authors' knowledge the only application of meta-optimization in non-destructive testing with use of wavelet-based algorithm was presented in [26], where the authors performed electromagnetic measurements with appropriate post-processing in order to detect and identify cracks in walls of nuclear fission reactors.

In this study the authors developed an existing hybridized algorithm based on the results of the previous study [18]. The application of meta-optimization to the wavelet-based damage identification algorithm has several goals. In spite of the computational procedure implemented in [18] the authors determined common wavelets' parameters for all types of investigated damages. The parent optimization sub-algorithm in the meta-optimization algorithm was based on a double criterion problem, which allows to obtain a front of optimal non-dominated solutions dependent on the accuracy of damage identification vs. the computation time. Thus, one may decide which strategy should be applied, for instance, the worse solution with quick data processing or the best solution with long-time data processing.

II. Damage identification

A. Wavelet-based algorithm

The algorithm of damage detection and identification was based on spatial FrDWT, which uses the two-dimensional Mallat's multi-resolution analysis, where B-spline scaling functions of fractional order $\beta_\tau^\alpha(x)$ constitute a space of the square-integrable subspaces $L^2\left(\mathbb{R}^2\right)$ and form a sequence of functional spaces V_i in the form:

$$\{0\} \subset \ldots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \ldots \subset L^2\left(\mathbb{R}^2\right). \tag{1}$$

The form of a scaling function of fractional order $\beta_\tau^\alpha(x)$ is defined by two parameters [15]: $\alpha \in \mathbb{R}$, which is an order of scaling function, and $\tau \in \mathbb{R}$, which is a shift parameter, and is as follows:

$$\beta_\tau^\alpha(x) = \sum_{k=0}^{\infty}(-1)^k \left| \begin{array}{c} \alpha+1 \\ k-\tau \end{array} \right| \rho_\tau^\alpha(x-k), \tag{2}$$

where

$$\rho_\tau^\alpha(x) = -\frac{\cos\pi\tau}{2\Gamma\left(\alpha+1\right)\sin\left(\pi\alpha/2\right)}\left|x\right|^\alpha - \frac{\sin\pi\tau}{2\Gamma\left(\alpha+1\right)\cos\left(\pi\alpha/2\right)}\left|x\right|^\alpha \operatorname{sgn}(x), \tag{3}$$

$\Gamma\left(\alpha+1\right)$ is the Euler γ-function, which allows for fractional factorization. The scaling $\beta_\tau^\alpha(x)$ and wavelet $\psi_\tau^\alpha(x)$ functions hold two-scale relations [15], [16]. For the cases when $\alpha \notin \mathbb{Z}$ and $\tau = \left(\alpha+1\right)/2$ the integer-valued B-spline wavelets can be obtained.

Following the method of complexification of B-spline wavelets of fractional order based on generation of Hilbert transform pairs of them proposed in [16] it is possible to obtained direction-oriented 2D complex wavelets. The complexification is based on a combination of wavelets with the same order α, but different shift parameters τ in the form:

$$\psi_\tau^\alpha(x) = \psi_\tau^\alpha(x) + j\psi_{\tau+1/2}^\alpha(x), j^2 = -1. \tag{4}$$

Considering that the analytic wavelet has a form [16]:

$$\psi^\alpha(x) = \psi + j\mathcal{H}\left\{\psi\right\}, \tag{5}$$

where $\mathcal{H}$ denotes a Hilbert transform, the 2D complex wavelets have the following form [16]:

$$\begin{aligned}
\psi_1\left(\mathbf{X}\right) &= \psi(x)\phi(y) + j\mathcal{H}\psi(x)\phi(y),\\
\psi_2\left(\mathbf{X}\right) &= \psi(x)\mathcal{H}\phi(y) + j\mathcal{H}\psi(x)\mathcal{H}\phi(y),\\
\psi_3\left(\mathbf{X}\right) &= \phi(x)\psi(y) + j\phi(x)\mathcal{H}\psi(y),\\
\psi_4\left(\mathbf{X}\right) &= \mathcal{H}\phi(x)\psi(y) + j\mathcal{H}\phi(x)\mathcal{H}\psi(y),\\
\psi_5\left(\mathbf{X}\right) &= 2^{-1/2}\left(\psi(x)\psi(y) - \mathcal{H}\psi(x)\mathcal{H}\psi(y)\right)\\
&\quad +2^{-1/2}j\left(\psi(x)\mathcal{H}\psi(y) + \mathcal{H}\psi(x)\psi(y)\right),\\
\psi_6\left(\mathbf{X}\right) &= 2^{-1/2}\left(\psi(x)\psi(y) + \mathcal{H}\psi(x)\mathcal{H}\psi(y)\right)\\
&\quad +2^{-1/2}j\left(\psi(x)\mathcal{H}\psi(y) - \mathcal{H}\psi(x)\psi(y)\right),
\end{aligned} \tag{6}$$

where $\mathbf{X} = (x, y)$ denotes 2D signal and $2^{-1/2}$ is used for scaling the wavelets. The wavelets (6) are oriented along the primal direcrions: $\theta_1 = \theta_2 = 0$, $\theta_3 = \theta_4 = \pi/2$, $\theta_5 = \pi/4$ and $\theta_6 = 3\pi/4$, which allows for increasing detectability of damages located in these directions. The decomposition graphical example can be found in [18].

Considering that discrete-type wavelet transforms could be expressed as a set of high-pass and low-pass filters, the decomposition procedure could be presented in the form of pairs of filters along the specific directions of a signal $\mathbf{X}$. Considering (4) and (6) one can obtain six complex subbands $\mathbf{w}^P$, $P = 1, \ldots, 6$, after the decomposition (see [16], [18] for details).

In order to ensure sensitivity of an algorithm to all possible orientations of spatial damages the real parts of $\mathbf{w}^P$ are normalized according to the Euclidean metric. Moreover, considering the strong dependence between magnitudes of displacements of modal shapes and obtained coefficients after decomposition it is suitable to consider multiple modal shapes during the analysis. The resulted post-processing expression with respect to M considered modes takes a form:

$$\mathbf{W} = \sum_{M} \left| \sum_{P} \Re \left(\mathbf{w}_M^P \right)^2 \right|. \tag{7}$$

Based on the above-presented algorithm the damage identification procedure was performed. As was mentioned earlier, the wavelets were defined by α and τ, thus these parameters were selected for the optimization procedure. The goal of optimization is to find suitable values for α and τ, which allow for obtaining of best results in detection and localization of damages.

B. Optimization procedure for searching values of α and τ

The main goal of the optimization procedure is to adjust the fractional order α and the shift factor τ in order to obtain the best properties of the damage identification algorithm. In the previous paper [18], the authors proved that the single optimization method could be successfully applied in this kind of tasks. Hence, the optimization problem can be written as follows:

$$\begin{array}{ll} \text{Minimize} & U(\alpha, \tau) \\ \text{subject to} & \alpha^{(L)} \le \alpha \le \alpha^{(U)}, \tau^{(L)} \le \tau \le \tau^{(U)} \end{array} \tag{8}$$

where $\alpha^{(L)}, \alpha^{(U)}, \tau^{(L)}, \tau^{(U)}$ are the lower and upper values of the boundary constraints that should be chosen taking into account the properties of the wavelet. The global criterion method [27] is used to create a single objective function $U(\alpha, \tau)$. In this way, an indirect utility function can be expressed in its simplest form as the weighted exponential sum:

$$U(\alpha, \tau) = c_1 \left[1 + \sum_{i=1}^{T(p)} \max(\mathbf{W}_i) \right]^{-\lambda} + c_2 \left[\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}^* \right]^{\lambda} \tag{9}$$

where c_i are weights indicating the relative significance of elements in the sum, the exponent λ determines the extent to which a method is able to capture all of the Pareto-optimal points for either convex or non-convex criterion spaces, detail coefficients in the matrix $\mathbf{W}_i$ are computed using the recurrence relation proposed in [18], $w_{i,j}^*$ is an element of the matrix $\mathbf{W}_{T(p)}$. The function $T(p)$ can be defined using the following expression:

$$T(p) = \text{card}\left\{ w_{i,j} : \forall i, j \in \{1, 2, \ldots, n\} \;\; w_{i,j} \ge p \max(\mathbf{W}) \right\} \tag{10}$$

where n is the size of the matrix $\mathbf{W}$, whereas p is the ratio between the greatest magnitudes and the other detail coefficients in the matrix $\mathbf{W}$. The value of this parameter should be chosen arbitrarily from the range $[0, 1]$.

It is very important to have the physical interpretation of the formulated objectives. The first component of the weighted exponential sum (9) describes the values of detail coefficients that have the greatest magnitudes. This criterion can be interpreted as follows. If the damage occurred somewhere in the composite plate then the result of this would be locally affected on the values of displacements of modal shapes. The second component of the sum is correlated with the blurring of the regular form or forms which indicate the damage. It can be stated that the values of $w_{i,j}^*$ in the matrix represent the blurring of the regular form.

The authors showed in their previous paper [18] that the problem, which has been formulated in the form of (9) could be solved using heuristic optimization algorithms. As it was presented in their work, an evolutionary algorithm, a direct search algorithm, a simulated annealing algorithm and a particle swarm optimization algorithm could be adopted. The main problem in this kind of approaches is to find the compromise between the time computational complexity of an algorithm and the accuracy of a solution. These factors are strongly dependent on properties and values of the relevant parameters (behavioural parameters) of these algorithms.

III. META-OPTIMIZATION METHOD

The idea of meta-optimization which is also known in the literature as super-optimization or hyper-heuristic is to apply one optimization technique to adjust another optimization technique. In this paper, the meta-optimization strategy corresponding to the data flow diagram presented in Fig. 1 is used in order to search the space of behavioural parameters. As it can be seen, the meta-optimization algorithm (MAC) evaluates a meta-objective function whereas the main optimization algorithm (OAE) computes the cost function in order to find an optimal solution with the minimum time complexity and maximum accuracy.

Meta-optimization concept can be realized in different ways, however one of the most promising approaches employs the multi-objective optimization algorithm. Consequently, the main purpose of the meta-optimization process is to tune values of behavioural parameters of the main optimization

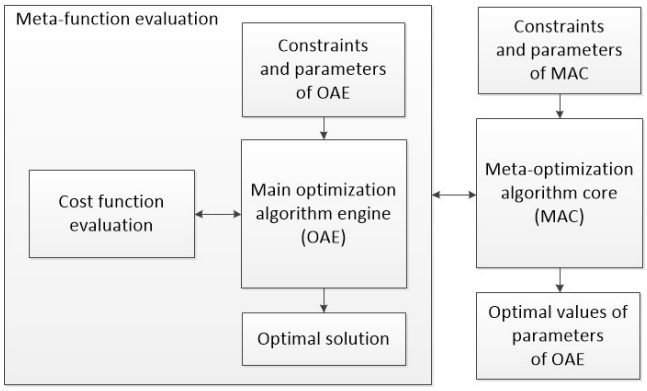


Fig. 1: A data flow diagram of the meta-optimization method

algorithm in order to minimize a multiple objective function $\mathbf{U}$. This function can be formulated taking into account two fundamental criteria. The first criterion (MBF) corresponds to the estimation of the accuracy of a solution, whereas the second one (FES) represents the time computational complexity of the main optimization algorithm. When one assumes that both objectives are not conflicted then the multi-objective meta-optimization problem can be stated as follows:

$$\begin{array}{ll} \text{Minimize} & \mathbf{U}(\xi) = [\text{MBF}(\xi)\ \ \text{FES}(\xi)] \\ \text{subject to} & \Omega(\xi) \end{array} \tag{11}$$

where ξ is the set of properties of the main algorithm, Ω represents boundaries and constraints in the meta-optimization process. The accuracy of the solution that is found by the main optimization algorithm, can be computed as the mean value of best scores of the cost function evaluations. On the other hand, the time complexity is approximated using the total number of cost function evaluations in the same algorithm.

Generally, multi-objective optimization problems do not have single global solution, and therefore there is the need to investigate a set of points, each of which satisfies the objectives. Due to this, in the present study, predominant Pareto optimality concept is mainly used. A solution is Pareto optimal if there is no other solution that improves at least one objective function without detriment another function [27]. It is often viewed the same as a non-dominated solution.

It is reasonable to expect that each of multi-objective versions of soft computing methods indicated in the previous section to be applicable in the task of meta-optimization. Nevertheless, the authors propose to use a much less complicated algorithm in the main optimization engine, while a more advanced approach in the meta-optimization core. In such manner, it is possible to obtain general values of relevant parameters of the main algorithm that can easily be implemented in the embedded system of the end-user device.

IV. Research results

The advantages and limitations of the proposed meta-optimization method were attempted in two separate experiments. The aim of the first case study was to validate the performance of the meta-optimization approach over a set of well-practised test functions. The second experiment dealt with the useful application of the elaborated method for wavelet-based damage identification in composite structures. It was decided that, the engine of the main optimization algorithm was prepared using the particle swarm optimization algorithm (PSO-OAE), while the core of the meta-optimization process was implemented by means of the multi-objective evolutionary algorithm (MOEA-MAC). MOEAs are known in the literature as the heuristic methods for solving optimization problems, which are based on the natural selection process that mimics biological evolution. The MOEA recommended in [28] is utilized herein to solve the meta-optimization problem defined as (11). Well-known and often practised genetic operators for multi-objective optimization are applied to obtain the convergence of a solution. In such manner, the problem of finding values of behavioural parameters is solved by computing the Pareto front, hence the set of evenly distributed non-dominated optimal solutions are determined. PSO is also classified into heuristic approaches, however this is a population-based stochastic optimization technique, which is inspired by simulation of social behaviour. In this paper, PSO proposed by [29] is adopted and applied to search for the optimal values of α and τ. The both optimization algorithms were implemented in the MATLAB® environment using Genetic Algorithm and Particle Swarm Optimization Toolboxes.

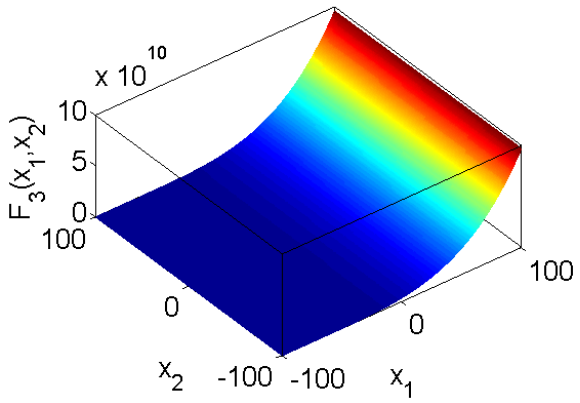
A. Benchmark tests

In the first step, the proposed method was verified using selected benchmark functions proposed in the CEC'2008 Special Session and Competition on Large Scale Global Optimization [30]. Due to the nature of the main problem formulated in this study, the authors decided to select the following test functions:

- F_3 : Shifted Rosenbrock's function

$$F_3(\mathbf{x}) = \sum_{i=1}^{D-1} \left(100\left(z_i^2 - z_{i+1}\right)^2\right) \tag{12}$$

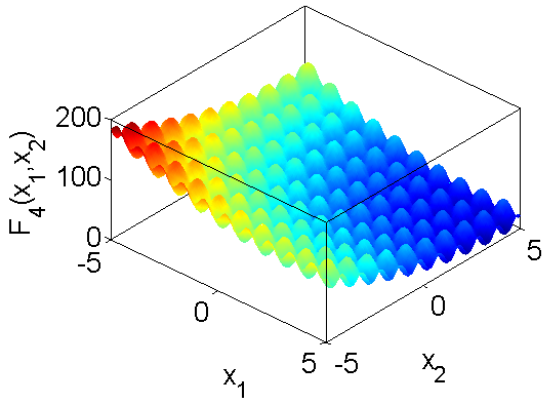
where $\mathbf{z} = \mathbf{x} - \mathbf{o} + \mathbf{1}, \mathbf{x} = [x_1, x_2, \ldots, x_D]$, $\mathbf{x} \in [-100, 100]^D$, $\mathbf{o} = [o_1, o_2, \ldots, o_D]$ is the shifted global optimum $\mathbf{x}^* = \mathbf{o}, F_3(\mathbf{x}^*) = 0$.



- F_4 : Shifted Rastrigin's function

$$F_4(\mathbf{x}) = \sum_{i=1}^{D} \left(z_i^2 - 10\cos(2\pi z_i) + 10\right) \qquad (13)$$

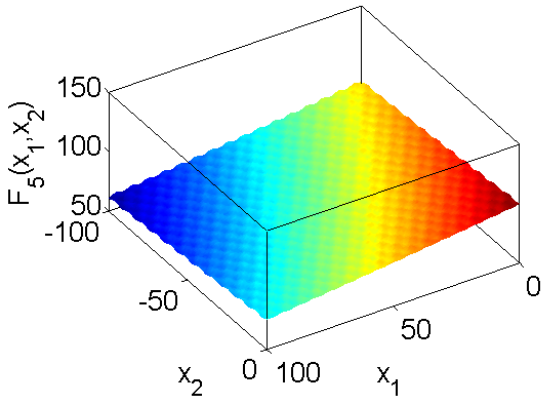
where $\mathbf{x} \in [-5,5]^D$, $\mathbf{z}$ is the same as in the previous function, the shifted global optimum $\mathbf{x}^* = \mathbf{o}, F_4(\mathbf{x}^*) = 0$.



- F_5 : Shifted Griewank's function

$$F_5(\mathbf{x}) = \sum_{i=1}^{D} \frac{z_i^2}{4000} - \prod_{i=1}^{D} \cos\left(\frac{z_i}{\sqrt{i}}\right) + 1 \qquad (14)$$

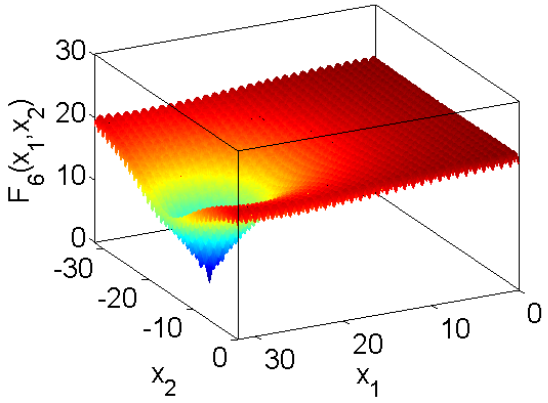
where $\mathbf{z} = (\mathbf{x} - \mathbf{o})$, $\mathbf{x} \in [-600,600]^D$, the shifted global optimum $\mathbf{x}^* = \mathbf{o}, F_4(\mathbf{x}^*) = 0$.



- F_6 : Shifted Ackley's function

$$F_6(\mathbf{x}) = -20\exp\left(-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D} z_i^2}\right) - \exp\left(\frac{1}{D}\sum_{i=1}^{D}\cos(2\pi z_i)\right) + 20 + e \qquad (15)$$

where $\mathbf{x} \in [-32,32]^D$, $\mathbf{z}$ is the same as in F_5 function, the shifted global optimum $\mathbf{x}^* = \mathbf{o}, F_6(\mathbf{x}^*) = 0$.



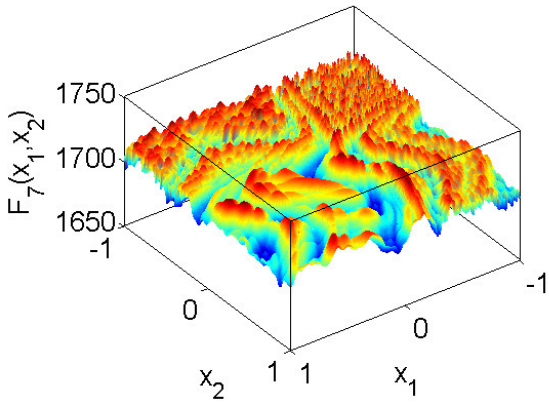
- F_7 : FastFractal 'DoubleDip' function

$$F_7(\mathbf{x}) = \sum_{i=1}^{D} \lambda_1\left(x_i + \lambda_2\left(x_{(i \bmod D)+1}\right)\right) + 1720 \qquad (16)$$

where
$\lambda_1(x) \approx \sum_{k=1}^{3}\sum_{1}^{2^{k-1}}\sum_{1}^{\hat{o}_2}\lambda_3\left(x, \hat{o}_1, \frac{1}{2^{k-1}(2-\hat{o}_1)}\right)$,
$\lambda_2(y) = 4\left(y^4 - 2y^3 + y^2\right)$,

$$\lambda_3(x,c,s) = \begin{cases} \left(-6144(x-c)^6 + 3088(x-c)^4 \right. \\ \left. -392(x-c)^2 + 1\right)s, x \in (-0.5, 0.5) \\ 0, \text{otherwise} \end{cases},$$

and $\mathbf{x} \in [-1,1]^D$, the global optimum is unknown, $F_7(\mathbf{x}^*)$ is also unknown, $\hat{o}_1$ is a double precision variable, pseudo-randomly chosen, with seed o_1, with equal probability from the interval [0,1], $\hat{o}_2$ is an integer variable, pseudo-randomly chosen, with seed o_2, with equal probability from the set $\{0,1,2\}$.



The task of the main optimization was defined as a continuous minimization problem. As it was mentioned above, the extreme was found with the help of the PSO-OAE. The cost function in the PSO-OAE was computed using one of the benchmark functions (12-16). In this algorithm only few parameters of the algorithm are relevant to guarantee, as far as possible, to find the optimal solution of the problem. Therefore, behavioural parameters such as the population size, the total number of generations, the social and cognitive

attraction coefficients were taken into account during the meta-optimization process realized by means of MOEA-MAC. For each function, the lower and upper boundaries of the PSO-OAE were assumed corresponding to functions' properties (12-16), whereas in the case of MOEA-MAC boundaries were declared in such a way, that the population size was equal from D to $2D$, the total number of generations from 5 to 100, the cognitive and social attraction from 0 to 4. The heuristic rules given in the literature were employed to get the best results from the MOEA. The fitness function was declared following to (11), where MBF was computed by averaging the best fitness function (for ten trials) and FES was obtained as the product of the population size and the total number of generations of the OAE. It was decided that individuals in the population of the MOEA were composed of genes representing real numeric values of behavioural parameters (the integer parts of the population size and generations parameters were used during the computations). The total number of generations of MOEA was set to 30. The population size of this algorithm was equal 40. The feasible population method was adapted to create a random well-dispersed initial population that satisfies all bounds in (11). Fitness scaling was realized using the rank method, whereas the selection of the parents to the next generation was achieved by applying the stochastic uniform method. Additionally, two reproduction options – the elite count and crossover fraction were chosen. The first one specifies the number of individuals that are guaranteed to survive to the next generation (it was equal 2). The second deals with the fraction of the next generation, other than elite children, that are produced by crossover. It was decided to use a heuristic crossover operator where the user-defined parameter was set to 1.2, and the crossover probability was equal 0.8. The remaining individuals are mutation children and they were obtained using the adaptive feasible method.

TABLE I: Optimal values of behavioural parameters for benchmark functions

Function	D	Cognitive attraction	Social attraction	Generations	Population size
F_3	2	2.180	0.706	74	4
	100	1.019	1.164	99	196
F_4	2	0.789	2.267	69	4
	100	1.393	1.319	96	100
F_5	2	0.712	1.125	88	4
	100	1.362	1.039	99	190
F_6	2	0.350	0.826	100	4
	100	1.290	1.375	81	188
F_7	2	0.988	0.920	67	4
	100	1.055	1.474	99	142

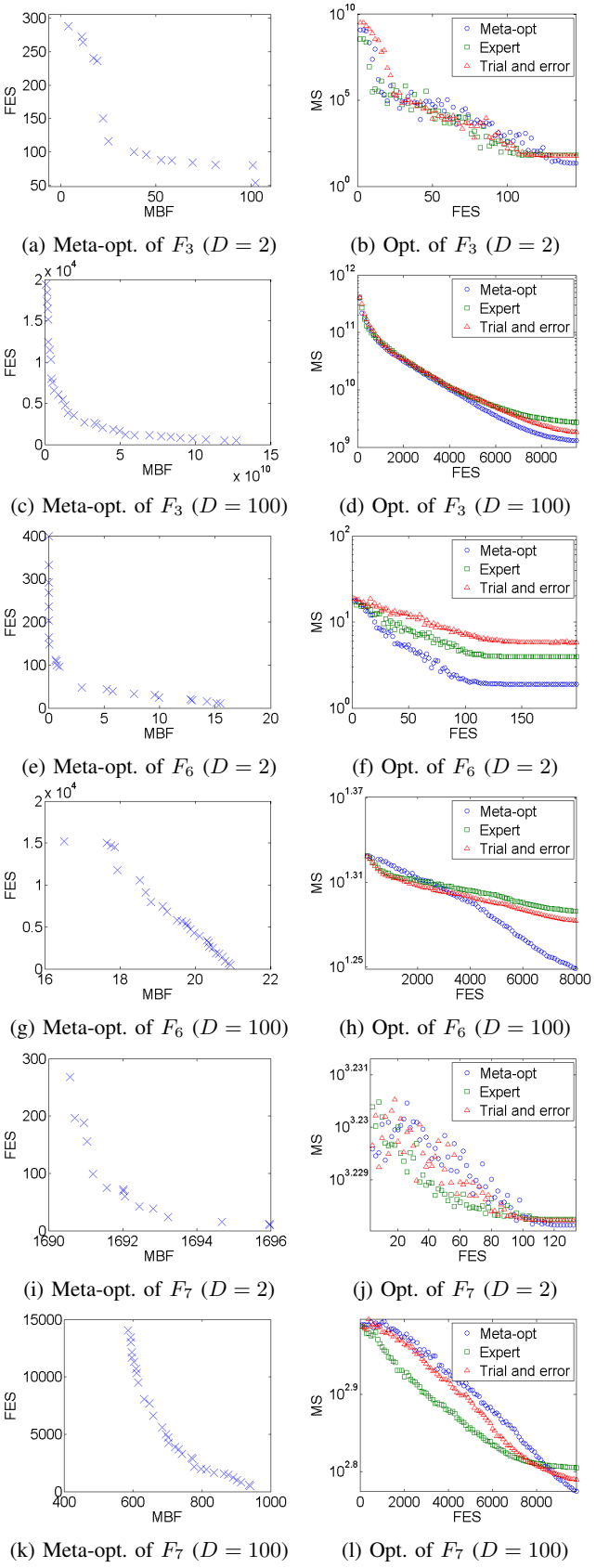
The meta-optimization process was carried out for two cases $D = 2$ and $D = 100$. The achieved results are presented in Tab. I. Besides, in Figs. 2(a,c,e-k) there are given graphs with the visualisation of selected Pareto fronts (for functions F_3, F_6 and F_7, $D = 2$ and $D = 100$, respectively) based on which the optimal values of behavioural parameters were chosen. In this case study, the authors selected non-dominated optimal solutions that were characterized by the highest accuracy of the cost function (in a statistic sense) with the minimum as possible time complexity of the algorithm. In order to have much more understandable and comparable results the tuning of behavioural parameters was also carried out with the use of expert's knowledge and trial and error procedure. In the first case, the suggestions proposed in [31] were applied (cognitive attraction = 0.5, social attraction = 1.25). In the second case values of behavioural parameters were changed several times for obtaining satisfactory solutions. The optimization process was run ten times for each case and afterwards the results were averaged. Overall, the comparison results of meta-optimization (○) and classic strategies (□, △) for adjusting behavioural parameter values were included in Tab. II. The most important statistic measures such as AVG and STD show that the best option is to find optimal values of behavioural parameters by means of the meta-optimization method. It is also confirmed by results presented in Figs. 2b,d,f-l (for functions F_3, F_6 and F_7, $D = 2$ and $D = 100$, respectively). These plots demonstrate mean values of the best scores of the cost function (MS) vs. the number of function evaluations (FES) for investigated cases. Each of these examples illustrates the effectiveness of the proposed meta-optimization method when it is compared to classic approaches.

TABLE II: Optimization results for different selection strategies of the behavioural parameter values

Function	Case	MAX	MIN	AVG	STD
F_3, $D = 2$	○	5.167E+01	2.607E-01	2.260E+01	2.018E+01
	□	1.470E+02	5.549E-02	6.384E+01	5.887E+01
	△	1.512E+02	3.662E-01	6.182E+01	5.185E+01
F_3, $D = 100$	○	3.032E+09	5.057E+08	1.274E+09	7.103E+08
	□	5.000E+09	1.382E+09	2.675E+09	1.345E+09
	△	3.734E+09	7.893E+08	1.773E+09	8.167E+08
F_4, $D = 2$	○	8.955E+00	1.644E-03	2.451E+00	3.071E+00
	□	2.487E+01	1.079E-06	6.369E+00	7.404E+00
	△	2.487E+01	3.473E-07	3.880E+00	7.627E+00
F_4, $D = 100$	○	9.684E+02	7.938E+02	8.869E+02	6.705E+01
	□	1.078E+03	9.148E+02	1.010E+03	4.789E+01
	△	1.115E+03	8.727E+02	1.005E+03	7.889E+01
F_5, $D = 2$	○	2.440E-01	1.972E-02	7.867E-02	7.536E-02
	□	3.254E-01	2.932E-02	1.156E-01	1.026E-01
	△	3.428E-01	8.386E-03	9.869E-02	1.232E-01
F_5, $D = 100$	○	3.916E+02	1.784E+02	2.490E+02	6.166E+01
	□	4.518E+02	1.879E+02	3.076E+02	7.058E+01
	△	4.788E+02	2.579E+02	3.769E+02	7.286E+01
F_6, $D = 2$	○	1.890E+01	4.756E-07	1.890E+00	5.977E+00
	□	1.993E+01	2.050E-06	3.965E+00	8.359E+00
	△	2.030E+01	7.201E-06	5.799E+00	9.361E+00
F_6, $D = 100$	○	2.052E+01	1.569E+01	1.774E+01	1.468E+00
	□	2.021E+01	1.898E+01	1.949E+01	4.244E-01
	△	2.004E+01	1.735E+01	1.919E+01	8.287E-01
F_7, $D = 2$	○	1.693E+03	1.690E+03	1.691E+03	6.175E-01
	□	1.694E+03	1.690E+03	1.691E+03	1.057E+00
	△	1.692E+03	1.690E+03	1.691E+03	8.257E-01
F_7, $D = 100$	○	6.319E+02	5.455E+02	5.944E+02	2.676E+01
	□	7.220E+02	5.453E+02	6.383E+02	4.941E+01
	△	6.691E+02	5.575E+02	6.163E+02	3.442E+01

B. Description of the damage identification problem

The testing data was achieved during experimental measurements (modal analysis) of artificially damaged square composite plates clamped on the edges. The damages with depth of 0.5 mm (ca. 19% of total thickness) were included using numerical milling machine. In the first case there was through-the-length crack, in the second case there was a spatial square damage and in the last case there were multiple



(a) Meta-opt. of F_3 ($D = 2$) (b) Opt. of F_3 ($D = 2$)

(c) Meta-opt. of F_3 ($D = 100$) (d) Opt. of F_3 ($D = 100$)

(e) Meta-opt. of F_6 ($D = 2$) (f) Opt. of F_6 ($D = 2$)

(g) Meta-opt. of F_6 ($D = 100$) (h) Opt. of F_6 ($D = 100$)

(i) Meta-opt. of F_7 ($D = 2$) (j) Opt. of F_7 ($D = 2$)

(k) Meta-opt. of F_7 ($D = 100$) (l) Opt. of F_7 ($D = 100$)

Fig. 2: The comparison results of using meta-optimization, expert, trail and error procedures in selection of the behavioural parameter values

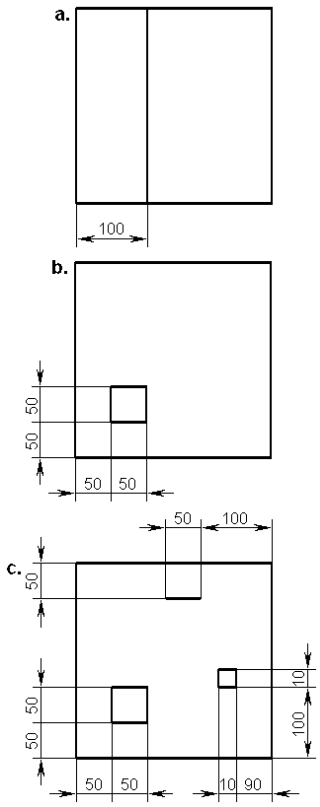


Fig. 3: Damaged plates considered in the analysis

damages: one small crack and two square spatial damages. The specific dimensions of damaged plates were presented in Fig. 3. During the scanning procedure the displacements in the net of 64×64 equidistant points were collected. The details of experimental setup and performing measurements can be found in [18].

First five modal shapes of each investigated case were considered in further analysis. Then, the collected data was exported to MATLAB® environment.

C. Results of damage identification

In this case study, the meta-optimization process was carried out on data collected using finite element (FE) analysis. The numerical models were prepared according to the geometry specification presented in Fig. 3 using MSC.Marc/Mentat® FE commercial software. The plates were modelled as 3D structures with the lay-up of a laminate and respective material properties presented in [17] and meshed using 8-node hexagonal elements. The boundary conditions were the same as for experimental study, i.e. all of the edges were clamped. The analyses were defined as normal mode evaluation, where the displacements in normal direction to the surface of a plate in 64×64 equidistant points of the first five modal shapes were considered for further studies. Due to the above-presented problem definition the numerical data was used as training data.

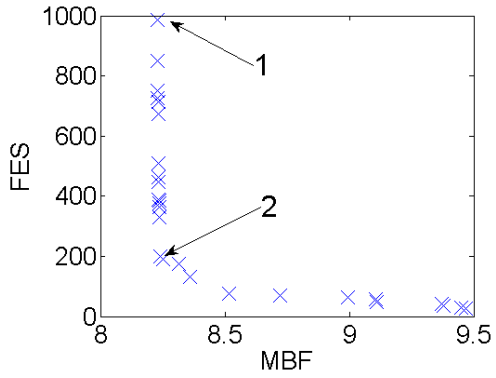


Fig. 4: Pareto front with optimal solutions obtained in the meta-optimization process

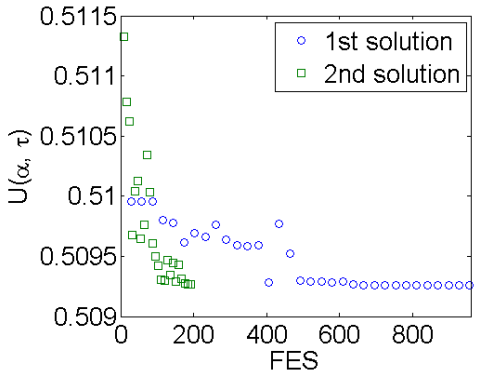
TABLE III: Optimal values of behavioural parameters for the damage identification purpose

No.	Cognitive attraction	Social attraction	Generations	Population size	FES	MBF
1	0.4324	0.8390	34	29	986	8.2307
2	0.1353	1.0078	25	8	200	8.2411

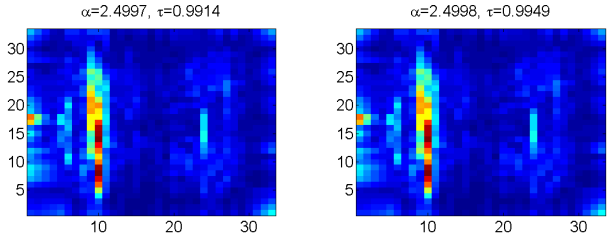
The extreme was found with the help of the PSO-OAE. The cost function in the PSO-OAE was evaluated using the indirect utility function (9) for three damaged plates at the same time. The lower and upper boundaries for α and τ in the PSO-OAE were assumed taking into account wavelet's properties ($\alpha^{(L)} = \tau^{(L)} = 0$, $\alpha^{(U)} = 2.5$,$\tau^{(U)} = 6$). Behavioural parameters were selected during the meta-optimization process realized by means of MOEA-MAC. The boundaries for these variables were declared in such a way, that the population size as well as the total number of generations were equal from 5 to 35, the cognitive and social attraction were equal from 0 to 4. The rest of the features of MOEA-MAC were selected in the same way as in the previous case.

The key results from the meta-optimization were shown in Fig. 4. This plot presents the Pareto front that means the set of non-dominated solutions. Due to the form of the plot it was possible that two optimal solutions were chosen for the further analysis. The values of behavioural parameters for these cases were included in Tab. III. It should be easily noted, that almost the same values of MBF can be achieved with the smaller number of function evaluations FES.

The damage identification experiments and the main optimization process were also repeated for real-world data. The values of behavioural parameters in this instance were the same as for the numerical data. Despite this, it was enough to obtain the high performance of damage identification for real measurements. Figs. 5a, 6a, 7a illustrate the main optimization processes (PSO-OAE) conducted applying the 1st and 2nd set of optimal values of behavioural parameters. In this way, it was possible for each case to obtain such values of parameters for which the total number of function evaluations was not larger than 200 to be enough to find the final solution.

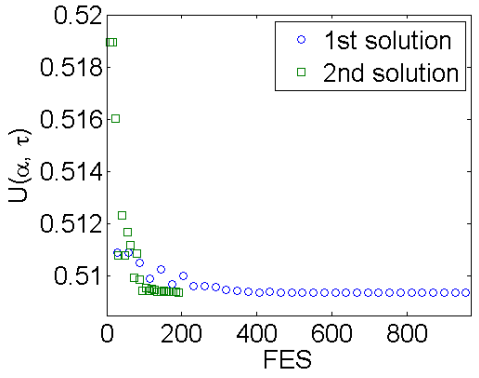


(a)

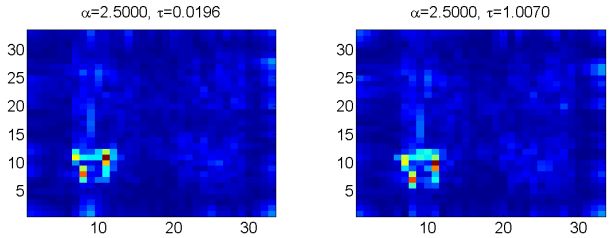


(b) for the 1st solution (c) for the 2nd solution

Fig. 5: The comparison of two non-dominated solutions from meta-optimization in the identification of the first damage

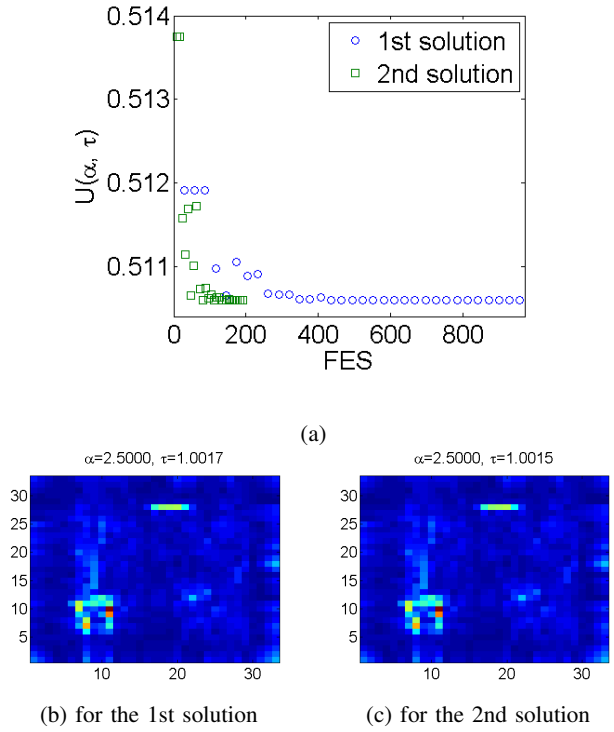


(a)



(b) for the 1st solution (c) for the 2nd solution

Fig. 6: The comparison of two non-dominated solutions from meta-optimization in the identification of the second damage



(b) for the 1st solution (c) for the 2nd solution

Fig. 7: The comparison of two non-dominated solutions from meta-optimization in the identification of the third damage

From the Pareto fronts presented in Figs. 5a-7a one can conclude that the compromise between time complexity of computations and the accuracy of the final solution has been reached. It is also conducted by the obtained results (see Figs. 5b,c-7b,c), where the optimized parameters of a wavelet in solutions obtained by the proposed method have almost identical values with those obtained using typical optimization procedure (1st solution), however these values were obtained much earlier in the 2nd solution than in the 1st one. It could be also noticed that the obtained values for α and τ cause that the applied wavelet has a great filtering ability and simultaneously short effective support, which cause that all of the damages were detected and located properly. The optimization algorithm solves the problem with selection of wavelet parameters, which influences much on the ability of damage detection and localization, i.e. in the case of underestimated value of α the obtained sets of coefficients are highly biased by the rests of a filtered signal due to inappropriate filtration, while in the case of overestimation of α the resulted sets of coefficients are blurred in the locations of singularities (damages) due to the power losses of a wavelet with the wide effective support, which makes the decision process about damage presence and location difficult and unambiguous.

V. Conclusions

In this paper, the authors proposed a new method for finding optimal values of behavioural parameters of the optimization procedure that is used in order to identify damages in composite structures. The proposed approach is based on the meta-optimization concept. The novelty of the proposed method depends on that the meta-optimization can be realized using multi-objective cost functions.

The preliminary verification of the elaborated approach in optimization tasks was carried out for well-known benchmark functions. The results achieved in this part of investigations demonstrate the capabilities of the approach for solving different kinds of optimization problems. The fundamental verification was conducted for the experimental data measured during tests on the artificially damaged composite plates. The problem of optimization of wavelet parameters applied for the structural damage assessment in composites was studied before by the authors, but the meta-optimization approach allows to obtain several new advantages. The parameters of optimization algorithm do not require to be known since they are determined by the meta-optimization procedure. This excludes the difficulties of determination of these parameters, which are often difficult to achieve and automates the method. Moreover, different cases of damages were analyzed together, which allows for the determination of global parameters of the applied fractional B-spline wavelets and guarantee the best possible results for the damage detection and localization problems. Finally, the compromise between the time complexity of computations and the accuracy of the final solution could be reached in the automated manner.

References

[1] C.-C. Chang and L.W. Chen, "Damage Detection of a Rectangular Plate by Spatial Wavelet Based Approach," *Appl. Acoust.,* vol. 65, 2004, pp. 819-832, http://dx.doi.org/10.1016/j.apacoust.2004.01.004.

[2] E. Douka, S. Loutridis and A. Trochidis, "Crack Identification in Plates Using Wavelet Analysis," *J. Sound Vib.,* vol. 270, 2004, pp. 279-295, http://dx.doi.org/10.1016/S0022-460X(03)00536-4.

[3] Y. Huang, D. Meyer and S. Nemat-Nasser, "Damage Detection with Spatially Distributed 2D Continuous Wavelet Transform," *Mech. Mater.,* vol. 41, 2009, pp. 1096-1107, http://dx.doi.org/10.1016/j.mechmat.2009.05.006.

[4] W. Fan and P. Qiao, "A 2D Continuous Wavelet Transform of Mode Shape Data for Damage Detection of Plate Structures," *Int. J. Solids Struct.,* vol. 46, 2009, pp. 4379-4395, http://dx.doi.org/10.1016/j.ijsolstr.2009.08.022.

[5] S. Zhong and S.O. Oyadiji, "Crack Detection in Simply-Supported Beams Without Baseline Modal Parameters by Stationary Wavelet Transform," *Mech. Syst. Signal Process.,* vol. 21, 2007, pp. 1853-1884, http://dx.doi.org/10.1016/j.ymssp.2006.07.007.

[6] A. Katunin, "Modal-Based Non-Destructive Damage Assessment in Composite Structures Using Wavelet Analysis: A Review," *Int. J. Compos. Mater.,* vol. 3, 2013, pp. 1-9, http://dx.doi.org/10.5923/s.cmaterials.201310.01.

[7] A. Katunin and F. Holewik, "Crack Identification in Composite Elements with Non-Linear Geometry Using Spatial Wavelet Transform," *Arch. Civ. Mech. Eng.,* vol. 13, 2013, pp. 287-296, http://dx.doi.org/10.1016/j.acme.2013.02.003.

[8] L.H. Yam, Y.J. Jan and J.S. Jiang, "Vibration-Based Damage Detection for Composite Structures Using Wavelet Transform and Neural Network Identification," *Compos. Struct.,* vol. 60, 2003, pp. 403-412, http://dx.doi.org/10.1016/S0263-8223(03)00023-0.

[9] M. Rucka and K. Wilde, "Neuro-Wavelet Damage Detection Technique in Beam, Plate and Shell Structures with Experimental Validation," *J. Theor. Appl. Mech.,* vol. 48, 2010, pp. 579-604.

[10] H. Hein and J. Feklistova, "Computationally Efficient Delamination Detection in Composite Beams Using Haar Wavelets," *Mech. Syst. Signal Process.,* vol. 25, 2011, pp. 2257-2270, http://dx.doi.org/10.1016/j.ymssp.2011.02.003.

[11] M. Krawczuk, A. Żak and W. Ostachowicz, "Genetic Algorithms in Fatigue Crack Detection," *J. Theor. Appl. Mech.,* vol. 39, 2001, pp. 815-823.

[12] J. Xiang and M. Liang, "A Two-Step Approach to Multi-Damage Detection for Plate Structures," *Eng. Fract. Mech.,* vol. 91, 2012, pp. 73-86, http://dx.doi.org/10.1016/j.engfracmech.2012.04.028.

[13] J. Dumont, A. Hernández and G. Carrault, "Improving ECG Beats Delineation with an Evolutionary Optimization Process," *IEEE Trans. Bio-Med. Eng.,* vol. 57, 2010, pp. 607-615, http://dx.doi.org/10.1109/TBME.2008.2002157.

[14] A. Katunin, "Crack Identification in Composite Beam Using Causal B-spline Wavelets of Fractional Order," *Model. Eng.,* vol. 15, 2013, pp. 57-63.

[15] M. Unser and T. Blu, "Fractional Splines and Wavelets," *SIAM Rev.,* vol. 42, 2000, pp. 43-67, http://dx.doi.org/10.1137/S0036144598349435.

[16] K.N. Chaudhury and M. Unser, "Construction of Hilbert Transform Pairs of Wavelet Bases and Gabor-Like Transforms," *IEEE Trans. Signal Process.,* vol. 57, 2009, pp. 3411-3425, http://dx.doi.org/10.1109/TSP.2009.2020767.

[17] A. Katunin, "Damage Identification in Composite Plates Using Two-Dimensional B-spline Wavelets," *Mech. Syst. Signal Process.,* vol. 25, 2011, pp. 3153-3167, http://dx.doi.org/10.1016/j.ymssp.2011.05.015.

[18] A. Katunin and P. Przystałka, "Damage Assessment in Composite Plates Using Fractional Wavelet Transform of Modal Shapes with Optimized Selection of Spatial Wavelets," *Eng. Appl. Artif. Intell.,* vol. 30, 2014, pp. 73-85, http://dx.doi.org/10.1016/j.engappai.2014.01.003.

[19] R.E. Mercer and J.R. Sampson, "Adaptive Search Using a Reproductive Meta-Plan," *Int. J. Syst. Cybern.*, vol. 7, 1977, pp. 215-228, http://dx.doi.org/10.1108/eb005486.

[20] J.J. Grefenstette, "Optimization of Control Parameters for Genetic Algorithms," *IEEE Trans. Syst. Man Cybern.*, vol. 16, 1986, pp. 122-128, pp. 215-228, http://dx.doi.org/10.1109/TSMC.1986.289288.

[21] A.J. Keane, "Genetic Algorithm Optimization in Multi-Peak Problems: Studies in Convergence and Robustness," *Artificial Intelligence in Engineering*, vol. 9, 1995, pp. 75-83, http://dx.doi.org/10.1016/0954-1810(95)95751-Q.

[22] T. Back, "Parallel Optimization of Evolutionary Algorithms," *Proc. Int. Conf. on Evolutionary Computation*, " 1994, pp. 418-427, http://dx.doi.org/10.1007/3-540-58484-6_285.

[23] S.K. Smit and A.E. Eiben, "Comparing Parameter Tuning Methods for Evolutionary Algorithms," *Proc. IEEE Congress on Evolutionary Computation (CEC)*, 2009, pp. 399-406, http://dx.doi.org/10.1109/CEC.2009.4982974.

[24] M. Meissner, M. Schmuker and G. Schneider, "Optimized Particle Swarm Optimization (OPSO) and Its Application to Artificial Neural Network Training," *BMC Bioinformatics*, vol. 7, 2006, 125, http://dx.doi.org/10.1186/1471-2105-7-125.

[25] J. Branke and J.A. Elomari, "Meta-Optimization for Parameter Tuning with a Flexible Computing Budget," *Proc. 14th Annual Conf. on Genetic and Evolutionary Computation (GECCO12)*, Terence Soule (Ed.), New York, USA, 2012, pp. 1245-1252, http://dx.doi.org/10.1145/2330163.2330336.

[26] K. Miya, M. Uesaka and Y. Yoshida, "Applied Electromagnetics Research and Application," *Prog. Nucl. Energ.,* vol. 32, 1998, pp. 179-194, http://dx.doi.org/10.1016/S0149-1970(97)00015-2.

[27] R.T. Marler and J.S. Arora, "Survey of Multi-Objective Optimization Methods for Engineering," *Struct. Multidiscip. O.,* vol. 26, 2004, pp. 369-395, http://dx.doi.org/10.1007/s00158-003-0368-6.

[28] K. Deb, "Multi-Objective Optimization Using Evolutionary Algorithms," Wiley, 2009.

[29] S.M. Mikki and A.A. Kishk, "Particle Swarm Optimization: A Physics-Based Approach," Morgan and Claypool, 2008.

[30] K. Tang, X. Yao, P. N. Suganthan, C. MacNish, Y. P. Chen, C. M. Chen, and Z. Yang, "Benchmark Functions for the CEC'2008 Special Session and Competition on Large Scale Global Optimization," Technical Report, Nature Inspired Computation and Applications Laboratory, USTC, China, http://nical.ustc.edu.cn/cec08ss.php, 2007.

[31] M. Clerc and J. Kennedy, "The Particle Swarm - Explosion, Stability, and Convergence in a Multidimensional Complex Space," IEEE T. *Evolut. Comput.,* vol. 6, 2002, pp. 58-73, http://dx.doi.org/10.1109/4235.985692.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 439–446
DOI: 10.15439/2014F336
ACSIS, Vol. 2

Optimisation using Natural Language Processing: Personalized Tour Recommendation for Museums

Mayeul Mathias *, Assema Moussa *, Fen Zhou *‡, Juan-Manuel Torres-Moreno *‡§,
Marie-Sylvie Poli †‡, Didier Josselin *‡¶, Marc El-Bèze *‡, Andréa Carneiro Linhares ‖, Françoise Rigat **
{mayeul.mathias, assema.moussa}@alumni.univ-avignon.fr
{fen.zhou, juan-manuel.torres, marie-sylvie.poli, didier.josselin, marc.elbeze}@univ-avignon.fr
andrea.linhares@ufc.br, francoise_rigat@yahoo.it
*LIA, Université d'Avignon et des Pays de Vaucluse, France.
†CNE, Université d'Avignon et des Pays de Vaucluse, France.
‡FR 3621 Agorantic - CNRS Université d'Avignon et des Pays de Vaucluse, France.
§École polytechnique de Montréal, (Québec) Canada
¶UMR 7300 ESPACE - CNRS, France
‖Universidade Federal do Ceará, Brazil.
**Università degli Studi di Torino, Italy.

***Abstract*—This paper proposes a new method to provide personalized tour recommendation for museum visits. It combines an optimization of preference criteria of visitors with an automatic extraction of artwork importance from museum information based on Natural Language Processing using textual energy. This project includes researchers from computer and social sciences. Some results are obtained with numerical experiments. They show that our model clearly improves the satisfaction of the visitor who follows the proposed tour. This work foreshadows some interesting outcomes and applications about on-demand personalized visit of museums in a very near future.**

I. Introduction

MUSEUMS are no longer only institutions that acquire, store and expose our heritage. Going to a museum is a learning activity but also an enjoyment for visitors. With the emergence of the Web, curators and cultural mediators decided to get involved in collaborative and numerical culture to attract a larger public. Today, almost all museums have a website but few of them allow the visitors to prepare their visit in the best conditions.

Some art, science and society museums are collaborating with research laboratories to develop new technologies that improve services in museums in response to the desires of existing and potential visitors.

However, there are still difficulties, epistemological barriers, to study the expectations and the intentions of different publics, including online visitors. Knowing why people want to come and visit museums could allow automatic systems to suggests their tour, save their time and give them the best of the knowledge of the exhibited arts.

Among all possibilities, a recommendation system for personalized routing is by far one of the best improvements. Indeed, some museums exhibit thousands of artworks and it is not conceivable for a visitor to admire all of them because he might spend time in front of artworks which do not match his interests and he might not be able to see other more interesting artworks due to tiredness or a lack of time. A few museums, as The Louvre, offer a recommendation system[1] but they are limited to the selection of a route in a pre-established set. Moreover, in this particular case, the personalization is restricted to the selection of a theme and the duration of the visit in a set of no more than 10 themes and 4 different durations.

It is essential to propose a personalized route for each visitor or group of visitors according to their interests while taking into account their constraints such as limited schedule, physical handicap or a list of artworks to include of the tour. This operation may also reduce unuseful moves (avoid round trips). But to calculate an optimal tour, we need to assess the visitor interest for each artwork by asking his preferences.

Modeling the preferences with random distributions may not reflect reality because curators take care of the scenography (therefore the coherence) of each room. So we worked on prefered artists (the visitor can select a set of interesting artists) and we propose to use the artworks description to highlight a kind of intrinsic interest from the point of view of the museum. Indeed, the description displayed to the visitor should show how significant is the artwork for the museum. We valuate each item by analyzing their description (with Automatic Text Summarization) and use it as a base value, considering that even without any preference, some artworks are more interesting than others.

The Musée de l'Orangerie

Due to the time needed to extract and check all the data we worked on this small museum to test our model.

The Musée de l'Orangerie (Museum of Orangerie), in Paris (France), regroups 144 artworks from 14 artists in 10 exhibition rooms. The website[2] of the museum supplies a map (shown in Fig. 1) and indexes information about all the

[1]http://www.louvre.fr/en/parcours
[2]http://musee-orangerie.fr

978-83-60810-58-3/$25.00

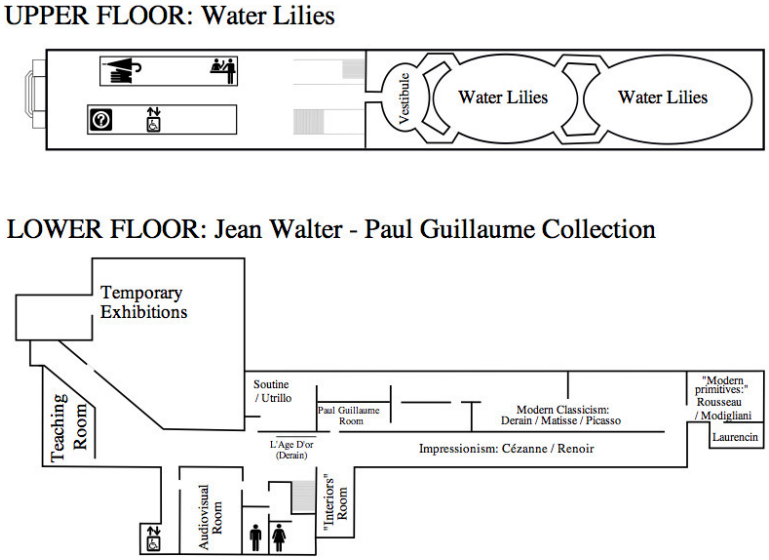


Fig. 1. Map of the Musée de l'Orangerie

artworks including the name of the artist, a description of the artwork and its date of creation.

Paper organization

The remaining of this paper is organized as follows. We review the related work in Section II. In Section III, we present a Natural Language Processing (NLP) based approach to compute artwork interest. The Personalized Tour Recommendation Problem is presented in Section IV. To solve this problem, we develop an Integer Linear Programming based method to solve the tour optimization problem in Section V and define a model to represent the visitor preferences in Section VI. The simulations are conducted and numerical results are demonstrated in Section VII. Finally, we conclude the paper in Section VIII.

II. Related work

A first model developed in 2010 [1] proposes to formulate the visitor routing problem as an extension of the open shop scheduling problem (in which each visitor group is a job and each interesting room is a machine). Each visitor group has to pass through all rooms but it is impossible for two groups of visitors to be simultaneously in the same room. This restriction can lead to non optimal or infeasible solutions if there are more visitor groups than rooms in the museum (which is the case if we consider each single visitor as a group).

Relying on the constraint programming model [2], we propose to reduce the number of used variables. In [2], they generate a route by calculating the smallest number K of steps required to cross the museum (to visit all the rooms). This model requires that each artwork is represented as K variables (one per step). Due to the fact that museums often have several thousands of artworks, it leads to a huge number of variables. Moreover they use mathematical distributions to simulate a visitor profile which does not necessary reflect reality (in museums, artworks are often grouped in a room because they are related to each other, a configuration that a random distribution as they used cannot represent).

In 2013, some works [3] used the visiting style of visitors (the way a visitor go from an artwork to another) but their model requires two matrices of size N^2 (where N is the number of artworks). The first one indicates the accessibility to an artwork from another (if they are in the same room or in two rooms directly connected) and the second one contains the distance between two artworks. However as the number of artworks is always greater than the number of rooms, most of the museums are modeled as two sparse matrices with duplicated data (in a room, it is often allowed to freely move between artworks). This makes the use of constraint programming expensive.

III. Artwork description analysis using Textual Energy

Our idea is to use the description of each artwork as an independent measure of their interest. Indeed, two similar artworks (same theme, support, artist) will produce the same result but may be very different. By analyzing the description provided by the museum, we tried to differentiate them.

Automatic Text Summarization (ATS) techniques by extraction [4], [5], [6] allow to rank a set of textual segments (sentences, paragraphs etc.) depending on a measure of similarity. Textual Energy algorithm (Enertex) converts a textual document into a physical object and use Statistical Physics to measure its energy [7]. This energy, to which we should refer as Textual Energy, is then computed and apply to summarization. The physical model of Textual Energy gives rise to a single non iterative algorithm of low complexity. Therefore Textual Energy allows to redefine sentence ranking on simple and efficient matrix operations. The resulting algorithms are

much easier to apply to large texts and give better results without using any post-processing.

A. Starting point: Hopfield Model

Hopfield's approach [8], [9] was based on magnetic Ising model to build a Neural Network (ANN) with pattern learning capabilities. The capacities and limitations of this ANN (an associative memory), were well established in a theoretical framework [8], [9]: the patterns must not be correlated to obtain free error recovery, the system saturates quickly and only a little fraction of the patterns can be stored correctly. Despite these major drawbacks, Hopfield contributed to ANN theory by introducing the concept of energy by analogy with magnetic systems. A magnetic system is a set of N spins like small magnets that can adopt several orientations. The simplest model is the dipole one or Ising model where there are only two opposite possible orientations: up ($\uparrow$ or +1) or down ($\downarrow$ or 0). Ising magnetic model was used in a large variety of systems that can be completely described by a set N of binary variables [10] with 2^N possible configurations (patterns). The spins are equivalent to neurons that can interact following Hebb's rule[3]:

$$J_{i,j} = \sum_{\mu=1}^{P} (s_{\mu,i} \times s_{\mu,j}) \tag{1}$$

$s_{\mu,i}$ and $s_{\mu,j}$ are the states of neurons i and j in the pattern μ. The summation concerns the P patterns to store. This rule of interaction is local, because $J_{i,j}$ depends only on the states of the connected unities. It has the capacity to store and to recover certain number of configurations of the system, because the Hebb rule transforms these configurations into attractors (minimal local) of the energy function [8]:

$$E_\mu = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (s_i \times J_{i,j} \times s_j) \tag{2}$$

The fundamental concept of magnetic energy is a function of the system configuration, that is, of the state of activation or non-activation of its units. The concept of energy induces a type of interaction. If we present a pattern ν, every spin will undergo a local field: $h^i = \sum_{j=1}^{N} J^{i,j} s^j$ induced by the energy of the others spins. Therefore the total energy of the new system made of the new pattern inserted into the previous system reflects the interaction between the pattern and the initial system.

We shall focus on theoretical objects that are usually considered in Statistical Physics. In magnetic system analysis, these are energy function distributions [11]. Hopfield himself used these functions to show that the recovery is convergent. Our Enertex system is entirely grounded on them.

B. Energy as a document similarity measure

The Vector Space Model (VSM) has also been applied to texts since [12] following a bag of word representation of sentences. In this model vectors represent sentences and a document gives rise to a matrix. We have used VSM to represent documents in our model magnetic system: a sentence (a row vector) is equivalent to a Ising spin chain and a document (a magnetic system) is represented by a matrix of P rows $\times$ N columns. Therefore, sentences can be studied as Ising spin chains. More formally, with a vocabulary of N words (terms) in a document, it is possible to represent a sentence as a chain of N spins, $i = 1, \cdots, N$. A document with P sentences is formed of P chains in the vector space Ξ of dimension N. In this paper, the description of each artwork is assimilated as a long pseudo-sentence. Therefore, a document (the collection of a museum) is constituted of a set of P (pseudo-)sentences.

Documents are preprocessed by removing functional words (by using a stop list), normalized and lemmatized [13], [14]. This preprocessing reduces considerably the document dimensionality. Let be $T = \{t_1, \ldots, t_N\}$ the set of remaining terms after this preprocessing. Once segmented into units, usually sentences, the text is represented by a set $S = \{\vec{s}_1, \ldots, \vec{s}_\mu, \ldots, \vec{s}_P\}$ where each $\vec{s}_\mu$ is the bag of words in segment μ. As usual in text vector model, we consider the matrix $S_{[P \times N]} = (s_{\mu,j})_{1 \leq \mu \leq P, 1 \leq j \leq N}$ of frequency/absence associated to H by:

$$s_{\mu,j} = \begin{cases} tf_{\mu,j} & \text{if } t_j \in \text{ sentence } \mu \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $tf_{\mu,j}$ is the term frequency of t_j in the sentence μ.

We therefore consider the presence of term t_j as a spin s_j $\uparrow$ with magnitude $tf_{\mu,j}$ (its absence by a $\downarrow$ respectively), and a description of each artwork (text segment) by a chain of N spins.

It is common to consider that these vectors are correlated according to the shared words. Here the introduction of the magnetic model induces moreover indirect interactions. In this model sentences that do not share any word could however interact because of the magnetic field generated by the other sentences of the document that form the global magnetic system.

We have studied the interactions between the terms and the sentences using Hebb's rule and Ising energy respectively. To obtain the matrix J of interactions between the N terms, we apply Hebb's rule (equation 1) in its matrix form:

$$J = S^T \times S \tag{4}$$

where $J_{i,j}$ is the number of co-occurrences of terms in sentences. The energy function (equation 2) of a (magnetic) system S is:

$$E = S \times J \times S^T = (S \times S^T)^2 \tag{5}$$

Each element $E_{\mu,\nu}$ represents the energy between sentences μ and ν. The values in the first matrix diagonal quantify the interaction energy between words into a sentence meanwhile the other values in the rest of the matrix show the interactions between distinct sentences. The sum of absolute values in one

[3]Hebb [9] suggested that synaptic connections change according to the correlation between neuronal states.

row gives the total energy of interaction of the corresponding sentence μ with the document:

$$E_\mu = \sum_\nu |e_{\mu,\nu}| \quad (6)$$

We use this energy value to rank sentences (description of artwork) by order of decreasing importance. The most energetic will be considered as the most important.

IV. Personalized Tour Recommendation Problem

The Personalized Tour Recommendation Problem (PTRP) can be viewed as an optimization problem and solved by optimization techniques. For this purpose, we first model the museum topology as a graph and then formulate the studied problem as an Integer Linear Programming (ILP) instance. Therefore, the optimal personalized tour can be obtained by solving the ILP model we propose.

A. Museum modeling

A museum is modeled as a 7-tuple $G = \langle V, A, E, X, P, r\rangle$ where:

- V is the set of vertices, each vertex is an exhibition room, an entrance or an exit of the museum.
- A is the set of arcs which connect different rooms. There is an arc $a_{ij} \in A$ between two vertices i and j, if we can go from room i to room j directly without passing through other rooms.
- E is the set of entrances of a museum, which is a subset of V, i.e. $E \subset V$.
- X is the set of exits of a museum, which is also a subset of V, i.e. $X \subset V$.
- P is the set of all artworks in the museum.
- r is a mapping function $P \to V$. For each artwork $p \in P$, $r(p)$ is the room containing p.

Some large museums may have several entrances and exits, that is why E and X are two subsets of V. We also admit that there is always a path from any entrance to an exit. We consider directed arcs and not edges because some museums may impose a flow direction for several reasons (minimizing congestion, pedagogical tour). Note that by definition of A, there is no incoming arc to any entrance and there is no any outgoing arc from any exit neither.

Application to the Musée de l'Orangerie: The Musée de l'Orangerie can be represented as the graph presented in Figure 2. We can see that there is only one entrance and one exit in the museum and they are located at the same place. Therefore, we consider the entrance and the exit as two different vertices in the graph to facilitate the model. The mapping between vertices and rooms is shown in Table I.

B. Personalized tour problem modeling

For the sake of satisfying the visitor maximally, a visit tour should be proposed according to the visitor's preferences and constraints.

A personalized tour problem can be defined as a 6-tuple $\langle G, I, R, u, t, T_{MAX}\rangle$ where:

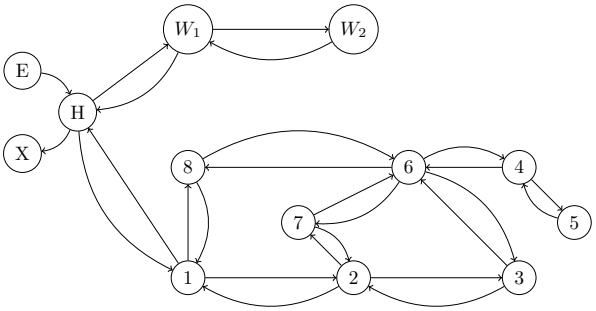


Fig. 2. A possible graph for the Musée de l'Orangerie

TABLE I
The Musée de l'Orangerie: vertices and rooms

Vertex	Room
E	Entrance
X	Exit
H	Hall
W_1	Water Lilies (first part)
W_2	Water Lilies (second part)
1	L'Age d'Or
2	Paul Guillaume Room
3	Impressionism
4	Modern primitives
5	Laurencin room
6	Modern classicism
7	Derain room
8	Soutine / Utrillo room

- G is the museum graph representing the museum topology as defined above.
- I is the set of artworks which have to be included in the tour.
- R is the set of artworks which have to be excluded of the tour.
- u is a mapping $u : P \to \mathbb{R}^+$. For each artwork $p \in P$, $u(p)$ denotes the interest of the visitor for the artwork p.
- t is a mapping $t : A \cup V \cup P \to \mathbb{R}^+$. For each room, arc and artwork, we have a time to spend. It can be the time needed to cross a room or an arc. It can also be the time to see an artwork.
- T_{MAX} is the maximum time that a visitor wants to spend in the museum.

A visit tour may be a simple path without any cycles (an elementary path) or a sophisticate path including cycles (a non-elementary path).

We define a tour as a sequence of pairs $\langle v, P_v\rangle$ where $v \in V$ and $P_v \subseteq r(v)$ (note that P_v may be $\emptyset$ because we can cross a room without seeing any artwork). A tour $T = (\langle v_1, P_{v_i}\rangle, ..., \langle v_n, P_{v_n}\rangle)$ is a solution to the personalized tour recommendation problem when:

1) The vertex v in the first element of T is an element of E (the tour starts by an entrance).
2) The vertex v in the last element of T is an element of X (the tour ends by an exit).
3) All consecutive elements T_1 and T_2 of T share the same vertex or an arc $a_{ij} \in A$ must exist from the vertex i of

T_1 to the vertex j of T_2.

4) The total time required to see all the artworks and pass through all the rooms (and ways) is not bigger than T_{MAX}.

Application to the Musée de l'Orangerie: we may have a visit tour like the following:

$$(\langle E, \emptyset\rangle, \langle H, \emptyset\rangle, \langle W_1, p_1\rangle, \langle W_2, p_5\rangle, \langle W_1, \emptyset\rangle, \langle H, \emptyset\rangle,$$
$$\langle 1, p_{10}\rangle, \langle 8, p_{103}\rangle, \langle 1, \emptyset\rangle, \langle H, \emptyset\rangle, \langle X, \emptyset\rangle)$$

In this tour, the visitor should cross the receiving hall H three times, exhibition room W_1 and 1 twice respectively. Although we may traverse a room several times, the visitor is supposed to visit the room only once. Consider for instance the exhibition room W_1, we may visit the selected artworks when we reach this room for the first time. The second time, we would just cross the room to visit another one.

V. Integer Linear Programming approach to solve the Optimal Personalized Tour Recommendation Problem

Before introducing an ILP model to solve the Personalized Tour Recommendation Problem, we define several decision variables:

- x_p equals 1 if the artwork $p \in P$ is included in the proposed tour, 0 otherwise.
- y_a equals 1 if arc $a \in A$ is crossed in the proposed tour, 0 otherwise.
- f_a denotes the number of rooms crossed when we arrive at arc $a \in A$ in the visit walk.
- z_v equals 1 if room $v \in V$ is traversed in the proposed tour (no matter whether we visit an artwork of this room or not), 0 otherwise.

Given a personalized tour problem (as defined in section IV), the objective function of the Optimal Personalized Tour Recommendation Problem (OPTRP) is to maximize the overall satisfaction of the proposed visit tour for the visitor:

$$\max \sum_{p\in P} x_p \times u(p) \qquad \text{OPTRP} \tag{7}$$

s.t.

$$\sum_{v\in E}\sum_{a\in\delta^+(v)} y_a = 1 \tag{8}$$

$$\sum_{v\in X}\sum_{a\in\delta^-(v)} y_a = 1 \tag{9}$$

$$\sum_{a\in\delta^+(v)} y_a = \sum_{a\in\delta^-(v)} y_a, \qquad \forall v \in V \setminus (E \cup X) \tag{10}$$

$$f_a \geq y_a, \qquad \forall a \in A \tag{11}$$

$$f_a \leq |V| \times y_a, \qquad \forall a \in A \tag{12}$$

$$\sum_{a\in\delta^-(v)} f_a = \sum_{a\in\delta^+(v)} f_a - z_v, \quad \forall v \in V \setminus (E \cup X) \tag{13}$$

$$\sum_{a\in\delta^+(v)} y_a + \sum_{a\in\delta^-(v)} y_a \geq z_v, \quad \forall v \in V \tag{14}$$

$$y_a \leq z_v, \qquad \forall v \in V, \forall a \in \delta^+(v) \cup \delta^-(v) \tag{15}$$

$$x_p \leq z_v, \qquad \forall v \in V, \forall p \in \{p : r(p) = v\} \tag{16}$$

$$x_p = 1, \qquad \forall p \in I \tag{17}$$

$$x_p = 0, \qquad \forall p \in R \tag{18}$$

$$\sum_{v\in V} z_v \times t_v + \sum_{a\in A} y_a \times t_a + \sum_{p\in P} x_p \times t_p \leq T_{MAX} \tag{19}$$

Constraints (8) and (9) ensure that the visitor should enter a museum from a unique entrance and finish the visit by a unique exit respectively (this model considers the case of multiple entrances and exits). Constraint (10) makes sure that a visitor should exit a room v after crossing or visiting it. Constraint (11) expresses that a visitor should have crossed at least a room before arriving at an arc a, while constraint (12) imposes that no flow is moving on the arc a, if it is not crossed in the visit tour. Constraint (13) means that if a room v is crossed in the tour, then the number of rooms crossed before arriving at this room equals the number of rooms crossed after leaving v minus one. Otherwise, they should be equal, since the room will not appear in the tour. Constraint (15) imposes that a room v should be crossed as long as one input arc or one outgoing arc is crossed. Constraint (14) ensures that a room v should not be crossed if none of the input arc or output link is used during the visit. Constraint (16) indicates that if a room v is not crossed, none of its artworks will be proposed for visiting. Constraints (17) and (18) ensure that an artwork should be included or excluded from the proposed tour if the visitor asks for it. The last constraint (19) guarantees that the time spent in front of the artworks and the time required to pass through rooms (and ways) does not exceed the available time for the visitor.

The ILP model we propose provides a visit tour starting from an entrance and terminating at an exit. In [2], authors also proposed an ILP model to plan the personalized visit. They divided the studied proposed into two sub-problems: first determine the number of moves (denoted as K) for a complete walk in the museum graph, and then solve the museum routing problem while maximizing visitor satisfaction. Since both of these sub-problems are NP-Hard, authors of [2] proposed to solve both of them by constraint programming. The complexity of their model depends a lot on K, which is generally large (at least equals to $|V|$). To compare the complexity of our model with the ILP mode in [2], the number of variables and constraints are listed in Table II.

VI. Visitor preferences modeling

The interest function u should reflect the satisfaction of the visitor for each artwork $p \in P$. The nearer to his preferences is an artwork p, the greater is $u(p)$.

Representation of artworks and the visitor preferences

We define C_p as the set of all characteristics of an artwork p and C as the union of all these sets (i.e. $C =$

TABLE II
COMPARISON OF ILP MODELS

Terms	OPTRP ILP	ILP [2]
Variables	x_p, y_a, f_a, z_v	x_p, $x_{p,k}$, $c_{i,j,k}$
Number of variables	$\|P\| + 2\|A\| + \|V\|$	$(\|P\| + \|A\|) \times K + \|P\|$
Constraints	(8)-(19)	(2)-(8) in [2]
Number of constraints	$3 + 4\|A\| + \|P\| + 3\|V\| + \|I\| - 2(\|E\| + \|X\|)$	$\|A\| \times (K-1) + \|P\| \times (2K-1) + K + 3$

$\{c|c \in C_p \forall p \in P\}$). A characteristic may be the theme, the type of support, the date of creation, the name of the artist or anything that identify an artwork.

We can represent any artwork $p \in P$ as a caracteristics vector $v_p = (v_{p_1}, ..., v_{p_n})$ in a vector-space of $|C|$ dimensions. Each element $v_{p_i} \in \mathbb{R}^+$ in the vector is a numerical value measuring the relevance of the artwork p to the associated characteristic. Additionally we define a vector v in the same vector-space as the vector representing the visitor preferences (where each element of v measures the interest of the visitor for the associated characteristic).

Measuring the interest for an artwork

To identify the interest for the visitor to an artwork, we compare v and v_p with the cosine similarity which calculate the angle between two vectors. The formula is the following:

$$similarity(v_p, v) = \frac{\sum_i^n v_{p_i} \times v_i}{\sqrt{\sum_i^n v_{p_i}^2} \times \sqrt{\sum_i^n v_i^2}} \qquad (20)$$

The resulting similarity ranges from 0, meaning that the visitor is not interested at all by the artwork, to 1 meaning that the artwork exactly matches his preferences.

In our model, we used $u(p) = similarity(v_p, v)$ where v_p and v are the vectors of the artwork p and the visitor preferences respectively.

VII. SIMULATIONS AND NUMERICAL RESULTS

We implemented the ILP model described in section V by using the IBM CPLEX 12 library[4].

The program takes several input parameters:

- The graph modeling the museum as defined in section IV
- The interest function f to use. This function produces interest vectors as defined in section VI
- The maximum duration that a visitor can spend in the museum

It outputs the proposed tour (as defined in section IV).

A. Intrinsic interest

As we saw in section III, the Enertex algorithm ranks the sentences of a document. We used Enertex as the following:

1) From the website of the museum, we created an XML file containing the following information for each artwork: the title, the artist name, the year and the description of the artwork.
2) We extracted data from the XML to produce a file where each pseudo sentence is a concatenation of title, artist and description.
3) The latter file is used as an input to Enertex with the query "musée orangerie peinture impressionniste postimpressionniste" to drive the balancing process of the system.

It produces a ranking for artworks depending on the information displayed by the museum (for each artwork, the result is a value ranging from 0 to 1).

Fig. 3 shows the ranking of the artworks in the *Musée de l'Orangerie* provided by Enertex.

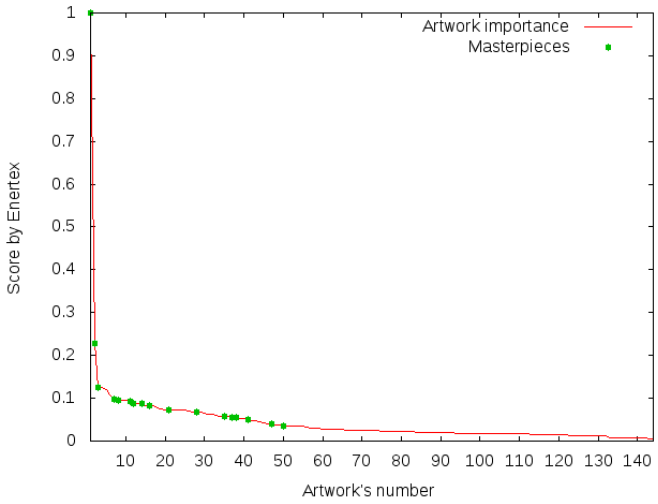


Fig. 3. Ranking by Enertex of the artworks in the Musée de l'Orangerie

As we can see, the resulting ranking is in agreement with the information provided by the museum. Indeed the masterpieces (according to the website of the museum) represent the most important artworks (which have the highest scores).

B. Interest functions

Four different interest function were designed to simulate the visitor preferences.

- f_1: produces the same vector $v = (1)$ for each artwork and visitor preferences
- f_2: produces a vector $v_p = (s_p)$ where s_p is the score given by Enertex for the artwork p and produces a vector $v = (1)$ as the visitor preferences.
- f_3: produces vectors $v = (v_1, ..., v_n)$ of size n equals to the number of artists. Each artwork is represented as a

[4]http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/

vector where $v_i = 1$ if the artwork is created by the artist i, 0 otherwise. The visitor preferences are represented as a vector where $v_i = 1$ if the visitor is interested by the artist i, 0 otherwise.

- f_4: produces vectors $v = v_{f_2} \| 10 \times v_{f_3}$ where v_{f_2} and v_{f_3} are the vectors produces by f_2 and f_3 respectively.

The first function defines the baseline: the visitor has no interest at all. The second uses the ranking provided by Enertex: the visitor wants to discover the most important artworks of the museum. The third uses the visitor preferences (a set of interesting artists). The last combines both visitor preferences and museum point of view (we multiply by 10 because we want to give more importance to the visitor preferences than to the museum point of view).

C. Evaluation

To evaluate the output tour, we measure the relevance percentage defined as :

$$rp = \frac{100 * \sum_{p \in T} f(p)}{\sum_{p \in P} f(p)} \tag{21}$$

where T is the set of artworks proposed in the tour (i.e. $T \subseteq P$) and f the interest function used (as saw above). The relevance percentage rp denotes a satisfaction rate of the visitor.

D. Results

For each function f, we ran the program with different time limits from 30 to 330 minutes (the time required to visit the entire collection) by steps of 15 minutes. For f_2 and f_3, we randomly pre-generated 5,000 combinations of 2, 3, 4 and 5 artists (i.e. the same combinations are used with f_2 and f_3) and calculated the arithmetic mean relevance percentage for each duration.

Figure 4 shows the evolution of rp for each function f. As expected, the first interest function f_1 produces a linear evolution of the relevance percentage (given that all artworks have the same interest, the tour includes the greatest number of artworks). With Enertex we are able to propose efficent tours to visitors who want to discover the museum (without particular preferences). The combination of both visitor preferences and intrinsic preferences produces the best results up to 49% of relevance improvement. It also appears that after 150 minutes, the improvement is less significant, we could assume that, from the visitor point of view, the optimal tour duration is about 2 hours and a half.

VIII. Conclusion and future works

This research tackles the problem of optimizing museum visits according to visitors preference and artwork importance. As a first milestone for next works taking into account the individual behavior in museum visits, it sets an original model combining computational optimisation and automatic learning via artificial intelligence. We first drew the optimization framework based on graph theory to depict the spatial organization of the museum (including rooms and paths), that requires an Integer Linear Programming to maximize the visitor overall satisfaction and to generate an optimal path, that is to say a series of rooms and artworks to be seen by the visitor. In complement, we compute an artwork description analysis by a natural language processing based on textual energy (using an algorithm called Enertex). This leads to ranking the different artworks according to the descriptions given by the museum, related to their artistic importance. Associating those two complementary approaches, we are then able to design optimal paths for visitors according to different interest functions based on artwork objective values assigned by museums.

Future works concern more subjective behavior of visitors depending on their profiles and leisure practices. Indeed, the project aims at finding relevant recommendations for optimal visit tours that rise a better fitness between the visitor wishes and the museum artistic supply. We can think about using natural language processing to generate the set of characteristics for all the artworks in a museum and calculate better interest vectors but also produce a summary of the proposed tour.

This information may advantageously be used by existing and potential visitors to refine the way they get involved in their cultural pratices. Indeed, it is admitted that the museum connoiseurs use to develop a critical mind about new services in a numerical society. Thence, aware visitors become able to appreciate the personalized routing recommendation system provided by their prefered museums.

Acknowledgments

The authors would like to thank the *Departement du Vaucluse* (France) and the *FR Agorantic* for the financial supports (projects *@MUSE* and *InfoMuse*).

References

[1] V. F. Yu, S.-W. Lin, and S.-Y. Chou, "The museum visitor routing problem," *Applied Mathematics and Computation*, vol. 216, no. 3, pp. 719–729, 2010.

[2] D. L. Berre, P. Marquis, and S. Roussel, "Planning personalised museum visits," in *ICAPS*, D. Borrajo, S. Kambhampati, A. Oddi, and S. Fratini, Eds. AAAI, 2013.

[3] I. Lykourentzou, X. Claude, Y. Naudet, E. Tobias, A. Antoniou, G. Lepouras, and C. Vassilakis, "Improving museum visitors' quality of experience through intelligent recommendations: A visiting style-based approach," in *Intelligent Environments (Workshops)*, ser. Ambient Intelligence and Smart Environments, J. A. Botía and D. Charitos, Eds., vol. 17. IOS Press, 2013, pp. 507–518.

[4] H. Luhn, "The Automatic Creation of Literature Abstracts," *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159–165, 1958.

[5] T. Sakai and K. Spärck-Jones, "Generic summaries for indexing in Information Retrieval," in *ACM Special Interest Group on Information Retrieval (SIGIR'01): 24th International Conference on Research and Development in Information Retrieval*. New Orleans, LA, USA: ACM, 2001, pp. 190–198.

[6] J.-M. Torres-Moreno, *Résumé automatique de documents : une approche statistique*. Hermes-Lavoisier (Paris), 2011.

[7] S. Fernández, E. SanJuan, and J. M. Torres-Moreno, "Textual Energy of Associative Memories: performants applications of ENERTEX algorithm in text summarization and topic segmentation," in *LNAI 4287, MICAI'07, Mexico*, 2007, pp. 861–871.

[8] J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences of the USA*, vol. 9, pp. 2554–2558, 1982.

[9] J. Hertz, A. Krogh, and G. Palmer, *Introduction to the theorie of Neural Computation*. Redwood City, CA: Addison Wesley, 1991.

[10] S. Ma, *Statistical Mechanics*. Philadelphia, CA: World Scientific, 1985.

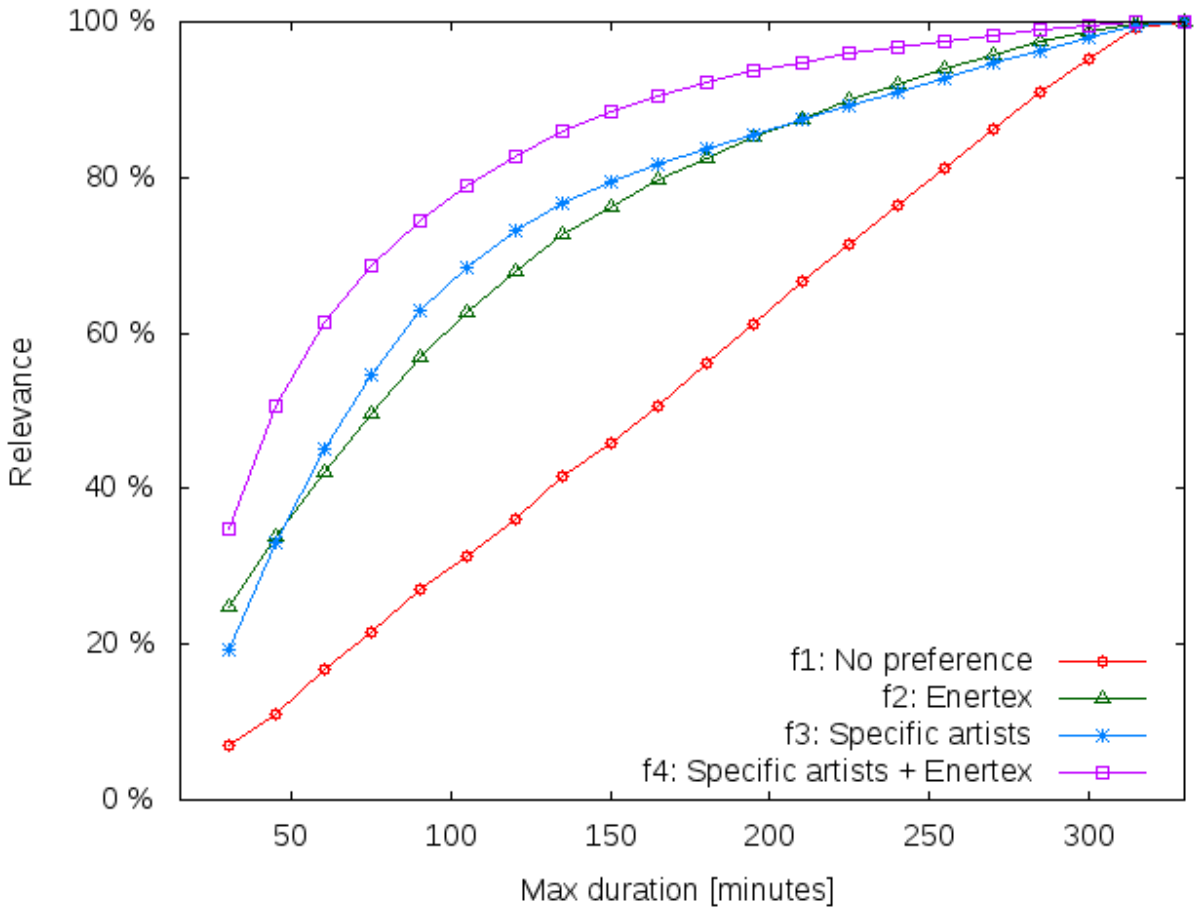


Fig. 4. Evolution of the relevance percentage

[11] A. Molina, J.-M. Torres-Moreno, E. SanJuan, G. Sierra, and J. Rojas-Mora, "Analysis and Transformation of Textual Energy Distribution," in *(MICAI), 2013 12th Mexican International Conference on Artificial Intelligence*. IEEE, 2013, p. 203–208.

[12] G. Salton and M. McGill, *Introduction to modern information retrieval*. Computer Science Series McGraw Hill Publishing Company, 1983.

[13] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, July 1980.

[14] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press, 1999. [Online]. Available: citeseer.ist.psu.edu/635422.html

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 447–456
DOI: 10.15439/2014F352
ACSIS, Vol. 2

Exploratory Equivalence in Graphs: Definition and Algorithms

Jurij Mihelič, Luka Fürst, and Uroš Čibej
University of Ljubljana, Faculty of Computer and Information Science
Tržaška cesta 25, SI-1000 Ljubljana, Slovenia
Email: {jurij.mihelic,luka.fuerst,uros.cibej}@fri.uni-lj.si

***Abstract*—Motivated by improving the efficiency of pattern matching on graphs, we define a new kind of equivalence on graph vertices. Since it can be used in various graph algorithms that explore graphs, we call it exploratory equivalence. The equivalence is based on graph automorphisms. Because many similar equivalences exist (some also based on automorphisms), we argue that this one is novel. For each graph, there are many possible exploratory equivalences, but for improving the efficiency of the exploration, some are better than others. To this end, we define a goal function that models the reduction of the search space in such algorithms. We describe two greedy algorithms for the underlying optimization problem. One is based directly on the definition using a straightforward greedy criterion, whereas the second one uses several practical speedups and a different greedy criterion. Finally, we demonstrate the huge impact of exploratory equivalence on a real application, i.e., graph grammar parsing.**

I. Introduction

GRAPHS are an ubiquitous format for structural-data representation and are gaining popularity in various scientific disciplines. They are used to represent diverse types of entities and relations between them in various areas, ranging from chemistry [1], [2], economy [3], politics [4], to popular culture [5]. Such representation enables a more general and global view on the data. Additionally, researchers may benefit from powerful theoretical tools developed in graph theory to extract new insights.

One of the most general problems on various graphs is search for patterns, i.e., finding occurrences of small graphs in larger graphs. In theory, this is known as the subgraph isomorphism problem and has been thoroughly studied, as this is one of the fundamental problems in theoretical computer science. The decision version of this problem is *NP*-complete, and the counting version of the problem is $\#P$-complete. Furthermore, no exponential-time algorithm with a lower bound better than the naive enumeration of pattern is known [6]. This makes the problem intrinsically hard. Despite these pessimistic results, various algorithms exist for finding patterns, a vast majority of them based on the branch-and-bound method (e.g., [7], [8]). In many practical instances, however, these algorithms perform much better than the expected worst-case scenario and are able to solve relatively large instances (e.g., patterns of 1000 vertices in graphs of 10,000 vertices, and even larger).

Despite the practical usability of the current algorithms, there is a large set of problem instances that are often very hard for all the search algorithms. These are graphs with a lot of symmetries, i.e., graphs with many automorphisms. Detecting these symmetries before the start of the search can speed up the algorithm by very large constants, since the search does not have to be repeated for the symmetrical vertices. The goal of this paper is to formally define an equivalence on graph vertices, called *exploratory equivalence*, that captures such symmetries in graphs and can be easily utilized in algorithms for finding patterns (e.g., subgraph isomorphism) in graphs. Since there can be many exploratory equivalences in a graph (and some capture more symmetries than others), we also define the corresponding optimization problem. Our work is based on the ideas already developed by Fürst *et al.* [9] for the purpose of improving the Rekers-Schürr parser [10] for context-sensitive graph grammars. However, while Fürst *et al.* recognized the concept of exploratory equivalence (under the name 'interchangeability'), they did not treat it in a general graph-theoretic and group-theoretic manner. Besides that, they did not consider the possibility of having multiple exploratory equivalences for a single graph, nor did they define the notion of optimal exploratory equivalence. In this paper, we address all of these issues.

Informally, if a group of k vertices in an unlabeled graph belong to the same exploratory equivalence class, then they are interchangeable in the following sense: if each of them were labeled with a unique label, their labels could be arbitrarily interchanged with each other without affecting the graph. The graph would remain isomorphic after any of the $k!$ possible interchanges. It is important to note that a single graph may have multiple exploratory equivalences, i.e., multiple ways of partitioning the graph vertex set into a set of exploratory equivalent classes. Among all possible exploratory equivalences for a given graph, the algorithms proposed in this paper seek the one that captures the largest number of symmetries. As we show later, this is the equivalence with the largest product of the factorials of the cardinalities of its equivalence classes.

Graph grammars [11] are production-based graph rewrite systems and are regarded as a generalization of well-known string-based formal grammars. The Rekers-Schürr parser is an algorithm that, for a given graph and a context-sensitive graph grammar, determines whether the graph belongs to the language generated by the grammar and returns a derivation of the graph in the grammar if this is the case. However, the algorithm may exhibit a heavily exponential behavior when presented with a grammar containing many symmetries. In particular, given a simple grammar for chemical formulas of linear alkanes, the algorithm failed to parse the structural formula of propane within several hours. By exploiting the symmetries in the grammar, the parser's performance is brought down to polynomial for several meaningful classes

978-83-60810-58-3/$25.00

of grammars [9]. For instance, the parsing of propane now takes less than a second. In general, however, the worst-case performance remains exponential, since the graph grammar parsing problem is *NP*-hard even for highly restricted graph grammar formalisms [12].

Symmetry reduction techniques are not unique to graph-related decision and optimization problems. Liberti [13], for instance, proposed a novel approach to symmetry reduction in branch-and-bound-based MIP (mixed integer programming) solvers. His approach was applied to the discretizable molecular distance problem in the field of organic chemistry [14].

The paper is structured as follows. In the next section, we briefly present definitions and notions used in the rest of the paper. The third section includes the definition of exploratory equivalence, the optimization problem of finding the best exploratory equivalence in a given graph, and an example demonstrating the introduced concepts. We also present the argument that exploratory equivalence does not belong to the class of well-known regular equivalences. The fourth section presents two heuristic algorithms for solving the optimization problem. In Section V, we briefly describe the relevant portion of the Rekers-Schürr parser, its improvement with regard to exploratory equivalence, and some experimental results. Finally, Section VI concludes the paper and gives some ideas for the future work.

II. Preliminaries

Given a (finite) set S, a family $\{P_1, P_2, \ldots, P_s\}$ of nonempty subsets of S is a *partition* of S if every element in S is exactly in one of the subsets, i.e., $P_i \subseteq S$ and $P_i \neq \emptyset$, where $1 \leq i \leq s$, $\bigcup_{1 \leq i \leq s} P_i = S$, and $P_i \cap P_j = \emptyset$ for all $1 \leq i, j \leq s$ where $i \neq j$. When the partition $\{P_1, P_2, \ldots, P_s\}$ is given explicitly, we usually use $\{i \in P_1 \mid i \in P_2 \mid \ldots \mid i \in P_s\}$ as a short form, e.g., $\{\{1,2\},\{3\},\{4\}\}$ is shortened to $\{1,2 \mid 3 \mid 4\}$. In what follows, the order of the sets in a partition is often important. In such cases, we use the form $\langle i \in P_1 \mid i \in P_2 \mid \ldots \mid i \in P_s \rangle$, e.g., $\langle 1,2 \mid 3 \mid 4 \rangle$.

A *group* $\Gamma = (A, \circ)$ with the underlying set A and the binary operation $\circ$ on the elements of A is an algebraic structure satisfying the following conditions: *closure*, i.e., $x \circ y \in A$, *associativity*, i.e., $(x \circ y) \circ z = x \circ (y \circ z)$, *identity element* e, i.e., $\exists e \in A \forall x \in A : e \circ x = x \circ e = x$, and *inverse element*, i.e., $\forall x \in A \exists x^{-1} \in A : x \circ x^{-1} = x^{-1} \circ x = e$.

A *permutation* σ is a bijective function of a finite set S onto itself, i.e., $\sigma \colon S \to S$. Let $\Pi[S]$ denote the set of all permutations of the elements in the set S. Notice that the set $\Pi[S]$ together with the operation of function composition forms a group, which is called the *symmetric group*. Since all the groups discussed in this paper are subgroups of a symmetric group, we write as a group its underlying set only. Additionally, we also define $\Pi[n] = \Pi[\{1, 2, \ldots, n\}]$.

Let Γ be a subgroup of $\Pi[S]$. An element $i \in S$ is called a *fixed point* of the permutation $\sigma \in \Gamma$ if $\sigma(i) = i$. The set of all permutations for which i is a fixed point is a subgroup and is called the *stabilizer subgroup*, i.e.,

$$\mathrm{Stab}_\Gamma(i) = \{\sigma \in \Gamma \mid \sigma(i) = i\}.$$

Notice that all stabilizer subgroups include the identity permutation.

Now let us generalize the definition of a stabilizer from an element to a set. Given $P \subseteq S$, a stabilizer on P is a set of permutations which have a fixed point for all the positions in P:

$$\mathrm{Stab}_\Gamma(P) = \{\sigma \in \Gamma \mid \forall i \in P : \sigma(i) = i\}.$$

Equivalently, $\mathrm{Stab}_\Gamma(P)$ can also be defined in terms of intersections of $\mathrm{Stab}_\Gamma(i)$, where $i \in P$, i.e.,

$$\mathrm{Stab}_\Gamma(P) = \bigcap_{i \in P} \mathrm{Stab}_\Gamma(i).$$

From the latter definition it is clear that $\mathrm{Stab}_\Gamma(P)$ also satisfies all four group conditions. We thus have the following theorem.

Theorem 1: Given a set S, a set $P \subseteq S$, and a subgroup Γ of the group $\Pi[S]$, $\mathrm{Stab}_\Gamma(P)$ is a subgroup of Γ.

We also write $\mathrm{Stab}_\Gamma(P)$ as $\mathrm{Stab}(\Gamma, P)$.

The set of all images of $i \in S$ under permutations of the group Γ is called the *group orbit* of i, i.e.,

$$\mathrm{Orbit}_\Gamma(i) = \{\sigma(i) \mid \sigma \in \Gamma\}.$$

Let $G = (V, E)$ denote a simple undirected graph, where $V = \{1, 2, \ldots, n\}$ is a set of vertices and $E \subseteq V \times V$ is a set of edges. When two graphs are considered, the second is usually denoted with $H = (U, F)$. To denote an edge $(i, j) \in E$, we usually use a shorter version $ij \in E$. A *neighborhood* of a vertex $i \in V$, i.e., a set of vertices adjacent to i, is denoted with $\mathcal{N}(i)$. More formally,

$$\mathcal{N}(i) = \{j \in V \mid ij \in E\}.$$

A *coloration* C of a graph G is an assignment of colors to the vertices V of G, i.e., a surjective function C from V onto $\{1, 2, \ldots, c\}$ for some c, where colors are denoted with integers from 1 to c. Any coloration defines a partition of the vertices V, and vice versa. If $S \subseteq V$, then the *spectrum* of S, denoted $C(S)$, is a set of all colors assigned to the vertices of S. If $S = \{i\}$ is a singleton, then $C(i) = C(S)$ denotes the color assigned to the vertex $i \in V$. A coloration C induces a graph partition $\{C^{-1}(1), C^{-1}(2), \ldots, C^{-1}(c)\}$, and vice versa. A coloration C_1 is *finer or equal* than a coloration C_2 (denoted $C_1 \preceq C_2$) if

$$\forall i, j \in V \colon C_2(i) < C_2(j) \implies C_1(i) < C_1(j).$$

This implies that each set of the C_1-induced partition is a subset of (or equal to) some set of the C_2-induced partition.

A graph *homomorphism* from a graph $G = (V, E)$ to a graph $H = (U, F)$ is a mapping $f \colon V \to U$ such that for each $ij \in E$ it also holds that $f(i)f(j) \in F$. Homomorphism $f \colon V \to U$ is usually denoted with $f \colon G \to H$. We also write $G \to H$ if there exists a homomorphism from G to H. A graph *isomorphism* is a bijective homomorphism, i.e., a mapping $f \colon G \to H$ such that $ij \in E$ if and only if $f(i)f(j) \in F$. We write $G \simeq H$ if there exists an isomorphism from G to H; such graphs G and H are called isomorphic. Since isomorphisms are bijective, every isomorphism also has an inverse. A graph *endomorphism* is a homomorphism whose domain is equal to its codomain, i.e., $f \colon G \to G$.

A graph *automorphism* is both an endomorphism and an isomorphism, i.e., a mapping $f \colon G \to G$ such that $ij \in E$ if

and only if $f(i)f(j) \in E$. Notice that every automorphism is a permutation. If identity is the only automorphism of a graph, we say that the graph is *rigid*. The set of all automorphisms of a graph G is denoted with

$$\mathrm{Aut}(G) = \{a \in \Pi[n] \mid G \simeq a(G)\}$$

and is called the *automorphism group* of a graph G. Constructing $\mathrm{Aut}(G)$ is at least as difficult as solving the graph isomorphism problem, since graphs G and H are isomorphic if and only if the disconnected graph formed by the disjoint union of G and H has an automorphism that swaps the two components. Several practical algorithms are known for finding $\mathrm{Aut}(G)$; the most well-known is probably NAUTY [15].

III. PROBLEM DESCRIPTION

As already mentioned in the introduction, our goal is to find equivalent (also called indistinguishable) vertices of a graph. There are many types of equivalences already discussed in the literature. We give several examples later in this section. Our definition of equivalence is associated with the algorithmic exploration of a graph; for example, when the task is to find a pattern graph that is a subgraph in another target graph. In particular, branch-and-bound search algorithms could exploit such equivalences by reducing the number of (partial) matches established between a set of equivalent vertices in the pattern graph and a corresponding set of vertices in the target graph. In the remainder of this section, we formally describe our type of equivalence and the problem of finding the corresponding equivalence classes. Additionally, we also discuss several other similar equivalences and argue that our type is novel.

First, let us define a few additional notions. Let S be a set, and let $P \subseteq S$ be a set of positions. We say that a permutation $\sigma_1 \in \Pi[P]$ is *covered* by a permutation $\sigma_2 \in \Pi[S]$ if the two permutations have the same image on the positions P, i.e.,

$$\sigma_1 \preceq \sigma_2 \equiv \forall i \in P\colon \sigma_1(i) = \sigma_2(i).$$

Observe that P is equal to the domain of σ_1.

Now let $A \subseteq \Pi[S]$. We say that a set A of permutations *covers* a set P of positions if every permutation of P is covered by a permutation in A. More formally,

$$\mathrm{cover}(A, P) \equiv \forall \sigma \in \Pi[P]\ \exists a \in A : \sigma \preceq a.$$

Given a graph $G = (V, E)$, we say that a partition $\{P_1, P_2, \ldots, P_s\}$ of V is *exploratory equivalent* if for all $1 \leq i \leq s$ the following two conditions hold:

$$\mathrm{cover}(A_{i-1}, P_i) \text{ and } A_i = \mathrm{Stab}(A_{i-1}, P_i), \tag{1}$$

where $A_0 = \mathrm{Aut}(G)$. The sets $P_1, P_2, \ldots, P_s$ are the equivalence classes. Notice that the order of classes regarding the partition $\{P_1, P_2, \ldots, P_s\}$ is irrelevant, but it is important when checking the conditions (1), since not all orders of $P_1, P_2, \ldots, P_s$ satisfy them. In this sense the exploratory equivalence is an algorithmic concept. In particular, an algorithm processing a vertex $u \in P_i$ may ignore all other vertices in P_i, since the automorphisms A_{i-1} cover all permutations of P_i. However, it is important to observe that equivalence classes are not independent. For example, when a vertex $u \in P_i$ is processed, this may influence the rest of the algorithm. Therefore, when determining the next class P_{i+1}, one must exclude the automorphisms corresponding to the already processed classes $P_1, P_2, \ldots, P_i$, which is the same as restricting to the automorphisms where the positions $P_1 \cup P_2 \cup \cdots \cup P_i$ are fixed points. That is the reason why in each step the automorphism group is restricted from A_{i-1} to $A_i = \mathrm{Stab}(A_{i-1}, P_i)$.

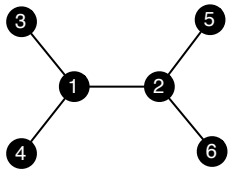


Fig. 1. An example graph with several exploratory equivalences.

Let us demonstrate the introduced concepts with an example. Consider the 6-vertex graph of Fig. 1. Its automorphism group consists of the following eight permutations (written in the one-line notation):

$$123456, 123465, 124356, 124365, \\ 215634, 215643, 216534, 216543. \tag{2}$$

There are twelve exploratory equivalent partitions of the graph. They are given in the form of a Hasse diagram (using the refinement relation $\preceq$ between two partitions) in Fig. 2. The

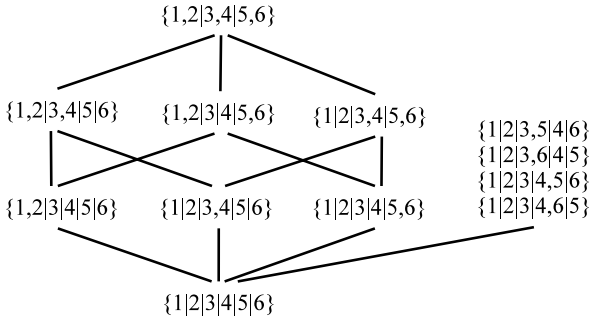


Fig. 2. Hasse diagram of all the exploratory equivalent partitions of the graph from Fig. 1. (The four partitions on the right-hand side are actually four separate vertices in the diagram.)

trivial partition ($\{1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6\}$ in the case of the graph of Fig. 1) is always exploratory equivalent. For the trivial partition, any ordering of its constituent sets satisfies the conditions (1). By contrast, for the exploratory equivalent partition $\{1, 2 \mid 3, 4 \mid 5, 6\}$, only the orderings $\langle 3, 4 \mid 5, 6 \mid 1, 2\rangle$ and $\langle 5, 6 \mid 3, 4 \mid 1, 2\rangle$ satisfy those conditions.

Corollary 1: Given a graph and its partition $\{P_1, P_2, \ldots, P_s\}$, let $A_0 = \mathrm{Aut}(G)$ and $A_i = \mathrm{Stab}(A_{i-1}, P_i)$ for $1 \leq i \leq s$. Then each A_i, where $1 \leq i \leq s$, is a subgroup.

Proof: $\mathrm{Aut}(G)$ is a group. By repeatedly applying Theorem 1, we know that A_i is a subgroup of A_{i-1}, for all $1 \leq i \leq s$. ∎

Now we are ready to define the problem. The input of the problem is a graph $G = (V, E)$ and its automorphism group $\mathrm{Aut}(G)$, and the goal of the problem is to find an exploratory equivalent partition $\{P_1, P_2, \ldots, P_s\}$ of V that maximizes the

product

$$\prod_{i=1}^{s} |P_i|!.$$

The reason for using the product of factorials in the objective function is that each class P_i covers $|P_i|!$ automorphic graphs, and the total number of automorphic graphs covered is thus the product above. In the following sections, we denote the problem with MAXEXPLOREQ.

In the paper [16], a large class of the so-called regular equivalences (called colorations therein) is surveyed. A coloration of a graph is *regular* when the equality of the spectra of two vertices implies the equality of the spectra of the corresponding neighborhoods. More formally, a coloration C of graph G is *regular* if and only if for all $i, j \in V$

$$C(i) = C(j) \implies C(\mathcal{N}(i)) = C(\mathcal{N}(j)).$$

Many different types of colorations are regular, e.g., strong and weak structural coloration, orbit coloration, perfect coloration, and exact coloration. See [16] for details. For example, coloring each orbit of $\mathrm{Aut}(G)$ gives orbit coloration. However, as it turns out, exploratory equivalence is not regular. To demonstrate this, consider again the graph from Fig. 1 and its exploratory equivalent partition $\{1, 2 \mid 3, 4 \mid 5, 6\}$, where the color of each class is different. It is easy to see that it is not regular, since $C(1) = C(2)$ but $C(\{3, 4\})$ is not equal to $C(\{6, 7\})$.

IV. ALGORITHM DESCRIPTION

In this section, we will describe two greedy algorithms for the MAXEXPLOREQ problem. The first algorithm is based on restricting the set of automorphisms to the stabilizer of the equivalent vertices found in one iteration. The second algorithm is more time-efficient owing to a faster detection of equivalent sets.

A. Greedy algorithm based on stabilizer restrictions

The first algorithm for the optimization problem MAXEXPLOREQ is based on the definition and will represent a reference algorithm that can be further improved. The idea of the algorithm is to start with the initial automorphism group, find one equivalence class of the partition, reduce the set of automorphisms only to the stabilizer of A, and recursively find new equivalence classes until the entire set of vertices is contained in the equivalence.

The input to this problem is the set of automorphisms (permutations) A and a set $V' \subseteq V$ of vertices not yet included in any equivalence class; initially V' is the entire set V.

If the set of automorphisms contains only the identity, then each vertex in V' represents a different equivalence class (i.e., no new indistinguishable vertices exist in the graph). If there is more than one automorphism in A, then at least two vertices are indistinguishable. At this point, the goal of the algorithm is to find a subset $S \subseteq V'$ that is covered by A. Usually, however, there are many possibilities for S, and different choices can lead to very different final solutions. The greedy criterion for this choice is the size of S, i.e., among many possibilities, the largest set S is chosen. When there are more sets with the same size, the algorithm chooses the one that has the largest stabilizer in A. The described algorithm is shown in more detail as Algorithm 1.

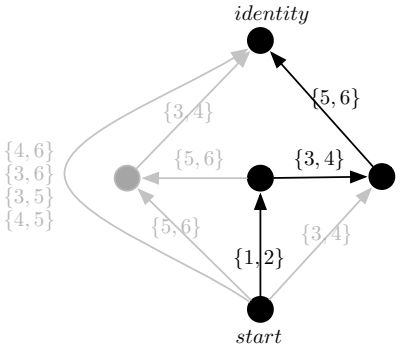


Fig. 3. The search space of Algorithm 1 for the graph in Fig. 1

Algorithm 1 Greedy algorithm for MAXEXPLOREQ based on stabilizer restrictions.

```
1:  function GREEDY1(A, V')
2:      if |A| = 1 then return singletons(V')
3:      bestP = ∅
4:      bestA = ∅
5:      for all P: P ⊆ V' ∧ cover(A, P) do
6:          A' ← Stab(A, P)
7:          if |P| > |bestP| ∨
8:             |P| = |bestP| ∧ |A'| > |bestA| then
9:              bestP ← P
10:             bestA ← A'
11:     return {bestP} ∪ GREEDY1(bestA, V' \ bestP)
```

To make this algorithm a little more clear, we will show its trace on the simple example graph of Fig. 1. The initial set of all automorphisms A is already shown in equation (2). From this set, the algorithm finds the equivalence class $\{1, 2\}$ and reduces A to the set $\mathrm{Stab}(A, \{1, 2\})$, which is:

$$A' = \{123456, 123465, 124356, 124365\}$$

In this automorphism group, it finds the equivalence class $\{3, 4\}$ and reduces the automorphisms to the stabilizer:

$$A' = \{123456, 123465\}$$

The final equivalence class from this group is $\{5, 6\}$, and the corresponding stabilizer contains only the identity. This yields the final result, namely the partition $\{1, 2 \mid 3, 4 \mid 5, 6\}$. If, at the moment when A' contained only the identity, the current partition did not include all six vertices of the graph, each of the missing vertices would be added as a singleton set to the equivalence. The entire search space for this example is shown in Fig. 3. Each vertex in this graph represents an automorphism group. The bottom vertex is the set of all automorphisms, and the top vertex is the set containing only the identity. Each edge represents a stabilization with the set that is written as the label of the edge. The bold vertices and edges are the ones that our algorithm follows.

Now we will discuss the correctness of the described algorithm.

Fig. 4. Two graphs on which Algorithm 1 returns a suboptimal solution. The left graph is the smallest counterexample in terms of the number of vertices, and the right one is the smallest counterexample in terms of the number of edges.

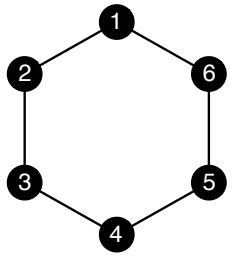


Fig. 5. A sample graph

Theorem 2: Algorithm 1 returns a partition of exploratory equivalent vertices.

Proof: Since the algorithm closely follows the definition, the proof is trivial. Each partition is covered by the automorphism group; the loop only iterates over the subsets that are covered. The second criterion from the definition is guaranteed by the recursion, since the set of automorphisms used in the recursion is only the stabilizer of the equivalence class found in the previous step. ■

Another question we need to address is the optimality of this algorithm. Unfortunately, the greedy criterion does not guarantee the optimality of the solution. We will demonstrate this by two examples shown in Fig. 4. These two examples were found by the exhaustive enumeration of all non-isomorphic connected graphs (starting with the smallest graph), and the graphs of Fig. 4 are the smallest examples where Algorithm 1 does not find an optimal solution. The optimal solution for the left graph in Fig. 4 is one with value 8 (partition $\{1,4 \mid 2,5 \mid 3,6\}$), whereas the algorithm returns a solution with value 6 (partition $\{1,3,5 \mid 2 \mid 4 \mid 6\}$). A similar situation occurs with the right graph, where the optimal solution is 8 (partition $\{1 \mid 2,3 \mid 4,5 \mid 6,7\}$), but the algorithm returns a suboptimal solution with value 6 (partition $\{2,4,6 \mid 1 \mid 3 \mid 5 \mid 7\}$).

Because of the exhaustive search over all subsets of V', the described algorithm is not very practical for larger graphs. In the next subsection, we will describe a more efficient algorithm that utilizes an incremental procedure to find the possible equivalence classes.

B. Greedy algorithm based on positional restriction of automorphisms

For a more convenient presentation of our second greedy algorithm, let us define a few auxiliary terms. The *positional restriction* of an automorphism (permutation) $a \in \Pi[S]$ to a set $R \subseteq S$ (denoted $\rho(a,R)$) is a partial function $a' \colon S \to S$ with $a'(i) = a(i)$ for all $i \in R$ and $a'(i)$ being undefined for all $i \in S \setminus R$. For example, $\rho((3,2,1,4),\{2,4\}) = (\uparrow,2,\uparrow,4)$. We use the one-line notation for representing automorphisms ($(1,2,3,4) \equiv 1234$) and the symbol $\uparrow$ for indicating the undefined values. Therefore, $a = (\uparrow,2,\uparrow,4)$ represents the fact that both $a(1)$ and $a(3)$ are undefined, whereas $a(2) = 2$ and $a(4) = 4$.

The positional restriction of a set of automorphisms $A \subseteq \Pi[S]$ to a set $R \subseteq S$ (denoted $\rho(A,R)$) is a set $\{\rho(a,R) \mid a \in A\}$. For example, $\rho(\{(1,2,3,4),(3,2,1,4)\}, \{2,4\}) = \{(\uparrow,2,\uparrow,4)\}$. As illustrated by this example, several automorphisms may collapse into one as a side-effect of a positional restriction.

For a given set S and a given set of (positionally unrestricted or restricted) set of automorphisms $A \subseteq \Pi[S]$, a *permofix* is a pair (P,F) such that the following conditions hold: (1) $P \subseteq S$, (2) $F \subseteq S$, (3) $P \cap F = \emptyset$, and (4) for each permutation $\sigma \in \Pi[P]$ there exists an automorphism $a \in A$ such that $a(i) = \sigma(i)$ for all $i \in P$ and $a(i) = i$ for all $i \in F$. In other words, a pair (P,F) is a permofix if there exists a set of automorphisms $A' \subseteq A$ that covers the set P (i.e., all permutations of P) and simultaneously fixes all elements of F. Given a permofix (P,F), the sets P and F will be called the *perm-set* and the *fix-set*, respectively. A k-permofix is a permofix (P,F) with $|P| = k$. The *potential* of a permofix (P,F) is the product $|P|!\,|F|!$. A permofix (P',F') is *contained* in a permofix (P,F) (denoted $(P',F') \sqsubseteq (P,F)$) if $P' \subset P$ (a *strict* subset) or $P' = P$ and $F' \subseteq F$.

As an example, consider the graph of Fig. 5. The 12 automorphisms of this graph are as follows:

$$\begin{aligned} a_1 &= (1,2,3,4,5,6) \\ a_2 &= (2,3,4,5,6,1) \\ a_3 &= (3,4,5,6,1,2) \\ a_4 &= (4,5,6,1,2,3) \\ a_5 &= (5,6,1,2,3,4) \\ a_6 &= (6,1,2,3,4,5) \\ a_7 &= (1,6,5,4,3,2) \\ a_8 &= (2,1,6,5,4,3) \\ a_9 &= (3,2,1,6,5,4) \\ a_{10} &= (4,3,2,1,6,5) \\ a_{11} &= (5,4,3,2,1,6) \\ a_{12} &= (6,5,4,3,2,1) \end{aligned} \tag{3}$$

For this graph, the pair $(\{1,3\},\{2,5\})$ is a permofix, since the automorphisms a_1 and a_9 cover both permutations of the set $\{1,3\}$ while fixing the elements 2 and 5. The pair $(\{1,3,5\},\emptyset)$ is a permofix as well, since the automorphisms a_1, a_3, a_5, a_7, a_9, and a_{11} collectively cover all permutations of the set $\{1,3,5\}$. We also have $(\{1,3\},\{2,5\}) \sqsubseteq (\{1,3,5\},\emptyset)$.

Given the set of automorphisms $A \subseteq \Pi[n]$ of a n-vertex graph, the algorithm works as a greedy iterative process. In each iteration, it produces the set of all permofixes in the current set of automorphisms (in the first iteration, this is the unrestricted set A) and greedily selects a permofix with the

largest potential. After making its selection, the algorithm positionally restricts all automorphisms to the fix-set of the selected permofix. The positionally restricted set of automorphisms serves as the input to the next iteration. The process stops once all automorphisms have become completely undefined functions. The output of the algorithm is a set composed of all perm-sets of the permofixes selected in individual iterations and of the singletons containing the individual vertices that are not present in any of the selected perm-sets. Later, we shall show that the algorithm's output is an exploratory equivalent partition of the vertex set.

The rationale for selecting a permofix with the highest potential is based on the following heuristics: Recall that the algorithm's goal is to find a partition $\{P_1, \ldots, P_s\}$ of $\{1, \ldots, n\}$ with a maximum value of $|P_1|! \ldots |P_s|!$. A permofix (P, F) is guaranteed to contribute at least a factor of $|P|!$ to the target product $|P_1|! \ldots |P_s|!$ (since the perm-set of the selected permofix is part of the algorithm's output), but it can potentially contribute up to $|P|!\,|F|!$. The optimal scenario takes place when the entire fix-set F serves as a perm-set of some permofix selected later in the process. Therefore, a permofix (P, F) having the largest value of $|P|!\,|F|!$ may potentially contribute the largest factor to the target product.

The pseudocode of the greedy algorithm based on positional restrictions of the automorphism set is shown as Algorithm 2.

To show that the output produced by the algorithm conforms to our problem definition, we shall first prove the following lemma:

Lemma 1: Each element of the set returned by the procedure GREEDY2 is a perm-set of the input set A of automorphisms.

Proof: The singletons are perm-sets by definition, so let us focus on the elements of the set $\mathcal{P}$ inside the procedure GREEDY2. In each iteration, the algorithm first applies the procedure FIND2PERMOFIXES to the current set of automorphisms A. This procedure returns a set of all pairs $(\{p, q\}, \{r_1, \ldots, r_t\})$ such that there exists an automorphism a with $a(p) = q$, $a(q) = p$, and $a(r_1) = r_1$, …, $a(r_t) = r_t$. By the definition of automorphism group, the set A always contains the identity automorphism a_{id} with the property $a_{\mathrm{id}}(p) = p$, $a_{\mathrm{id}}(q) = q$, and $a_{\mathrm{id}}(r_1) = r_1$, …, $a(r_t) = r_t$. The automorphisms a and a_{id} jointly form a proof that the pair $(\{p, q\}, \{r_1, \ldots, r_t\})$ is indeed a permofix.

The procedure EXTEND iteratively produces k-permofixes based on sets of $(k-1)$-permofixes in the set of automorphisms A. For $k = 3$, the procedure creates a pair $PF = (\{p, q, r\}, F_1 \cap F_2 \cap F_3)$ from the permofixes $PF_1 = (\{p, q\}, \{r\} \cup F_1)$, $PF_2 = (\{p, r\}, \{q\} \cup F_2)$, and $PF_3 = (\{q, r\}, \{p\} \cup F_3)$. Neglecting the sets F_1, F_2, and F_3 for the time being, the permofix PF_1 represents the permutation $(p\ q)(r)$ in the cycle notation. Likewise, PF_2 and PF_3 represent the permutations $(p\,r)(q)$ and $(q\,r)(p)$, respectively. Since $(A, \circ)$ is a group, the permutation $(p\,q)(r) \circ (p\,r)(q) \circ (q\,r)(p) = (p\,q\,r)$ has to be completely present in A; in other words, A has to contain an automorphism for each of the 3! permutations of the set $\{p, q, r\}$. Therefore, $\{p, q, r\}$ is a perm-set in A. The fix-set corresponding to this perm-set is (a superset of) the intersection of the fix-sets of PF_1, PF_2, and

Algorithm 2 Greedy algorithm based on positional restrictions

```
1:  function GREEDY2(A, V)
2:      // A: a set of automorphisms, V = {1, ..., n}
3:      P := ∅;
4:      W ::= V;
5:      while A contains at least one valid element do
6:          R := CLEANUP(FIND2PERMOFIXES(A));
7:          k := 3;
8:          repeat
9:              R' := R;
10:             R := CLEANUP(EXTEND(R, k));
11:             k := k + 1
12:         until R' = R;
13:         (P_m, F_m) := highest-potential permofix in R;
14:         P := P ∪ {P_m};
15:         W := W \ P_m;
16:         A := ρ(A, F_m)
17:     return P ∪ singletons(W)
18:
19: function FIND2PERMOFIXES(A)
20:     R := ∅;
21:     for all a ∈ A do
22:         for all (i, j): i ≠ j ∧ a(i) = j ∧ a(j) = i do
23:             P := {i, j};
24:             F := {k | a(k) = k};
25:             R := R ∪ {(P, F)}
26:     return R
27:
28: function EXTEND(R, k)
29:     for all P: P ⊆ {1, ..., n} ∧ |P| = k do
30:         F := {1, ..., n};
31:         i := 0;
32:         for all p ∈ P do
33:             if ∃F': (P \ {p}, {p} ∪ F') ∈ R then
34:                 F := F ∩ F';
35:                 i := i + 1
36:             else
37:                 break
38:         if i = k then
39:             R := R ∪ {(P, F)}
40:     return R
41:
42: function CLEANUP(R)
43:     for all (P, F) ∈ R do
44:         for all (P', F') ∈ R \ {(P, F)} do
45:             if (P', F') ⊑ (P, F) then
46:                 R := R \ {(P', F')}
47:     return R
```

PF_3. Consequently, PF is a permofix in A. This reasoning can be straightforwardly extended to the general case of $k > 3$. Therefore, every pair created by the procedure EXTENDS is a permofix in the current set of automorphisms.

The procedure CLEANUP does not produce anything new; it merely reduces the number of permofixes. For a permofix (P, F), all permofixes (P', F') with $(P', F') \sqsubseteq (P, F)$ are heuristically pronounced redundant. If $P' = P$ and $F' \subseteq F$, the permofix (P', F') is clearly superfluous. If $P' \subset P$, then the permofix (P, F) has been created from (P', F') within the EXTEND procedure.

The positional restriction can only reduce the set of permofixes. It is easy to see that if a pair (P, F) is a permofix in a positionally restricted set of automorphisms, then it is a permofix in the original set, too.

In summary, the set $\mathcal{R}$ consists of permofixes of the initial set of automorphisms A, and every element of the set returned from the procedure GREEDY2 is a perm-set of A. ■

In the following theorem, we show that the algorithm produces a solution to our problem, i.e., an exploratory equivalent partition of the vertex set.

Theorem 3: The procedure GREEDY2 returns an exploratory equivalent partition of the vertex set V.

Proof: Let $\{P_1, \ldots, P_s, \{i_1\}, \ldots, \{i_r\}\}$ be the result of the algorithm GREEDY2, where P_1, …, P_s are the perm-sets produced in individual iterations, and $\{i_1\}, \ldots, \{i_r\}$ are the singletons created from the vertices that do not belong to the set $P_1 \cup \ldots \cup P_s$. By construction, the elements of the output set are mutually disjoint sets that collectively cover the entire vertex set. The output set is thus a partition of the vertex set.

By definition, each of the produced perm-sets P_1, …, P_s is covered by the initial set of automorphisms $A_0 \equiv A$, i.e., we have cover(A_0, P_i) for all $i \in \{1, \ldots, s\}$. Let us now show that cover(Stab(A_0, P_s), P_{s-1}) also holds. The perm-set P_s has to be a subset of the fix-set F_{s-1}; otherwise, the algorithm would, at some earlier stage, have set $a_1(j) := \uparrow$, …, $a_{|A|}(j) := \uparrow$ for at least one $j \in P_s$ and hence could not produce P_s. By the definition of permofix, there exists a set of automorphisms that fixes F_{s-1} and simultaneously covers P_{s-1}. Since $P_s \subseteq F_{s-1}$, the same set of automorphisms also fixes P_s. Consequently, the set of automorphisms where P_s is fixed (i.e., Stab(A_0, P_s)) covers P_{s-1}. In the same manner, we can prove cover(Stab(Stab$(A_0, P_s), P_{s-1}), P_{s-2}$), etc. Therefore, the perm-sets P_s, P_{s-1}, P_{s-2}, …, P_1, together with the singleton sets formed by the missing elements, constitute an exploratory equivalent partition of the vertex set V. ■

In practice, the algorithm GREEDY2 is more efficient than GREEDY1. For each combination P of the current set of vertices, the first greedy algorithm checks whether P is covered by the current set of automorphisms (in other words, whether P is a perm-set in the current set of automorphisms). By contrast, the algorithm GREEDY2 generates candidate perm-sets (and the associated fix-sets) in an incremental fashion: a perm-set with k elements is generated by merging k perm-sets with $k - 1$ elements. If no k-element perm-sets are generated, the algorithm will not attempt to generate any $(k + 1)$-element perm-sets.

Let us illustrate the algorithm GREEDY2 with two examples. Consider the graph of Fig. 5. Given the set of its automorphisms as input (enumerated in Eq. 3), the algorithm produces the following 2-permofixes (after executing the procedure CLEANUP):

$$
\begin{array}{lll}
(\{1,2\},\emptyset) & (\{2,3\},\emptyset) & (\{3,6\},\emptyset) \\
(\{1,4\},\emptyset) & (\{2,5\},\emptyset) & (\{4,5\},\emptyset) \\
(\{1,6\},\emptyset) & (\{3,4\},\emptyset) & (\{5,6\},\emptyset) \\
(\{1,3\},\{2,5\}) & (\{1,5\},\{3,6\}) & (\{2,4\},\{3,6\}) \\
(\{2,6\},\{1,4\}) & (\{3,5\},\{1,4\}) & (\{4,6\},\{2,5\})
\end{array}
$$

The procedure EXTEND produces two 3-permofixes: $(\{1,3,5\},\emptyset)$ and $(\{2,4,6\},\emptyset)$. The procedure CLEANUP subsequently removes all permofixes (P, F) with $|P| = |F| = 2$. In the next step, the algorithm selects a permofix with the highest value of $|P|!\,|F|!$. This is either $(\{1,3,5\},\emptyset)$ or $(\{2,4,6\},\emptyset)$. In either case, the fix-set is empty, so the procedure RESTRICT sets all elements of all automorphisms to $\uparrow$. As a result, the algorithm immediately stops with the result $\{1,3,5 \mid 2 \mid 4 \mid 6\}$ (or $\{2,4,6 \mid 1 \mid 3 \mid 5\}$, depending on its selection). Among all exploratory equivalent partitions, these two both have the highest product of the factorials of the cardinalities of their constituent sets and hence represent two optimal solutions to the MAXEXPLOREQ problem.

The graph of Fig. 1 has 8 automorphisms:

$$
\begin{aligned}
a_1 &= (1,2,3,4,5,6) \\
a_2 &= (1,2,3,4,6,5) \\
a_3 &= (1,2,4,3,5,6) \\
a_4 &= (1,2,4,3,6,5) \\
a_5 &= (2,1,5,6,3,4) \\
a_6 &= (2,1,5,6,4,3) \\
a_7 &= (2,1,6,5,3,4) \\
a_8 &= (2,1,6,5,4,3)
\end{aligned}
$$

In the first iteration, the algorithm produces the following permofixes:

$$
\begin{array}{ll}
(\{3,5\},\emptyset) & (\{1,2\},\emptyset) \\
(\{3,6\},\emptyset) & (\{3,4\},\{1,2,5,6\}) \\
(\{4,5\},\emptyset) & (\{5,6\},\{1,2,3,4\}) \\
(\{4,6\},\emptyset) &
\end{array}
$$

The set of automorphisms contains no permofixes (P, F) with $|P| > 2$. Using the highest-potential criterion, the algorithm selects either the permofix $(\{3,4\},\{1,2,5,6\})$ or the permofix $(\{5,6\},\{1,2,3,4\})$. Let us assume that the former is selected; the latter permofix leads to the same output. After the selection, the set of automorphisms is positionally restricted with respect to the fix-set $\{1,2,5,6\}$):

$$
\begin{aligned}
a'_1 &= (1,2,\uparrow,\uparrow,5,6) \\
a'_2 &= (1,2,\uparrow,\uparrow,6,5) \\
a'_5 &= (2,1,\uparrow,\uparrow,3,4) \\
a'_6 &= (2,1,\uparrow,\uparrow,4,3)
\end{aligned}
$$

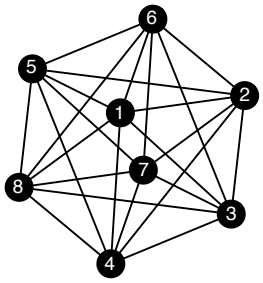


Fig. 6. The smallest graph on which Algorithm 2 returns a suboptimal solution.

The automorphisms a'_3 and a'_4 are equal to a'_1 and a_2', respectively, and an analogous situation occurs with the automorhpisms a'_7 and a'_8. In the second iteration, only two permofixes are produced: $(\{1,2\},\emptyset)$ and $(\{5,6\},\{1,2\})$. The latter has a greater potential than the former and is hence selected, restricting the set of automorphisms to $\{(1,2,\uparrow,\uparrow,\uparrow,\uparrow), (2,1,\uparrow,\uparrow,\uparrow,\uparrow)\}$. The restricted automorphisms give rise to the sole permofix $(\{1,2\},\emptyset)$, which is selected in the third iteration of the algorithm. The algorithm thus outputs the partition $\{3,4 \mid 5,6 \mid 1,2\}$, which is again an optimal solution to the MAXEXPLOREQ problem.

For a vast majority of input graphs, the algorithm GREEDY2 produces optimal exploratory equivalent partitions. The smallest graph (in terms of vertex count) with a suboptimal result is shown in Fig. 6. For this graph, the algorithm produces the partition $\{1,7 \mid 2,8 \mid 3,5 \mid 4,6\}$ with the target cardinality factorial product being $2!\,2!\,2!\,2! = 16$. The optimal solution, however, is the partition $\{1,2,3,4 \mid 5 \mid 6 \mid 7 \mid 8\}$ with the target product of $4! = 24$. In the first iteration, the algorithm produces 20 permofixes, two of which are $(\{1,2,3,4\},\ \emptyset)$ and $(\{1,7\},\{2,3,4,5,6,8\})$. The former permofix would lead to an optimal solution, but the algorithm chooses the latter, since $2!\,6! > 4!$. However, the fix-set of the selected permofix eventually contributes only $2!\,2!\,2!$ instead of the potential $6!$ to the target product, making the algorithm's first-iteration choice suboptimal.

Interestingly, the graphs of Fig. 4 are not counterexamples for the second greedy algorithm, and the graph of Fig. 6 is not a counterexample for the first algorithm. In contrast to the algorithm GREEDY1, the algorithm GREEDY2 considers the combined sizes of individual perm-sets and fix-sets when making greedy selections. In the right graph of Fig. 4, for example, the algorithm GREEDY2 has to choose between the permofix $(\{2,3\},\{1,4,5,6,7\})$ (or an equivalent permofix with potential $2!\,5!$) and the permofix $(\{2,4,6\},\emptyset)$ (or an equivalent permofix with potential $3!$). The first permofix is obviously preferable, leading to an optimal partition. Conversely, since the algorithm GREEDY1 considers perm-sets without the associated fix-sets, it prefers the perm-set $\{1,2,3,4\}$ over all 2-element perm-sets (regardless of the sizes of their associated fix-sets) when dealing with the graph of Fig. 6.

V. EXPLORATORY EQUIVALENCE AND THE IMPROVED REKERS-SCHÜRR PARSER

As we mentioned in the introduction, the concept of exploratory equivalence was developed by Fürst *et al.* [9] for the purpose of improving the Rekers-Schürr graph grammar parser [10], although the authors did not provide a rigorous graph-theoretic and group-theoretic definition of exploratory equivalence and did not consider the possibility of multiple exploratory equivalent partitions for a single graph. In this section, we show how a proper consideration of exploratory equivalence may lead to immense performance gains when parsing graphs against graph grammars.

The Rekers-Schürr graph grammar parser (both the original and the improved version) accepts a graph and a context-sensitive graph grammar on its input. A context-sensitive graph grammar (called just 'grammar' in the sequel) is a quadruple $(\mathcal{N}, \mathcal{T}, \mathcal{P}, \mathcal{A})$, where $\mathcal{N}$ is a set of *nonterminal* labels, $\mathcal{T}$ is a set of *terminal* labels, $\mathcal{P}$ is a set of *productions*, and $\mathcal{A}$ is a set of *axioms*. Each production p is a rule of the form $Lhs[p] ::= Rhs[p]$, where $Lhs[p]$ (the left-hand side – LHS) and $Rhs[p]$ (the right-hand side – RHS) are subgraphs of a graph $Union[p]$ whose elements (vertices and edges) have labels from $\mathcal{N} \cup \mathcal{T}$. The graph $Common[p] = Lhs[p] \cap Rhs[p]$ is called the *context* of the production. Let $Xlhs[p] = Lhs[p] \setminus Common[p]$ and $Xrhs[p] = Rhs[p] \setminus Common[p]$; note that $Xlhs[p]$ and $Xrhs[p]$ might not be proper graphs, since they may contain dangling edges. A sample production, as well as the graphs and the graph element sets associated with it, is shown in Fig. 7. In contrast to the graph depictions shown so far, the inscriptions inside the vertices represent vertex labels rather than vertex indices. The indices are displayed next to individual vertices. The yellow-colored vertices belong to the graph $Common[p]$ and hence to both the LHS and RHS simultaneously; this is also reflected in the fact that such vertices have the same index on both sides of the production.

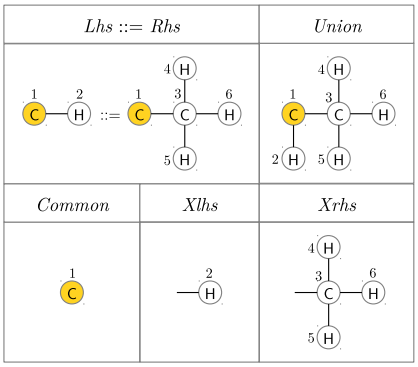


Fig. 7. A sample production and the associated graphs and graph element sets.

An *l-homomorphism* $h\colon Lhs[p] \to G$ for a production p is a graph homomorphism whose restriction to $Xlhs[p]$ is injective. An *l-occurrence* of a production p in a graph G is a graph $L' \sqsubseteq G$ such that $L' = h(Lhs[p])$ for some l-homomorphism h. The terms *r-homomorphism* and *r-occurrence* are defined symmetrically (with $Rhs[p]$ and $Xrhs[p]$ instead of $Lhs[p]$ and $Xlhs[p]$, respectively).

To *apply* a production p to a graph G, the following three steps are performed: (1) find an l-occurrence of p in G (let $h\colon Lhs[p] \to G$ be the associated l-homomorphism); (2) remove the elements $h(Xlhs[p])$ from the graph G; (3) attach fresh copies of the elements $Xrhs[p]$ to the elements

$h(Common[p])$ in the same way as the elements $Xrhs[p]$ are attached to the elements of $Common[p]$ within the graph $Rhs[p]$. A *derivation* of a graph G in a graph grammar is a sequence of production applications beginning with an axiom graph and ending with the graph G. The *language* of a graph grammar GG is the set of all terminally labeled graphs that have a derivation in GG. (A graph is *terminally labeled* if all of its elements are labeled by labels from the set $\mathcal{T}$.) A *parser* is an algorithm that, for a given graph G and a given graph grammar GG, determines whether G belongs to the language of GG and produces a derivation of G in GG if this is the case. Figure 8 shows a grammar for generating the structural formulas of linear alkanes. All graph labels belong to the set $\mathcal{T}$, including the 'non-label' — a fictitious label for unlabeled edges. Figure 9 displays the derivation of the propane graph in that grammar. The derivation starts with the axiom (the methane graph) and passes through the ethane graph.

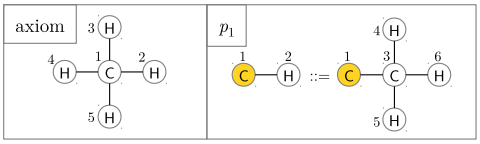


Fig. 8. A grammar for generating the structural formulas of linear alkanes.

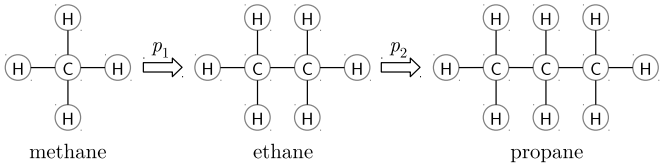


Fig. 9. Derivation of the propane graph in the grammar of Fig. 8.

The Rekers-Schürr parser works as a two-stage process. In the first stage, the input graph G is analyzed in order to obtain a partially ordered redundant set $\mathcal{S}$ of candidate production applications that might take part in a potential derivation of G. In the second stage, the parser tries to find, using backtracking if necessary, a sequence of production applications within the set $\mathcal{S}$ that constitutes a correct derivation of the graph G. The improvement by Fürst *et al.* pertains only to the first stage of the parsing process.

At the beginning of the first stage, the parser creates a graph $\overline{G}$ as a copy of the input graph G. After that, it iteratively searches the graph $\overline{G}$ for all r-occurrences of individual productions. For each discovered r-occurrence of a production p, the graph $\overline{G}$ is augmented by attaching fresh copies of the elements $Xlhs[p]$ to the r-occurrence, giving rise to a *production instance* — a homomorphic image of the entire production p that defines a candidate application of p in a potential derivation of G. The augmentation of the graph $\overline{G}$ might result in new r-occurrences among the added elements. The discover-and-augment cycle finishes once all r-occurrences of all productions have been discovered.

To guarantee the discovery of all r-occurrences, each RHS has to be matched against the graph $\overline{G}$ in all possible ways. In other words, all RHS-to-$\overline{G}$ r-homomorphisms have to be established, including different r-homomorphisms between a production and *each* of its r-occurrences. However, exploratory equivalence can make some (or all) of the r-homomorphisms between a production and its r-occurrence redundant. Let us assume that a production p contains k distinct vertices $v_1, \ldots, v_k$ with the following properties:

- either $v_1, \ldots, v_k \in Xrhs[p]$ or $v_1, \ldots, v_k \in Common[p]$;
- $v_1, \ldots, v_k$ constitute an equivalence class in at least one exploratory equivalent partition of the graph $Rhs[p]$;
- $v_1, \ldots, v_k$ constitute an equivalence class in at least one explorationally equivalent partition of the graph $Union[p]$.

Then it can be shown [9] that the set of r-homomorphisms $h\colon Rhs[p] \to \overline{G}$ established between the production p and the graph $\overline{G}$ can be safely restricted to those r-homomorphisms h for which $index(h(v_1)) < \ldots < index(h(v_k))$, where $index(v)$ is a unique index assigned to a vertex $x \in \overline{G}$. This rule reduces the number of established p-homomorphisms between the production p and each of its occurrences by a factor of $k!$. Since each discovered r-homomorphism is followed by an augmentation of the graph $\overline{G}$, immense performance gains can thus be attained. This rule can be straightforwardly extended to multiple non-singleton classes of an exploratory equivalent partition.

Consider the grammar of Fig. 8. The optimal exploratory equivalent partition for the axiom graph is $\{1 \mid 2,3,4,5\}$. This implies that we can employ the rule $h(2) < h(3) < h(4) < h(5)$ whenever searching for occurrences of the axiom graph in the graph $\overline{G}$. For the RHS of the production p_1, the optimal partition is $\{1 \mid 3 \mid 4,5,6\}$. Since the graph $Union[p_1]$ also has an exploratory equivalent partition in which the vertices 4, 5, and 6 are part of the same equivalence class, we can enforce the rule $h(4) < h(5) < h(6)$ for every r-homomorphism established between the RHS of the production p_1 and the graph $\overline{G}$. Because of the interleaved discover-and-augment cycle, the enforcement of these rules may significantly reduce the parsing time.

For the task of parsing the graphs of methane, ethane, and propane against the grammar of Fig. 8, Table I compares the duration of parsing without considering exploratory equivalence (EE) and the duration of parsing when exporatory equivalence is taken into account in the form of imposing constraints on r-homomorphisms between the RHSs and the graph $\overline{G}$. The experiments were conducted on a 3.40-GHz Intel Core i7 machine.

The difference between the two versions of the parser is striking. Without using the rules based on exploratory equivalence, the parser quickly succumbs to a combinatorial explosion as the size of the input graph increases; it took more than 11 hours to parse the graph of propane with 3 vertices C and 8 vertices H. By contrast, when exploratory equivalence is taken into account, the parser takes less than one second (0.989 seconds) even when parsing the graph $C_{30}H_{62}$ (30 vertices C, 62 vertices H). Asymptotically, for a graph with n vertices C, the original parser creates $\Omega(6^n)$ production instances (possibly much more than that), while the version that makes use of exploratory equivalence generates exactly

$12n - 7$ production instances. For many grammars containing symmetries in the sense of exploratory equivalence, the use of exploratory equivalence can reduce the asymptotical parsing time from exponential to polynomial (see [9] for additional examples).

TABLE I. THE TIME REQUIRED TO PARSE THE INDIVIDUAL GRAPHS OF FIG. 9 AGAINST THE GRAMMAR OF FIG. 8.

Graph	Without EE	With EE
methane (CH_4)	0.16	0.14
ethane (C_2H_6)	1.03	0.15
propane (C_3H_8)	41000.	0.16

VI. CONCLUSION

We introduced a novel type of graph equivalence, called exploratory equivalence because of its applicability to various graph search algorithms. Exploratory equivalence was defined as an automorphism-based equivalence relation on graph vertices. In contrast to our usual perceptions about equivalence, exploratory equivalence may induce several distinct vertex set partitions for a given graph.

In addition to defining exploratory equivalence itself, we have also introduced the concept of an optimal exploratory equivalent partition for a given graph. We presented two greedy algorithms for finding such a partition. Both algorithms produce optimal results for a vast majority of input graphs. For instance, considering all non-isomorphic graphs on 8 vertices, the second greedy algorithm produces an optimal partition for 11116 graphs out of 11117, the sole exception being the graph of Fig. 6. Among all non-isomorphic 9-vertex graphs, the algorithm produces suboptimal results for only 2 graphs out of 261080.

In subgraph search algorithms, exploratory equivalence can be employed to prevent or at least reduce multiple discoveries of individual occurrences of graph patterns in a given host graph. In the Rekers-Schürr graph grammar parser, this strategy may bring about immense performance gains, since each discovery of a graph in a host graph results in an augmentation of the same host graph.

A possible direction for the future work is a generalization of exploratory equivalence. As defined in this paper, exploratory equivalence can be regarded as a global relation between vertices. Informally, a pair of vertices may potentially belong to the same exploratory equivalence class only if the entire graph 'looks the same' from the viewpoint of both vertices. For this reason, exploratory equivalence is a fairly infrequent phenomenon for large random graphs, except for sets of leaf vertices attached to the same internal vertex. A natural generalization of 'global' exploratory equivalence is therefore a 'local' version of this concept, where only a limited neighborhood is inspected when determining the equivalence of a set of vertices. However, practical implications of such a definitions have yet to be discovered.

As shown in Section V, exploratory equivalence can be used to impose constraints on graph homomorphisms when searching for occurrences of a given pattern graph inside a given host graph. The purpose of such constraints is to eliminate multiple discoveries of the same occurrence. However, in some cases, the constraints induced by exploratory equivalence do not suffice to cover all automorphisms of the pattern graph. Consider, for example, the graph of Fig. 5. This graph has 12 automorphisms, but the optimal exploratory equivalent partition ($\{\{1, 3, 5 \mid 2 \mid 4 \mid 6\}\}$) only covers half of them. Consequently, the rule $h(1) < h(3) < h(5)$ still allows for two different isomorphisms between a pair of 6-cycles. Besides the constraints induced by the optimal exploratory equivalence, we would need another constraint to cover the rotational symmetry of the graph. The relationship between exploratory equivalence (and other types of equivalence) and graph search constraints is thus another promising direction for the future work.

REFERENCES

[1] D. K. Agrafiotis, V. S. Lobanov, M. Shemanarev, D. N. Rassokhin, S. Izrailev, E. P. Jaeger, S. Alex, and M. Farnum, "Efficient Substructure Searching of Large Chemical Libraries: The ABCD Chemical Cartridge," *J. Chem. Inf. Model.*, 2011. doi: 10.1021/ci200413e

[2] J. M. Barnard, "Substructure searching methods: Old and new," *J. Chemical Information and Computer Sciences*, vol. 33, no. 4, pp. 532–538, 1993. doi: 10.1021/ci00014a001

[3] M. O. Jackson, *Social and Economic Networks*. Princeton, NJ, USA: Princeton University Press, 2008. ISBN 0691134405, 9780691134406

[4] D. Knoke, *Political Networks: The Structural Perspective*, ser. Structural Analysis in the Social Sciences. Cambridge University Press, 1994. ISBN 9780521477628

[5] B. Hopkins, "Kevin Bacon and graph theory," *PRIMUS*, vol. 14, no. 1, pp. 5–11, 2004. doi: 10.1080/10511970408984072

[6] F. V. Fomin and D. Kratsch, *Exact Exponential Algorithms*. Springer, 2011.

[7] J. R. Ullmann, "An Algorithm for Subgraph Isomorphism," *J. Assoc. for Computing Machinery*, vol. 23, pp. 31–42, 1976. doi: 10.1145/321921.321925

[8] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento, "A (sub)graph isomorphism algorithm for matching large graphs." *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 10, pp. 1367–72, Oct. 2004. doi: 10.1109/TPAMI.2004.75

[9] L. Fürst, M. Mernik, and V. Mahnič, "Improving the graph grammar parser of Rekers and Schürr," *IET Software*, vol. 5, no. 2, pp. 246–261, 2011. doi: 10.1049/iet-sen.2010.0081

[10] J. Rekers and A. Schürr, "Defining and parsing visual languages with Layered Graph Grammars," *Journal of Visual Languages and Computing*, vol. 8, no. 1, pp. 27–55, 1997. doi: 10.1006/jvlc.1996.0027

[11] H. Ehrig, G. Engels, H.-J. Kreowski, G. Rozenberg, and U. Montanari, Eds., *Handbook of graph grammars and computing by graph transformation (Vols. 1.–3.)*. World Scientific, 1997–1999.

[12] G. Rozenberg and E. Welzl, "Boundary NLC graph grammars – basic definitions, normal forms, and complexity," *Information and Control*, vol. 69, no. 1–3, pp. 136–167, 1986. doi: 10.1016/S0019-9958(86)80045-6

[13] L. Liberti, "Automatic generation of symmetry-breaking constraints," in *COCOA*, ser. Lecture Notes in Computer Science, B. Yang, D.-Z. Du, and C. Wang, Eds., vol. 5165. Springer, 2008. doi: 10.1007/978-3-540-85097-7_31 pp. 328–338.

[14] A. Mucherino, C. Lavor, and L. Liberti, "Exploiting symmetry properties of the discretizable molecular distance geometry problem," *J. Bioinformatics and Computational Biology*, vol. 10, no. 3, 2012. doi: 10.1142/S0219720012420097

[15] B. D. McKay and A. Piperno, "Practical graph isomorphism, ii," *J. Symbolic Computation*, vol. 60, pp. 94–112, 2013. doi: 10.1016/j.jsc.2013.09.003

[16] M. G. Everett and S. P. Borgatti, "Regular equivalence: General theory," *Journal of mathematical sociology*, vol. 19, no. 1, pp. 29–52, 1994. doi: 10.1080/0022250X.1994.9990134

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 457–463
DOI: 10.15439/2014F92
ACSIS, Vol. 2

An adaptive branching scheme for the Branch & Prune algorithm applied to Distance Geometry

Douglas Gonçalves*, Antonio Mucherino*, Carlile Lavor†

*IRISA, University of Rennes 1, Rennes, France.
{douglas.goncalves,antonio.mucherino}@irisa.fr

†IMECC-UNICAMP, Campinas-SP, Brazil.
clavor@ime.unicamp.br

***Abstract*—The Molecular Distance Geometry Problem (MDGP) is the one of finding molecular conformations that satisfy a set of distance constraints obtained through experimental techniques such as Nuclear Magnetic Resonance (NMR). We consider a subclass of MDGP instances that can be discretized, where the search domain has the structure of a tree, which can be explored by using an *interval* Branch & Prune (*i*BP) algorithm. When all available distances are exact, all candidate positions for a given molecular conformation can be enumerated. This is however not possible in presence of interval distances, because a continuous subset of positions can actually be computed for some atoms. The focus of this work is on a new scheme for an adaptive generation of a discrete subset of candidate positions from this continuous subset. Our generated candidate positions do not only satisfy the distances employed in the discretization process, but also additional distances that might be available (the so-called pruning distances). Therefore, this new scheme is able to guide more efficiently the search in the feasible regions of the search domain. In this work, we motivate the development and formally introduce this new adaptive scheme. Presented computational experiments show that *i*BP, integrated with our new scheme, outperforms the standard *i*BP on a set of NMR-like instances.**

I. Introduction

LET $G = (V, E, d)$ be a simple weighted undirected graph where the vertices V represent the points of a Euclidean space and where $d : E \to \mathbb{R}_+$ assigns positive weights d_{uv} to edges (u, v) when the distance between u and v is available. The Distance Geometry Problem (DGP) [9] asks to find an embedding $x : V \to \mathbb{R}^3$ satisfying constraints based on the available edge weights, i.e. to find a conformation x in the Euclidean space such that:

$$\underline{d}_{uv} \leq \|x(u) - x(v)\| \leq \overline{d}_{uv}, \quad \forall (u, v) \in E, \tag{1}$$

where $\underline{d}_{uv}$ and $\overline{d}_{uv}$ denote, respectively, the lower and upper bounds for the distance d_{uv} ($\underline{d}_{uv} = \overline{d}_{uv}$ if d_{uv} is an exact distance).

One of the most interesting applications of the DGP arises in biology, where vertices of G represent atoms of a given molecule, and weighted edges provide the relative distances between some atom pairs. When molecules are concerned, the DGP is generally referred to as the Molecular DGP (MDGP) [3], [5]. The interested reader can make reference to a recent survey [9] and to an edited book [13] for additional information about the MDGP and methods for its solution.

The MDGP is generally formulated as a continuous optimization problem where the objective function is a penalty function capable of measuring the violation of the constraints. Under certain assumptions, the domain of this optimization problem can be discretized, so that it becomes combinatorial [7], [12]. The discrete search domain has the structure of a tree, where the candidate positions for a given atom of the molecule belong to the same layer of the tree. We employ an *interval* Branch & Prune (*i*BP) algorithm [8] for exploring such a tree with the aim of finding solutions to discretizable MDGPs. The reader is referred to Section II for more details about the discretization.

The basic idea behind the *i*BP algorithm is to construct the search domain of the optimization problem branch by branch (*branching phase*), and to verify, every time a new branch is added, whether it is feasible or not (*pruning phase*). Atomic positions are generated by intersecting 3 Euclidean objects (spheres and spherical shells), which we can define on each layer of the tree because of the discretization assumptions. When discovered, infeasible positions are pruned away, so that the search can be focused on the parts of the tree where there are feasible solutions. Only a subset of available distances is employed in the discretization process (the *discretization distances*), while others can be exploited for pruning purposes (the *pruning distances*).

In the discretization process, if all considered distances are exact, there can be at most two feasible positions for the current atom [7]. If some distances are represented by intervals, the feasible positions belong to a continuous Euclidean object, that can be discretized by sampling D candidate positions [8]. In this phase, the number D of chosen sample positions plays a very important role.

Experiments reported in previous publications (see for example [2], [8]) show in fact that the obtained results can be strongly influenced by the choice of D. If D is too small, only

978-83-60810-58-3/$25.00

infeasible branches may be generated, so that the whole tree is pruned and no solutions are found. On the other hand, if D is too large, the consequent combinatorial explosion might make the experiments too expensive. Finding a trade-off D value is not an easy task in general.

This paper presents a new scheme for an adaptive branching during the execution of the *i*BP algorithm, which is based on the idea of including, during the intersection of the Euclidean objects related to the known discretization distances, other objects, related to pruning distances, that might be available at the current layer. This way, it is possible to generate branches that are feasible, with respect to the pruning distances, up to the current layer.

The rest of the paper is organized as follows. In Section II, we will briefly discuss the discretization assumptions, present the *i*BP algorithm, and give some details about the generation of the coordinates of candidate positions at each iteration of *i*BP. In Section III, we will propose a new scheme, based on the intersection of several Euclidean objects, for the computation of candidate positions that are *all feasible* at the current layer. Section IV will show some experiments on artificially generated instances, while conclusions will be drawn in Section V.

II. The *i*BP algorithm

Let $G = (V,E,d)$ be an MDGP instance. The subclass of MDGP instances that we consider in this paper is defined as follows. Let $E' \subset E$ be the subset of edges for which their weights d are exact distances.

The interval *Discretizable DGP in dimension 3 (iDDGP$_3$).* Given a simple weighted undirected graph $G = (V,E,d)$, we say that G represents an instance of the *i*DDGP$_3$ if and only if there exists an order on the vertices of V verifying the following conditions:

(a) $G_C = (V_C, E_C) \equiv G[\{1,2,3\}]$ is a clique and $E_C \subset E'$;
(b) $\forall i \in \{4,\ldots,|V|\}$, there exists $\{i',i'',i'''\}$ such that
  1) $i''' < i$, $i'' < i$, $i' < i$;
  2) $\{(i'',i),(i',i)\} \subset E'$ and $(i''',i) \in E$;
  3) $d_{i',i'''} < d_{i',i''} + d_{i'',i'''}$.

Orders satisfying (a) and (b) are named "Discretization Orders". We refer to $\{i''',i'',i'\}$ as *reference atoms*, and to $d_{i''',i}$, $d_{i'',i}$ and $d_{i',i}$ as *reference distances*.

Notice that assumption (a) allows us to place the first 3 atoms uniquely, avoiding to consider congruent solutions that can be obtained by rotations and translations [7]. Assumption (b1) ensures the existence of three reference atoms for every $i > 3$, and assumption (b2) ensures that at most one of the three reference distances may be represented by an interval [12]. Finally, assumption (b3) avoids the reference atoms to be collinear. We remark that assumption (b3) cannot always be verified before the solution of an instance, because some of the necessary distances may not be available (the corresponding edges may not be in E). However, this assumption can fail to be satisfied with probability 0, and therefore we do not really need to verify it in advance [6]. Under these assumptions, the MDGP can be discretized, i.e. the instance at hand belongs to the *i*DDGP$_3$ class. In this case, the search domain becomes a tree, where nodes contain candidate atomic positions, organized layer by layer.

Algorithm 1 The *i*BP algorithm.

```
1:  iBP(i,n,d,D)
2:  if (i > n) then
3:    // one solution is found
4:    print current conformation;
5:  else
6:    // coordinate computation
7:    if (d_{i''',i} is an interval) then
8:      compute the two candidate arcs;
9:      add them to the list L;
10:   else
11:     compute the two candidate positions;
12:     add them to the list L;
13:   end if
14:   for h = 1,...,|L| do
15:     if (L(h) is an arc) then
16:       take D samples from the arc; set N = D;
17:     else
18:       set N = 1;
19:     end if
20:     // verifying the feasibility of the computed positions
21:     for k = 1,...,N do
22:       if (x_i^{h,k} is feasible) then
23:         iBP(i+1,n,d,D);
24:       end if
25:     end for
26:   end for
27: end if
```

We employ an *interval* Branch & Prune (*i*BP) algorithm [8] for the solution of discretizable instances. Alg. 1 is a sketch of this algorithm. The *i*BP algorithm performs a recursive search on the tree which represents the search domain. At each recursive call, candidate positions for the current atom are computed by exploiting the coordinates of previously placed atoms and the distance information ensured by the discretization assumptions. When all reference distances are exact, then two candidate positions are computed. When one of the references is an interval, two *feasible arcs* are rather identified (see Fig. 1).

In the algorithm call, i is the current atom for which candidate positions are currently searched, n is the total number of atoms forming the considered molecule, d is the list of available distances (exact and interval distances), and D is the discretization factor, i.e. the number of sample points that are taken from the arcs when the distance $d_{i''',i}$ is represented by an interval (see assumption (b2)). In the algorithm (see lines 9 and 12), we make use of a list L of positions and arcs, from which candidate positions are extracted.

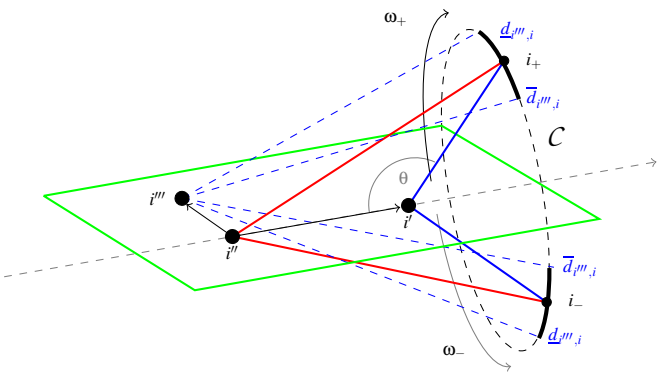


Fig. 1. The two feasible arcs (in bold, black) obtained by intersecting two spheres and one spherical shell.

When working on the atom i, feasible positions for its three reference atoms $\{i''', i'', i'\}$, on the current tree branch, are already available. These reference atoms define a local coordinate system centered at i' [4], [14]. The possible positions for the atom i verifying $d_{i',i}$ and $d_{i'',i}$ can be described by two angles θ_i and ω_i. Using $d_{i'',i'}$ and the cosine law, we can obtain a value for $\theta_i \in [0,\pi]$. Thus, the circle $\mathcal{C}$ of possible positions for atom i (see Fig. 1) can be described in terms of ω_i:

$$x_i(\omega_i) = x_{i'} + U_{i'} w_i, \tag{2}$$

where

$$w_i = \begin{bmatrix} -d_{i',i}\cos\theta_i \\ d_{i',i}\sin\theta_i\cos\omega_i \\ d_{i',i}\sin\theta_i\sin\omega_i \end{bmatrix},$$

$\omega_i \in [0, 2\pi]$, and $U_{i'}$ is the rotation (change of basis) matrix from the local system at i' to the canonical system of coordinates [4].

If $d_{i''',i}$ is exact, at most two values for ω_i, say $\{\omega_i^+, \omega_i^-\}$, can be computed. Two possible positions x_i^+ and x_i^- can be therefore identified for the atom i. These positions are symmetric with respect to the plane defined by the reference atoms. If $d_{i''',i}$ is instead an interval, then two disjoint and symmetric candidate arcs are obtained, as shown in Fig. 1. They correspond to two intervals, $[\underline{\omega}_i^+, \overline{\omega}_i^+]$ and $[\underline{\omega}_i^-, \overline{\omega}_i^-]$, for the angle ω_i. By selecting D equidistant angles in $[\underline{\omega}_i^+, \overline{\omega}_i^+]$ and other D equidistant angles in $[\underline{\omega}_i^-, \overline{\omega}_i^-]$, $2 \times D$ atomic positions for the current atom i can be computed.

In the standard *i*BP, the feasibility of these candidate atomic positions is verified by exploiting the so-called pruning distances. The Direct Distance Feasibility (DDF) is the pruning device that, for each candidate position related to the current atom i, verifies whether the inequality

$$\underline{d}_{ij} - \varepsilon \leq ||\mathrm{x}_i - \mathrm{x}_j|| \leq \overline{d}_{ij} + \varepsilon, \tag{3}$$

is satisfied for each atom $j < i$ that is not involved in the discretization, where $\varepsilon > 0$ is a given tolerance. This way, however, a large number of generated positions may be pruned and only a few of them may be actually feasible. The scheme we propose in this paper aims at overcoming this issue.

Finally, we remark that an essential pre-processing step, before applying the *i*BP, is to find a discretization order for the vertices of the graph G that allow to satisfy the assumptions in the *i*DDGP$_3$ definition. This preprocessing step can be performed efficiently, in polynomial time [11], so that the necessary assumptions can be fulfilled by graphs related to proteins.

III. Adaptive branching in *i*BP

The discretization of the two candidate arcs, used in the standard *i*BP algorithm when interval data are available, represents the simplest way to deal with imprecise information about the distances [8]. The candidate arcs are discretized by considering a finite number of samples in the two intervals $[\underline{\omega}_i^+, \overline{\omega}_i^+]$ and $[\underline{\omega}_i^-, \overline{\omega}_i^-]$, and then a new branch is created for each of them. If D is the discretization factor (see Alg. 1), $2 \times D$ positions are generated, and $2 \times D$ new branches are added to the tree at the current layer. When considering this approach, it is expected that at least one of such samples is able to fulfill the pruning distance constraints at the current layer.

The value given to D plays a critical role. On the one hand, too small values can generate trees where no solutions can be found (all branches are pruned, because no positions are compatible to the pruning distances). On the other hand, too large D values can drastically increase the width of the tree. Unfortunately, no upper bound on D can theoretically be defined: in the case only one specific singleton in the given arcs is actually feasible, only an infinite number of samples could guarantee that this singleton can be discovered. However, this is the worst case scenario: nondegenerate subarcs generally result to be feasible w.r.t. the available pruning distances.

In the standard *i*BP, after the generation of candidate atomic positions, their feasibility is verified by employing pruning devices, such as DDF (see Section II). There are two extreme situations:

1) all positions are feasible: this suggests that we could consider a smaller D value without harming the computations;
2) all positions are infeasible: since a finite number of samples on the two arcs are taken, this information does not allow us to discriminate between "the two arcs are infeasible" and "the chosen samples are infeasible".

The adaptive scheme that we propose was conceived for tailoring the branching phase of the *i*BP algorithm so that all computed candidate positions are feasible at the current layer. The basic idea is to identify, before the branching phase of the algorithm, the subset of positions (if it exists) on the two candidate arcs that is feasible with respect to all pruning distances to be verified at the current layer.

Let us consider expression (2), which is able to give the Cartesian coordinates of the atom i as a function of the torsion angle ω_i. For simplifying the notations, we will omit, in the following, the subscripts of the angles θ_i and ω_i.

In case the distance $d_{i''',i}$ is represented by an interval, i.e. $d_{i''',i} \in [\underline{d}_{i''',i}, \overline{d}_{i''',i}]$, two candidate arcs can be computed (see

Section II). These two arcs correspond to the two interval torsion angles $[\underline{\omega}^+,\overline{\omega}^+] \subset [0,\pi)$ and $[\underline{\omega}^-,\overline{\omega}^-] \subset [\pi,2\pi)$. All points in those two arcs satisfy therefore the interval distance $[\underline{d}_{i''',i},\overline{d}_{i''',i}]$, as well as the two exact distances $d_{i'',i}$ and $d_{i',i}$. However, there can be pruning distances (between already placed atoms and i) that we could exploit for tightening these two arcs. Therefore, instead of using these distances for pruning pre-computed positions, our idea is to exploit pruning distances for tightening the two arcs before sampling, so that all generated positions can be feasible (at least at the current layer).

Tightening the feasible arcs

Let us suppose there is an $h \in \{j < i \mid j \notin \{i''',i'',i'\}\}$, such that the pruning distance $d_{h,i}$ is known. Solutions to the equation

$$d_{h,i} = \|x_h - x_i(\omega)\| \tag{4}$$

give the values for the angle ω that are compatible with the distance $d_{h,i}$. By squaring equation (4), and by using (2), we obtain

$$\begin{aligned} d_{h,i}^2 &= \|x_h - x_i(\omega)\|^2 \\ &= \|x_h - (x_{i'} + U_{i'} w_i)\|^2 \\ &= \|x_h - x_{i'}\|^2 - 2\langle x_h - x_{i'},\ U_{i'} w_i\rangle + \|U_{i'} w_i\|^2, \end{aligned}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product between two vectors. Since $U_{i'}$ is an orthogonal matrix, we have

$$d_{h,i}^2 = \|x_h - x_{i'}\|^2 - 2\langle x_h - x_{i'},\ U_{i'} w_i\rangle + d_{i',i}^2.$$

Let $v = x_h - x_{i'}$ and let $\hat{x},\hat{y},\hat{z}$ be the columns of $U_{i'}$. Then:

$$\begin{aligned} d_{h,i}^2 &= \|v\|^2 - 2\langle v,\ U_{i'} w_i\rangle + d_{i',i}^2 \\ &= \|v\|^2 + d_{i',i}^2 - 2\langle v, (-d_{i',i}\cos\theta)\hat{x} + \\ &\quad (d_{i',i}\sin\theta\cos\omega)\hat{y} + (d_{i',i}\sin\theta\sin\omega)\hat{z}\rangle \\ &= \|v\|^2 + d_{i',i}^2 - 2\big(\langle v,\hat{x}\rangle(-d_{i',i}\cos\theta) + \\ &\quad \langle v,\hat{y}\rangle(d_{i',i}\sin\theta)\cos\omega + \langle v,\hat{z}\rangle(d_{i',i}\sin\theta)\sin\omega\big). \end{aligned}$$

If we set

$$A = 2\langle v,\hat{y}\rangle(d_{i',i}\sin\theta), \tag{5}$$

$$B = 2\langle v,\hat{z}\rangle(d_{i',i}\sin\theta),$$

$$\Delta = \|v\|^2 + d_{i',i}^2 + 2\langle v,\hat{x}\rangle(d_{i',i}\cos\theta),$$

and

$$C = \Delta - d_{h,i}^2,$$

we obtain the following equation:

$$A\cos\omega + B\sin\omega = C. \tag{6}$$

Solving $A\cos\omega + B\sin\omega = C$

In order to solve equation (6), we consider the following approach. We set

$$A = R\cos\alpha, \tag{7}$$

$$B = R\sin\alpha, \tag{8}$$

and, in order to obtain R, we square and sum the two equations (7) and (8):

$$A^2 + B^2 = R^2\cos^2\alpha + R^2\sin^2\alpha = R^2(\cos^2\alpha + \sin^2\alpha) = R^2.$$

If we consider the positive square root (R can be seen as the length of a triangle side), we have

$$R = \sqrt{A^2 + B^2}.$$

If $A \neq 0$, we can divide (8) by (7), and obtain

$$\frac{B}{A} = \frac{\sin\alpha}{\cos\alpha} = \tan\alpha,$$

or, equivalently

$$\alpha = \tan^{-1}\left(\frac{B}{A}\right).$$

The correct quadrant for α can be identified by checking the signs of $\cos\alpha$ and $\sin\alpha$.

Notice that, when both A and B are zero, we can have either no solutions or an infinite number of solutions. When $A = B = 0$, then $v = x_h - x_{i'}$ is on the $\hat{x}$ axis, because $\sin\theta \neq 0$ (assumption (b3)) and $d_{i',i} > 0$ (see equation (5)). Atoms h,i'',i' are therefore aligned and the sphere centered in x_h does match with the whole dashed circle $\mathcal{C}$ (when there are infinite solutions) or does not (when there are no solutions). If $A = 0$ and $B \neq 0$, then $\cos\alpha = 0$ and α is either $\pi/2$ or $-\pi/2$, depending on the sign of B.

When $A \neq 0$, from equations (6), (7) and (8), we can obtain

$$\begin{aligned} A\cos\omega + B\cos\omega &= R\cos\alpha\cos\omega + R\sin\alpha\sin\omega \\ &= R\cos(\omega - \alpha), \end{aligned}$$

and hence

$$R\cos(\omega - \alpha) = C,$$

which is

$$\omega = \alpha \pm \cos^{-1}\left(\frac{C}{R}\right). \tag{9}$$

Therefore, we usually have two solutions for equation (6) in $[0,2\pi)$. There are two exceptions. When $C = R$, we have only one solution; when $C/R \notin [-1,1]$, there are no intersection points.

Solutions to equation (6) (and therefore to equation (4)) provide the points where the sphere, centered at x_h and with radius $d_{h,i}$, intersects the circle $\mathcal{C}$ in Fig. 1. Those points are the extreme points of the feasible arcs: they define feasible intervals for the angle ω. Fig. 2 shows some possible intersections between the spherical shell centered in x_h (having minimum radius $\underline{d}_{h,i}$ and maximum radius $\overline{d}_{h,i}$) with the dashed circle $\mathcal{C}$.

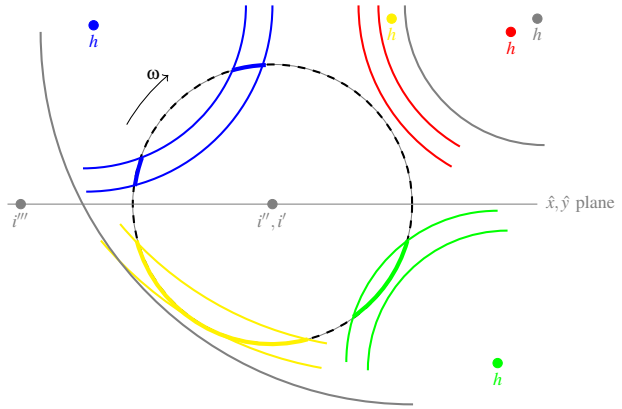


Fig. 2. Possible intersections between the spherical shell related to the distance $d_{h,i}$ and the circle of candidate positions related to $d_{i',i}$ and $d_{i'',i}$.

Managing different scenarios

The feasible positions for the current atom i can be obtained by intersecting the two arcs (computed by using the discretization distances, in bold in Fig. 1) and several spherical shells, each of them defined by considering a pruning distance between $h < i$ and i. In order to perform this intersection, the following two equations need to be solved

$$A\cos\omega + B\sin\omega = \Delta - \underline{d}^2_{h,i}, \tag{10}$$

$$A\cos\omega + B\sin\omega = \Delta - \overline{d}^2_{h,i}, \tag{11}$$

for every pruning distance $d_{h,i}$. There are three situations that can occur while performing the intersections (i.e. while solving equations (10) and (11)).

Both equations have no solutions: If both equations (10) and (11) have no solutions, then the entire candidate circle is either completely valid, or completely invalid. If we can find at least one value for ω such that

$$\underline{d}^2_{h,i} < \|x_h - x_i(\omega)\|^2 < \overline{d}^2_{h,i},$$

then the entire circle $\mathcal{C}$ is feasible w.r.t. the distance $d_{h,i}$. If not, it is sufficient to verify whether one of these 2 equations is satisfied

$$\max_{\omega\in[0,2\pi]} \|x_h - x_i(\omega)\|^2 < \underline{d}^2_{h,i},$$

$$\min_{\omega\in[0,2\pi]} \|x_h - x_i(\omega)\|^2 > \overline{d}^2_{h,i},$$

for stating that the entire circle is infeasible.

Only one equation has solutions: Let us suppose that only equation (10) has solutions. In this case, the resulting intersection is an interval $[\underline{\omega}, \overline{\omega}]$ whose extreme points are the solutions of equation (10). In order to find the right orientation of the arc on the circle $\mathcal{C}$, we define the function

$$F(\omega) = \|x_h - x_i(\omega)\|^2 = \Delta - A\cos\omega - B\sin\omega,$$

and we consider its derivative

$$F'(\omega) = A\sin\omega - B\cos\omega. \tag{12}$$

The orientation at an extreme point (solution of (10)) is the one for which $F(\omega)$ increases, and this information is given by (12) evaluated in such an extreme point. Notice that we might need to add 2π to one of the extreme points in order to have $\overline{\omega} > \underline{\omega}$. The analysis in the case in which only equation (11) has solutions is analogous.

Let $[\underline{\omega^*}, \overline{\omega^*}]$ be the obtained interval for ω. If this interval has an empty intersection with the two initial arcs in $\mathcal{C}$, then there are no feasible positions, and the current branch of the search domain can be pruned. If this intersection is instead non-empty, then the result provides the interval for ω that is feasible w.r.t the discretization distances, as well as the pruning distance $d_{h,i}$. Notice that, when more than one pruning distance is available, the same procedure can be repeated as many times as the number of available pruning distances.

Both equations have solutions: When both equations (10) and (11) have solutions, we obtain four values for ω: two from equation (10) and other two from equation (11). Two intervals can be therefore defined for ω, related to two arcs in $\mathcal{C}$. Both arcs need to be intersected with the initial arcs. The procedure to apply is analogous to the one presented for the previous case.

iBP and the new adaptive scheme

After considering all pruning distances, after performing all intersections, the final result provides a list of arcs on $\mathcal{C}$ that are feasible with all the distances that can be verified at the current layer of the *i*BP tree. All positions that can be taken from these arcs are feasible at the current layer: all of them generate a new branch and may serve as a reference for computing new candidate positions on deeper layers of the tree. In order to integrate the *i*BP algorithm with this adaptive scheme, there are two main changes to be performed on Alg. 1. On line 8 and 11, the adaptive scheme needs to be invoked for taking into consideration the information about the pruning distances. Moreover, line 22 needs to be removed, because this verification is not necessary anymore (unless other pruning devices rather than DDF are employed).

IV. Computational Experiments

Experiments of Nuclear Magnetic Resonance (NMR) [10] are able to provide estimates of some relative distances between pairs of atoms of a molecule. We present in this section some computational experiments on artificially generated NMR instances, where we compare the standard *i*BP algorithm to the new *i*BP integrated with our adaptive scheme for the generation of feasible atomic positions (accordingly to all available distances at the current tree layer). In this work, we do not consider real NMR data because the experiments here presented have the only aim of showing the advantages in using this new adaptive scheme. Later on, this scheme will be integrated in a more general framework capable of dealing with real NMR data. All codes were written in C programming language and all the experiments were carried out on an Intel Core 2 Duo @ 2.4 GHz with 2GB RAM, running Mac OS

Instance			*i*BP w/out adaptive scheme			*i*BP with adaptive scheme		
name	$\|V\|$	$\|E\|$	D	*i*BP calls	Time	D	*i*BP calls	Time
1niz	68	328	5	7668930	9.93	5	105543	0.15
2jnr	96	443	5	17410	0.02	5	16989	0.02
2pv6	110	558	7	174651	0.27	5	181020	0.24
1zec	122	622	6	1194478	1.92	5	932428	1.53
2mla	130	681	5	323354	0.54	5	136547	0.25
2me1	135	687	6	2813983	4.30	5	1415331	2.35
2me4	135	681	5	1533970	2.36	5	249096	0.40
1dsk	140	733	6	3746764	5.34	6	1091745	1.52

TABLE I
EXPERIMENTS ON OUR ARTIFICIALLY GENERATED NMR INSTANCES.

X. The codes have been compiled by the GNU C compiler v.4.0.1 with the `-O3` flag.

The instances that we consider in the experiments have been generated as follows. We consider a subset of proteins from the Protein Data Bank (PDB) [1] that are related to human immunodeficiency. Together with the coordinates of the atoms available on the PDB, we suppose having the chemical structure of the protein, i.e. information about bond lengths and angles. Once the coordinates are loaded from the PDB files, we compute all distances between atom pairs belonging to the protein backbone, and we add a distance in our instances if the computed distance is between:

1) two bonded atoms (considered as exact);
2) two atoms that are bonded to a common atom (considered as exact);
3) two atoms belonging to a quadruplet of bonded atoms forming a torsion angle (considered as an interval);
4) two hydrogen atoms (considered as an interval, if the distance belongs the interval [2.5,5] Å).

We remark that the first 3 items are related to the chemical structure of the molecule; only the last item concerns distances that simulate NMR data. The distances that are derived from the information mentioned in item 3 are generally intervals; however, when one of the possible torsion angles is related to the peptide bond (that connects pairs of consecutive amino acids), the distance is considered as exact, because the peptide bond forces all atoms to lie on the same plane. Interval distances coming from torsion angles are computed so that all possible values for the torsion angle are allowed. The interval distances related to item 4 have instead length equal to 2Å, and their bounds were generated so that the *true* distance is randomly placed inside the interval. After the computation of the distance information, the atoms in every instance have been reordered by considering the discretization order published in [11], which is valid for every protein backbone.

In Table I we compare the performance of the previous version of *i*BP [8] with our new one, where the adaptive branching scheme presented in Section III is implemented. For each instance, we report the label of the corresponding file on the PDB, the total number $|V|$ of atoms and the number $|E|$ of available distances. Moreover, for each *i*BP version, we report the number D of samples to be taken from each candidate arc, the number of *i*BP calls and the CPU time in seconds, that are necessary to find one solution. In the DDF pruning device (equation (3)), the used tolerance ε is 10^{-3}.

The D values in Table I are actually the smallest ones for which *i*BP could find at least one solution in a given time limit (10 seconds in these experiments). When using our adaptive branching scheme, the D value never increased and it was reduced in some cases. This was expected because our adaptive scheme is able to guide the sample points in the feasible regions of the candidate arcs. Even if the computation of the intersections may increase the computational cost for single *i*BP recursive calls, the overall CPU time for each experiment is lower when the adaptive scheme is employed. This is due to the fact that, when the branching phase in *i*BP is adaptive, only feasible coordinates are generated: there are no useless computations (i.e. computed positions that are immediately discarded).

V. CONCLUSIONS

We proposed a new adaptive branching scheme that was integrated in the *i*BP algorithm to solve discretizable MDGPs with interval data. When interval data are used in the discretization process, candidate positions for the current atom are generally represented by two candidate arcs. By exploiting the interval pruning distances that can be verified at the current layer, we can guide the branching phase of the *i*BP algorithm to take samples only on the feasible regions of the candidate arcs.

As it was assessed by our computational experiments, this approach improves the overall performances of the *i*BP algorithm, thereby improving its robustness. Using the intersections of the spherical shells defined by the pruning distances with the candidate arcs provided by the discretization, we avoid the generation of useless samples in infeasible portions of the candidate arcs.

However, it is important to mention that, as in the previous *i*BP version, the presented scheme does not guarantee that the chosen sample positions can lead to feasible positions at further layers: our scheme ensures the feasibility only up to the current layer. Predicting the compatibility of sample positions with the atoms that *follow* the current one is a topic of future research.

VI. ACKNOWLEDGMENTS

We are thankful to Brittany Region (France) which funded a 1-year postdoc for DG at IRISA, University of Rennes 1

(stratégie d'attractivité durable). This work is partially supported by the ANR project ANR-10-BINF-03-01 "Bip:Bip". CL is also thankful to FAPESP and CNPq for financial support.

REFERENCES

[1] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, *The Protein Data Bank*, Nucleic Acid Research **28**, 235–242, 2000.

[2] V. Costa, A. Mucherino, C. Lavor, A. Cassioli, L.M. Carvalho, N. Maculan, *Discretization Orders for Protein Side Chains*, to appear in Journal of Global Optimization, 2014.

[3] G.M. Crippen and T.F. Havel, *Distance Geometry and Molecular Conformation*, John Wiley & Sons, New York, 1988.

[4] D.S. Gonçalves, A. Mucherino, *Discretization Orders and Efficient Computation of Cartesian Coordinates for Distance Geometry*, to appear in Optimization Letters, 2014.

[5] T.F. Havel, *Distance Geometry*, D.M. Grant and R.K. Harris (Eds.), Encyclopedia of Nuclear Magnetic Resonance, Wiley, New York, 1701-1710, 1995.

[6] C. Lavor, J. Lee, A. Lee-St.John, L. Liberti, A. Mucherino, M. Sviridenko, *Discretization Orders for Distance Geometry Problems*, Optimization Letters **6**(4), 783–796, 2012.

[7] C. Lavor, L. Liberti, N. Maculan, A. Mucherino, *The Discretizable Molecular Distance Geometry Problem*, Computational Optimization and Applications **52**, 115–146, 2012.

[8] C. Lavor, L. Liberti, A. Mucherino, *The interval Branch-and-Prune Algorithm for the Discretizable Molecular Distance Geometry Problem with Inexact Distances*, Journal of Global Optimization **56**(3), 855–871, 2013.

[9] L. Liberti, C. Lavor, N. Maculan, A. Mucherino, *Euclidean Distance Geometry and Applications*, SIAM Review **56**(1), 3–69, 2014.

[10] T.E. Malliavin, A. Mucherino, M. Nilges, *Distance Geometry in Structural Biology: New Perspectives*. In: [13], 329–350, 2013.

[11] A. Mucherino, *On the Identification of Discretization Orders for Distance Geometry with Intervals*, Lecture Notes in Computer Science **8085**, F. Nielsen and F. Barbaresco (Eds.), Proceedings of Geometric Science of Information (GSI13), Paris, France, 231–238, 2013.

[12] A. Mucherino, C. Lavor, L. Liberti, *The Discretizable Distance Geometry Problem*, Optimization Letters **6**(8), 1671–1686, 2012.

[13] A. Mucherino, C. Lavor, L. Liberti, N. Maculan (Eds.), *Distance Geometry: Theory, Methods and Applications*, Springer, 2013.

[14] H.B. Thompson, *Calculation of Cartesian Coordinates and their Derivatives from Internal Molecular Coordinates*, Journal of Chemical Physics **47**, 3407, 1967.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 465–470
DOI: 10.15439/2014F99
ACSIS, Vol. 2

Higher-Order Quantum-Inspired Genetic Algorithms

Robert Nowotniak, Jacek Kucharski
Institute of Applied Computer Science
Lodz University of Technology
18/22 Stefanowskiego St., 90-924 Lodz, Poland
Email: {rnowotniak,jkuchars}@kis.p.lodz.pl

***Abstract*—This paper presents a theory and an empirical evaluation of Higher-Order Quantum-Inspired Genetic Algorithms. Fundamental notions of the theory have been introduced, and a novel Order-2 Quantum-Inspired Genetic Algorithm (QIGA2) has been developed. Contrary to all QIGA algorithms which represent quantum genes as independent qubits, in higher-order QIGAs quantum registers are used to represent genes strings, which allows modelling of genes relations using quantum phenomena. Performance comparison has been conducted on a benchmark of 20 deceptive combinatorial optimization problems. It has been presented that using higher quantum orders is beneficial for genetic algorithm efficiency, and the new QIGA2 algorithm outperforms the old QIGA algorithm tuned in highly compute-intensive metaoptimization process.**

I. Introdution

RESEARCH on quantum-inspired computational intelligence techniques was started by Narayann[1] in 1996, and the first proposal of Quantum-Inspired Genetic Algorithm (QIGA1) has been presented by Han and Kim in [2]. Quantum-Inspired Genetic Algorithms belong to a new class of artificial intelligence techniques, drawing inspiration from both evolutionary[3] and quantum[4] computing. Current literature on the subject consists of about a few hundreds scientific papers. Only a few papers attempt to theoretically analyse the properties of that class of algorithms. Among those there are i.a. [22,28], which has been emphasized in conclusions of recent comprehensive surveys [18,29].

In QIGA algorithms, representation and genetic operators are based on computationally useful aspects of both biological evolution and unitary evolution of quantum systems. QIGA algorithms use quantum mechanics concepts including qubits and superposition of states. QIGA algorithms have been successfully applied to a broad range of search and optimization problems[5,6,7]. The algorithms have demonstrated their particular efficacy for solving complex optimization problems. Recent years have witnessed successful applications of Quantum-Inspired Genetic Algorithms in a variety of fields, including image processing[8,9,10], flow shop scheduling[11,12], thermal unit commitment[13,14], power system optimization[15,16], localization of mobile robots[17] and many others.

For a current and comprehensive survey of Quantum-Inspired Genetic Algorithms and the necessary background of Quantum Computing and Quantum-Inspired Computational Intelligence techniques, the reader is referred to [1,2,18,29].

This work was supported in part by PL-Grid Infrastructure

This paper is structured as follows. In Section 1, an introductory background and the most important references for the subject field have been given. In Section 2, the theory of Higher-Order Quantum-Inspired Genetic Algorithms has been presented. In Section 3, details of the original Order-2 Quantum-Inspired Genetic Algorithm have been provided. In Section 4, experimental results have been provided and evaluated. In Section 5, the article has been briefly summarized, final conclusions have been drawn, and also possible directions for future research have been suggested.

II. Theory of Higher-Order Quantum-Inspired Genetic Algorithms

Let $N \in \mathbb{N}^+$ denote the length of chromosomes in the algorithm (i.e. problem size), X – search space of the optimization problem, Q – quantum population (a set of quantum individuals in QIGA algorithm), and P – classical population (a set of elements in X space). Let us assume that each individual in the algorithm consists of a single quantum chromosome.

We introduce the following new notions.

Definition 1 (quantum order $r \in \mathbb{N}^+$): the size of the biggest quantum register used in the algorithm.

$$1 \leq r \leq N \tag{1}$$

We say an algorithm is Order-r, if r is the size of the biggest quantum register used in that algorithm. All Quantum-Inspired Genetic Algorithms that use independent qubits to represent binary genes are Order-1. All existing algorithms, presented in the literature so far are Order-1 in terms of this theory. To simplify the further discussion, let us assume that all quantum registers used in the algorithm have the same size.

Definition 2 (relative quantum order w): – the ratio of quantum order r to quantum chromosomes length N (problem size) in the algorithm.

$$w = \frac{r}{N} \in (0, 1] \tag{2}$$

If a certain QIGA algorithm uses a representation of solutions based on 100 independent qubits (binary quantum genes), the relative quantum order for that algorithm is $w = \frac{1}{100}$. If the size of a problem (the number of binary variables) is $N = 60$, and the representation is based on 3-qubit registers, then the relative quantum range is $w = \frac{3}{60} = 0.05$ etc.

The algorithms characterised by $w = 1$ are "true" quantum algorithms, where a single quantum register contains all the

978-83-60810-58-3/$25.00

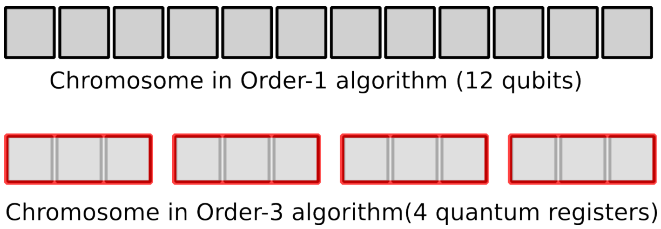


Fig. 1. Examples of chromosomes of length $N = 12$ for Order-1 and Order-3 algorithms. Consecutive genes are joined to r-qubit quantum registers. In Order-1 algorithm, the chromosome consists of 12 independent qubits, each one is a unit vector in 2-dimensional space. In Order-3 algorithm, the chromosome consists of 4 quantum registers, each one is a unit vector in $2^3 = 8$-dimensional space.

binary variables. For $w = 1$, when the number of binary variables (the size of the problem N) grows linearly, the cost of simulation grows exponentially (which corresponds to a simulation of a real quantum computer).

Definition 3 (quantum factor $\lambda \in [0, 1]$): For a given algorithm, the quantum factor is defined as a ratio of the dimension of space in a given class of algorithms to the dimension of space of the full quantum register of N qubits. Additionally, if there are no quantum elements in the algorithm (e.g. a simple genetic algorithm SGA[30], operating in a discrete space of binary strings), then $\lambda = 0$.
Thus, the numerical value of the factor is expressed as:

$$\lambda = \frac{2^r \frac{N}{r}}{2^N} = \frac{2^r}{w2^N} \tag{3}$$

where r is the quantum order of an algorithm and N is the problem size. The 2^r in the numerator of the above formula corresponds to the dimension of the state space in the r-qubit quantum register (the biggest quantum register used in an algorithm of that class). Such quantum register codes a 2^r-point probability distribution (it shows the probability of choosing one from 2^r elements of a solution space X). 2^N corresponds to the dimension of the state space of a quantum register containing all N qubits.

In Order-1 algorithms, chromosomes consist of N independent qubits. According to the Quantum Computing theory the state of each qubit is described by a unit vector in a 2-dimensional space ($|q\rangle = [\alpha \ \ \beta]^T$), so the space dimension for the chromosomes in such algorithms is $2^r \frac{N}{r} = 2N$.

In Order-2 algorithms, chromosomes consist of $\frac{N}{2}$ size-2 quantum registers. The state for each register is described by a unit vector in a 4-dimensional space ($|q\rangle = [\alpha_0 \ \alpha_1 \ \alpha_2 \ \alpha_3]^T$). Therefore, the dimension of space for the chromosomes in such algorithms is also $2^2 \frac{N}{2} = 2N$. However, in Order-1 algorithms only one qubit coordinate might be independently modified (one degree of freedom), while in Order-2 algorithms the same can be done with 3 out of 4 coordinates of the 2-qubit quantum register state. Consequently, it allows for modelling of relations between two neighbouring genes joined in a common register.

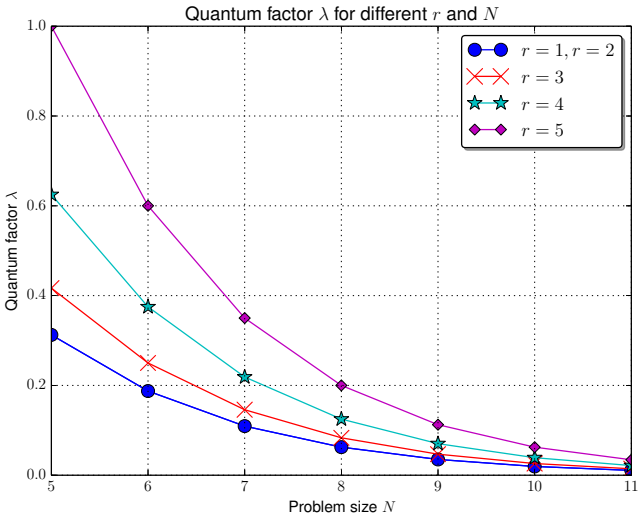


Fig. 2. Quantum factor λ for different problem size N and different quantum orders $r \in \{1, 2, 3, 4, 5\}$

For even Higher-Order algorithms ($r \geq 3$), simulating quantum element makes the algorithm exponential computational complexity. Relationship between quantum factor λ, quantum order r for growing problem size N has been presented in Figure 2.

It should be noted that for $r = 1$ (all regular Order-1 QIGA algorithms):

$$\lambda = \frac{2^1 \frac{N}{1}}{2^N} = \frac{2 \cdot N}{2^N}$$

Thus, for example, in an algorithm coding solutions in the form of 10-element strings of independent qubits, $\lambda = \frac{20}{2^{10}} \approx 0.02$. It means that the size of space in such algorithm comprises 2% of the full quantum register state space, which would include 10 binary variables. Together with the increase of size of a problem N and for a constant quantum order $r = 1$, the quantum factor decreases exponentially and becomes $\lambda < 10^{-10}$ for $N = 50$.

For that reason, for a constant quantum order $r = 1$ (QIGA Order-I quantum-inspired algorithms) and for an increasing size of a problem N, the quantum factor λ has a limit that equals zero:

$$\lim_{\substack{r=1 \\ N \to \infty}} \lambda = \lim_{\substack{r=1 \\ N \to \infty}} \frac{2 \cdot N}{2^N} = 0$$

However, for $r = N$ (typical quantum algorithms)

$$\lambda = \frac{2^N \frac{N}{N}}{2^N} = \frac{2^N}{2^N} = 1 \tag{4}$$

For $\lambda = 1$, when the number of variables (the size of a problem N) grows linearly, the cost of simulation grows exponentially (which corresponds to a full simulation of a real quantum computer).
Thus, algorithms can be classified according to quantum factor λ value as follows:

Algorithm 1 Order-2 Quantum-Inspired Genetic Algorithm

1: $t \leftarrow 0$
2: Initialize quantum population $Q(0)$
3: **while** $t \leq t_{max}$ **do**
4: $t \leftarrow t + 1$
5: Generate $P(t)$ by observing quantum pop. $Q(t-1)$
6: Evaluate classical population $P(t)$
7: Update $Q(t)$
8: Save best classical individual to b
9: **end while**

1) $\lambda = 0$ – a classical algorithm without any quantum elements, operating in a discrete finite space (e.x. SGA[30] operating in finite discrete binary strings space).
2) $\lambda \in (0, 1)$ – a quantum-inspired algorithm, like QIGA1 (order $r = 1$), or higher-order algorithm.
3) $\lambda = 1$ – a "true" quantum algorithm which requires either a real quantum level hardware, or an exponential complexity simulation on classical computer.

Order-r Quantum-Inspired Genetic Algorithms are capable of modelling relations between separate genes which are joined into the same quantum register of size r. This allows the algorithm to work better for deceptive combinatorial optimization problems and to better solve strong epistasis in deceptive problems. This is presented empirically the next sections of the paper.

III. Order-2 Quantum-Inspired Genetic Algorithm

In this section, a novel Order-2 Quantum-Inspired Genetic Algorithm (QIGA2) has been presented. The algorithm has been developed based on the theory of higher-order quantum-inspired algorithms presented in the previous section.

Pseudocode of the algorithm has been presented in Algorithm 1, and in general it is very similar to a typical evolutionary algorithm scheme. The general principle of operation of the algorithm is very similar to the initial QIGA 1 algorithm, but instead of independent qubits modelling successive binary genes, the QIGA 2 algorithm uses 2-qubit quantum registers representing successive pairs of genes.

In each generation of the algorithm a classic population P (a set of elements from the solution space X) is sampled through observation of quantum states of the quantum population Q i.e. $|P|$-times repeated sampling of the space X according to probability distributions stored in Q. The classical population P is then evaluated exactly as in a typical evolutionary algorithm. The quantum population Q, however, is updated in consecutive generations in such a way that it increases the probability of sampling the best solution b neighbourhood, which has been recorded in previous generations of P.

The key new elements distinguishing QIGA2 from the previous Order-1 algorithms are the modified method of representing solutions and the new genetic operators working in a space of a higher dimension and described by 4×4 unitary matrices in the quantum-mechanic sense. Both original elements have been described in the next subsections respectively.

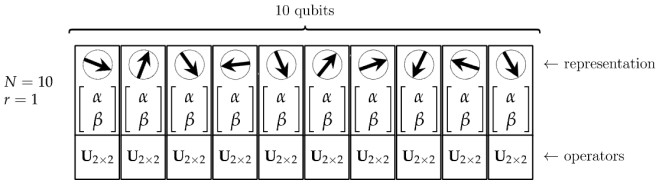


Fig. 3. In QIGA1, representation is based on isolated qubits / binary quantum genes

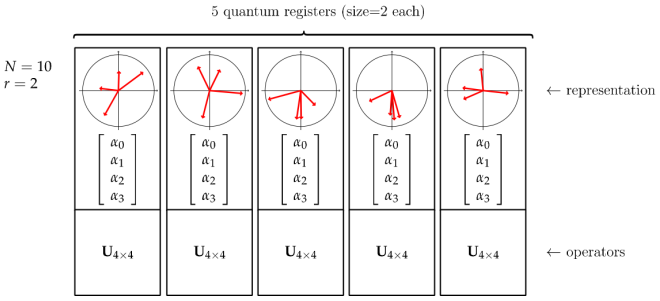


Fig. 4. In QIGA2, quantum registers are used to represent pairs of genes

A. Representation of solutions in QIGA2

The fundamental difference between the already existing QIGA1 and QIGA2 algorithms lies in the way they represent solutions. In QIGA1 algorithms, quantum genes are modelled with qubits i.e. two-level quantum systems $|q\rangle = \alpha|0\rangle + \beta|1\rangle = [\alpha \quad \beta]^T$ which are able to code two-point probability distributions. It corresponds to a possibility of each gene to have a value 0 or 1 with a probability of $|\alpha|^2$ and $|\beta|^2$ accordingly. It has been depicted in Figure 3.

In the authors' QIGA2 algorithm, the representation of solutions is based on using the adjacent 2-qubit quantum registers. For that purpose the adjacent genes are consecutively paired. The corresponding 2-qubit registers $|q\rangle = [\alpha_0 \quad \alpha_1 \quad \alpha_2 \quad \alpha_3]^T$ code 4-point probability distributions. So, in a single quantum register 4 values of probability $|\alpha_0|^2$, $|\alpha_1|^2$, $|\alpha_2|^2$, $|\alpha_3|^2$ are recorded. These are probabilities of having a value of 00, 01, 10 and 11 for each given pair of genes accordingly. It is presented in Figure 4. Similarly to QIGA1 algorithms, the proposed QIGA2 uses only the real parts of probability amplitudes. It ignores the imaginary part of amplitudes $\alpha_0, \ldots, \alpha_3$.

At the stage of the $Q(0)$ base population initialization, all genes can be given the value of $q_{ij} = [\frac{1}{2} \quad \frac{1}{2} \quad \frac{1}{2} \quad \frac{1}{2}]^T$, which corresponds to a situation when the algorithm samples the entire solution space X with the same probability.

B. Order-2 quantum genetic operators

The second original element of the QIGA2 algorithm is the use of genetic operators. In the QIGA1 algorithm genetic operators are created by unitary 2×2 quantum gates (thanks to the limiting of the amplitudes to a set $\mathbb{R}$, they become just matrices of a normalised state vector rotation on a plane). By contrast, in the QIGA2 algorithm the genetic operators can be described by 4×4 quantum gates in the quantum-mechanical sense.

Algorithm 2 Observation of genes pair in QIGA2

Require: $q_{ij} = [\alpha_0 \quad \alpha_1 \quad \alpha_2 \quad \alpha_3]^T$ – quantum register of 2 qubits

```
r ← uniformly random number from [0,1]
if r < |α0|^2 then
    p ← 00
else if r < |α0|^2 + |α1|^2 then
    p ← 01
else if r < |α0|^2 + |α1|^2 + |α2|^2 then
    p ← 10
else
    p ← 11
end if
```

Algorithm 3 Update of quantum genes states in QIGA2

```
for i in 0, ..., |Q| − 1 do
    for j in 0, ..., N/2 do
        q' = [0 0 0 0]^T
        bestamp ← j-th pair of binary genes in b as decimal
        sum ← 0
        for amp in {0, 1, 2, 3} do
            if amp ≠ bestamp then
                q'[amp] ← μ · q_ij
                sum ← sum + (q'[amp])^2
            end if
        end for
        q'[bestamp] ← sqrt(1 − sum)
        q_ij ← q'
    end for
end for
```

The pseudocode for the operation of measuring the states of a 2-qubit quantum register $q_{ij} = \alpha_0|00\rangle + \alpha_1|01\rangle + \alpha_2|10\rangle + \alpha_3|11\rangle = [\alpha_0 \quad \alpha_1 \quad \alpha_2 \quad \alpha_3]^T$ coding a pair of classic binary genes is presented in the Algorithm 2. The observation function returns strings of binary genes 00, 01, 10 and 11 with a probability of $|\alpha_0|^2$, $|\alpha_1|^2$, $|\alpha_2|^2$ oraz $|\alpha_3|^2$ respectively.

Algorithm 3 presents the pseudocode of the proposed new genetic operator (observing the state of a 2-qubit quantum gene) in QIGA2. Index i of the main operator's loop iterates through all the individuals in the quantum population $q_0, \ldots, q_{|Q|-1}$. Index j iterates through all the consecutive *pairs* of genes $j \in \{0, 1, \ldots N/2\}$ of a given quantum individual q_i. Within these loops, a new state q' of the quantum gene pair number j of the character q_i is calculated.

The update is performed in the following manner: If the amplitude $\alpha_{amp}(amp \in \{0, 1, 2, 3\})$ does not correspond to a j-th pair of bits of the currently best found individual b, the amplitude is decreased (amplitude contraction) according to the rule: $q_{ij}[amp]' = \mu \cdot q_{ij}[amp]$, where $\mu \in (0, 1)$ is the algorithm's parameter. The amplitude of a pair of bits on position j in the best individual b is modified to preserve the normalization condition of the state vector (i.e. unit sum of probabilities $\sum_{amp=0}^{3} |\alpha_{amp}|^2 = 1$).

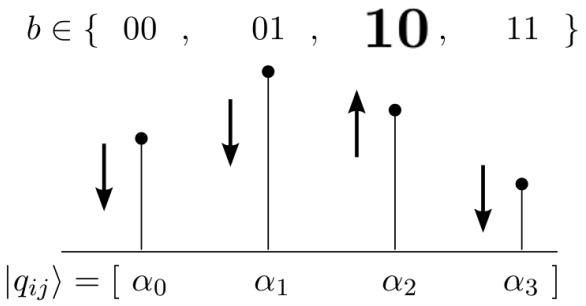


Fig. 5. The new quantum genetic operator idea in QIGA2

Based on empirical experiments it has been established that the the best efficacy of an algorithm is achieved for the parameter value $\mu \approx 0.99$. In order to further increase the efficacy, the value of the parameter μ in the QIGA2 algorithm might be subject to metaoptimalisation (similarly to [19,20,21,31]).

The way the new operator works is illustrated in Figure 5. The vertical bars represent probability amplitudes $|\alpha_0|^2, |\alpha_1|^2, |\alpha_2|^2, |\alpha_3|^2$. If on the position $j \in \{0, 1, \ldots, \frac{N}{2}\}$ of the individual b there is a pair of bits 10, all the amplitudes get contracted by the factor of μ, except for α_2 which will increase. If on the position j of the individual b there is a pair of bits 00, all the amplitudes get contracted by the factor of μ, except for α_0, which will increase etc. Therefore, the only amplitude that increases is the one that corresponds to the j-th pair of bits in the best individual b. This makes the algorithm converge to the best individual b gradually, but also doing global exploration of the search space X.

Simplicity is an unquestionable advantage of the QIGA2 algorithm. It is not only simpler than QIGA1, but also less complicated than its later modified variants, whose authors also tried improve on the efficacy of the original algorithm. It should be noted that in QIGA2 the use of the Lookup Table (used in the original Han's QIGA1 algorithm[2]) has been eliminated completely.

IV. Numerical experiments

For empirical comparison of the algorithms performance, there was used a benchmark consisting of a broad set of 20 recognized combinatorial optimization problems of different sizes $N \in \{48, 90, \ldots, 1000\}$, encoded in the form of the NP-complete SAT. Objective of the combinatorial optimization process was to find a binary string that have maximum fitness value. The benchmark has been taken from [32], and all details about the test functions are available there.

The compared algorithms were SGA[30], the original QIGA1[2], the QIGA1 tuned in meta-optimization process[31] and the authors' QIGA2. Numerous publications to date present that QIGA1 is more effective than other modern stochastic search methods and hence its comparison to other algorithms has been omitted in this paper as it has been assumed to be superior to other newest algorithms.

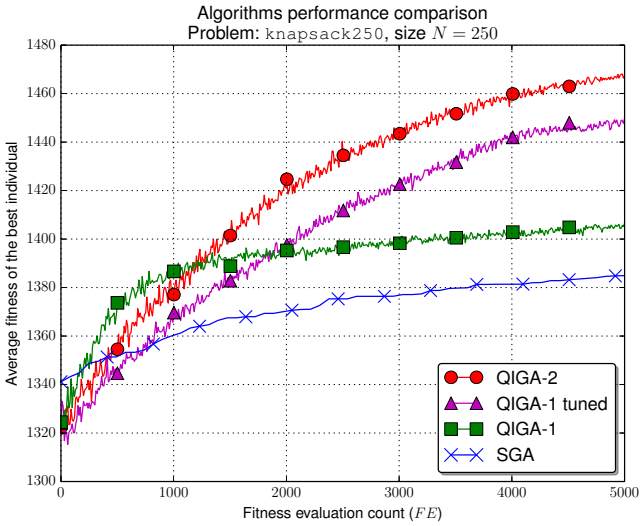


Fig. 6. Detailed comparison of the algorithms for a selected problem knapsack, size $N = 250$

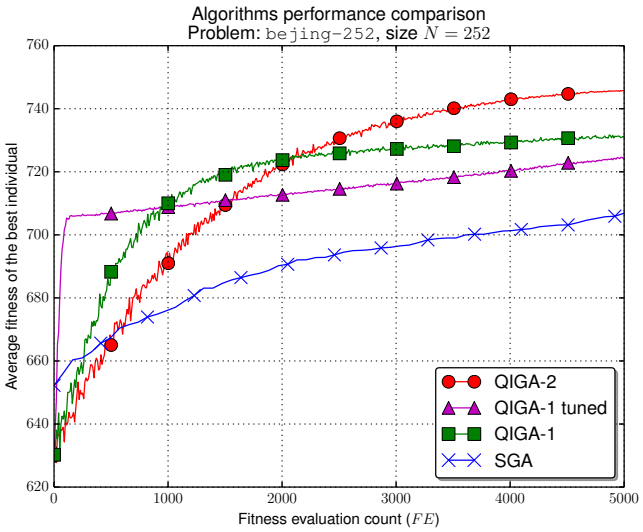


Fig. 7. Detailed comparison of the algorithms for a selected problem bejing, size $N = 252$

The classic SGA algorithm was run with its typical parameters values taken from [30]: the population size was set to 100 individuals (binary solutions), evolving for 50 generations. Thus, the total number of fitness evaluations was equal in all algorithms, and the stopping criterion was maximum number of fitness evaluations $MaxFE = 5000$. In SGA, single point crossover operator with probability $P_c = 0.65$ and mutation operator with probability $P_m = 0.05$ were used. The selection was based on the roulette wheel method. Implementation of SGE algorithm was taken from the external PyEvolve library [26] The parameters for the original QIGA1 algorithm were taken from [2] as were the parameters for the tuned QIGA1, where the only changed parameters were those that had been meta-optimized. The QIGA2 algorithm was run with the value of the parameter $\mu = 0.9918$. For each of the test problem, each algorithm was run 50 times.

As a means for evaluating the algorithms efficacy the authors used the fitness value of the best individual after the number of generations which reached the 5000th call of the fitness evaluation function. Because of stochastic nature of evolutionary algorithms, that value was later averaged for 50 runs of a given algorithm.

In Table 1, the results for each algorithm are presented. **In 17 out of 20 test problems (85%), the authors' QIGA2 algorithm presented on average a better solution than both the original and the tuned QIGA1 algorithm.** Table 2 presents a ranking of the compared algorithms ordered according to the number of test problems for which a given algorithm achieved the best result comparing to algorithms. Figures 6-8 present a detailed comparison of the algorithms performance for three selected test problems of size $N = 250$, $N = 1000$ and $N = 252$. The graph shows the mean value of the best solution found by each of the algorithms versus number of calls of the individual fitness evaluation function.

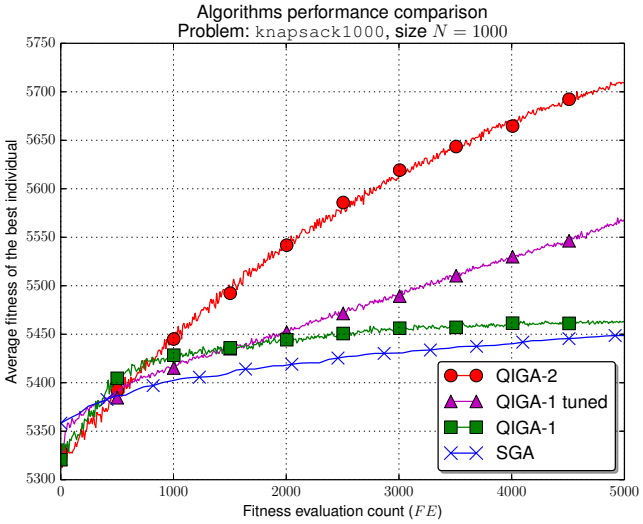


Fig. 8. Detailed comparison of the algorithms for a selected problem knapsack, size $N = 1000$

The presented data is averaged for 50 runs of each algorithm.

Thanks to the simplification of the algorithm and, specifically, owing to the elimination of the LookupTable, also **the implementation of QIGA2 algorithm is 15-30% faster than that of the QIGA1** (the algorithms were implemented in the same programming languages, with the same compiler versions and on the same hardware platforms).

V. Conclusions

In this paper, fundamentals of Higher-Order Quantum-Inspired Genetic Algorithms have been presented. The authors' original QIGA2 algorithm has been created on the basis of this theory. The paper introduces a new way of representing solutions using adjacent quantum registers and

TABLE I
ALGORITHMS EFFICACY COMPARISON FOR VARIOUS PROBLEMS OF DIFFERENT SIZE $N \in \{48, \ldots, 1000\}$

Problem	Size N	SGA	QIGA-1	QIGA-1tuned	QIGA-2
anomaly	48	251.4	252.55	254.65	**255.25**
sat	90	284.9	289.2	293.2	**293.7**
jnh	100	826.15	831.05	**839.05**	836.05
knapsack	100	577.709	578.812	592.819	**596.476**
sat	100	408.6	413.6	418.6	**419.7**
bejing	125	297.35	302.1	305.35	**306.2**
sat-uuf	225	886.75	898.25	**921.65**	921.5
knapsack	250	1387.916	1406.528	1449.905	**1467.407**
sat1	250	981.45	995.15	1021.2	**1023.1**
sat2	250	982.95	994.6	1019.1	**1020.6**
sat3	250	984.2	994.3	**1021.3**	1019.7
bejing	252	709.85	731.0	724.4	**745.75**
parity	317	1141.65	1158.2	1179.35	**1180.75**
knapsack	400	2209.925	2222.160	2284.969	**2334.494**
knapsack	500	2803.266	2812.740	2869.774	**2929.469**
bejing	590	1263.8	1343.15	1284.0	**1353.2**
lran	600	2310.9	2330.35	2386.8	**2398.95**
bejing	708	1510.65	1605.9	1523.15	**1611.55**
knapsack	1000	5451.656	5462.718	5568.234	**5709.116**
lran	1000	3819.65	3848.4	3918.5	**3937.3**

TABLE II
RANKING OF THE COMPARED ALGORITHMS

Rank	Algorithm	No. of Best Solutions
1	QIGA2	17
2	QIGA-1 tuned	3
3	QIGA-1	0
4	SGA	0

a new genetic operator working in the space of a higher dimension in quantum-mechanical sense. Based on empirical data gathered from 20 varied deceptive test problems of diverse sizes $N \in \{48, \ldots, 1000\}$, it has been shown that the authors' QIGA2 algorithm achieves a better performance than both the original and the tuned QIGA1 algorithms. Consequently, it shows that using quantum order $r = 2$ is a method for improving the performance of Quantum-Inspired Genetic Algorithms. Further investigations may include the application of the presented theory of Higher-Order Quantum-Inspired Genetic Algorithms to a very important field of problems of numerical optimization.

REFERENCES

[1] A. Narayanan and M. Moore, "Quantum-inspired genetic algorithms", *Proc. IEEE Evolutionary Computation*, 61-66 (1996).

[2] K. H. Han and J. H. Kim, "Genetic quantum algorithm and its application to combinatorial optimization problem", *Proc. Congress on Evolutionary Computation*, 1354-1360 (2000).

[3] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*, Springer, 1996.

[4] M. Nielsen and I. Chuang, *Quantum computation and quantum information*, Cambridge University Press, 2000.

[5] P. Jantos, D. Grzechca, and J. Rutkowski, "Evolutionary algorithms for global parametric fault diagnosis in analogue integrated circuits", *Bull. Pol. Ac.: Tech.* 60 (1), 133-142 (2012).

[6] A. Slowik, "Application of evolutionary algorithm to design minimal phase digital filters with non-standard amplitude characteristics and finite bit word length", *Bull. Pol. Ac.: Tech.* 59 (2), 125-135 (2011).

[7] L. Chomatek and M. Rudnicki, "Application of genetically evolved neural networks to dynamic terrain generation", *Bull. Pol. Ac.: Tech.* 59 (1), 3-8 (2011).

[8] Ł. Jopek, R. Nowotniak, M. Postolski, L. Babout, and M. Janaszewski, "Application of Quantum Genetic Algorithms in Feature Selection Problem", *Scientific Bulletin of Academy of Science and Technology, Automatics* 13(3), 1219-1231 (2009).

[9] H. Talbi, M. Batouche, and A. Draa, "A quantum-inspired genetic algorithm for multi-source affine image registration", *Image Analysis and Recognition*, Springer, 147-154 (2004).

[10] H. Talbi, M. Batouche, and A. Draa, "A Quantum-Inspired Evolutionary Algorithm for Multiobjective Image Segmentation", *International Journal of Mathematical, Physical and Engineering Sciences* 1, 109-114 (2007).

[11] L. Wang, H. Wu, F. Tang, and D.Z. Zheng, "A hybrid quantum-inspired genetic algorithm for flow shop scheduling", *Advances in Intelligent Computing*, Springer, 636-644 (2005).

[12] B.B. Li and L. Wang, "A hybrid quantum-inspired genetic algorithm for multiobjective flow shop scheduling", *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics* 37, 576-591 (2007).

[13] Y.W. Jeong, J.B. Park, J.R. Shin, and K.Y. Lee, "A thermal unit commitment approach using an improved quantum evolutionary algorithm", *Electric Power Components and Systems* 37, 770-786 (2009).

[14] T. Lau, C. Chung, K. Wong, T. Chung, and S. Ho, "Quantum-inspired evolutionary algorithm approach for unit commitment", *IEEE Trans. Power Systems* 24, 1503-1512 (2009).

[15] L. Su-Hua, W. Yao-Wu, P. Lei, and X. Xin-Yin, "Application of quantum-inspired evolutionary algorithm in reactive power optimization", *Relay* 33, 30-35 (2005).

[16] J. G. Vlachogiannis and K. Y. Lee, "Quantum-inspired evolutionary algorithm for real and reactive power dispatch", *IEEE Trans. Power Systems* 23, 1627-1636 (2003).

[17] S. Jeżewski, M. Łaski, and R. Nowotniak, "Comparison of Algorithms for Simultaneous Localization and Mapping Problem for Mobile Robot", *Scientific Bulletin of Academy of Science and Technology, Automatics* 14, 439-452 (2010).

[18] G. Zhang, "Quantum-inspired evolutionary algorithms: a survey and empirical study", *Journal of Heuristics*, 1-49 (2010).

[19] J.J. Grefenstette, "Optimization of control parameters for genetic algorithms", *IEEE Trans. Systems, Man and Cybernetics* 16, 122-128 (1986).

[20] R. Nowotniak and J. Kucharski, "Meta-optimization of Quantum-Inspired Evolutionary Algorithm", *Proc. XVII Int. Conf. on Information Technology Systems*, (2010).

[21] M.E.H. Pedersen, *Tuning & Simplifying Heuristical Optimization*, University of Southampton, School of Engineering Sciences, 2010.

[22] R. Nowotniak and J. Kucharski, "Building Blocks Propagation in Quantum-Inspired Genetic Algorithm", *Scientific Bulletin of Academy of Science and Technology, Automatics* 14, 795-810 (2010).

[23] S. Luke, *Essentials of metaheuristics*, lulu.com, 2009.

[24] D. E. Goldberg, *Genetic algorithms in search, optimization, and machine learning*, Addison-Wesley Professional, 1989.

[25] K.H. Han and J.H. Kim, "Quantum-inspired evolutionary algorithm for a class of combinatorial optimization", *IEEE Trans. Evolutionary Computation* 6, 580-593 (2002).

[26] C.S. Perone, "PyEvolve: a Python open-source framework for genetic algorithms", *ACM SIGEVOlution* 4, 12-20 (2009).

[27] R. Durrett, *Probability: Theory and Examples*, International Thomson Publishing Company, 1996.

[28] R. Nowotniak, J. Kucharski, Convergence analysis of Quantum-Inspired Evolutionary Algorithms based on Banach fixed point theorem, Proceedings of the 2012 FIMB PhD students conference

[29] Manju, A., and M. J. Nigam. "Applications of quantum inspired computational intelligence: a survey." Artificial Intelligence Review (2012): 1-78.

[30] Holland, John H. Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence. U Michigan Press, 1975.

[31] R. Nowotniak, J. Kucharski. "GPU-based tuning of quantum-inspired genetic algorithm for a combinatorial optimization problem." Bulletin of the Polish Academy of Sciences: Technical Sciences 60.2 (2012): 323-330.

[32] H. H. Holger, and T. Stützle. "SATLIB: An Online Resource for Research on SAT." Proceedings of Theory and Applications of Satisfiability Testing, 4th International Conference (SAT 2000).

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 471–478

DOI: 10.15439/2014F88
ACSIS, Vol. 2

Change-Point Detection in Binary Markov DNA Sequences by the Cross-Entropy Method

Tatiana Polushina
Department of Clinical Science,
Faculty of Medicine and Dentistry,
University of Bergen, 7804,
NO-5020 Bergen, Norway
Email: t.polushina@gmail.com

Georgy Sofronov
Department of Statistics
Faculty of Science
Macquarie University
Sydney NSW 2109 Australia
Email: georgy.sofronov@mq.edu.au

***Abstract*—A deoxyribonucleic acid (DNA) sequence can be represented as a sequence with 4 characters. If a particular property of the DNA is studied, for example, GC content, then it is possible to consider a binary sequence. In many cases, if the probabilistic properties of a segment differ from the neighbouring ones, this means that the segment can play a structural role. Therefore, DNA segmentation is given a special attention, and it is one of the most significant applications of change-point detection. Problems of this type also arise in a wide variety of areas, for example, seismology, industry (e.g., fault detection), biomedical signal processing, financial mathematics, speech and image processing. In this study, we have developed a Cross-Entropy algorithm for identifying change-points in binary sequences with first-order Markov dependence. We propose a statistical model for this problem and show effectiveness of our algorithm for synthetic and real datasets.**

I. Introduction

THE eukaryotic genomes are packaged into nucleosomes, composed of approximately 147 base pairs. There are 4 different bases: adenine (A), cytosine (C), guanine(G) and thymine (T). We can consider different approaches to base partition that depend on chemical and physical structure. One type of separation is pyrimidines (T and C) and purine (A and G). The second type of separation is keto (T and G) and amino (A and C) groups. In this paper, we consider groups of complemental bases: GC and AT pairs.

In this study, we are interested in finding regions that differ from neighbouring ones in GC level. It is well-known that genomic sequences are nonhomogeneous with respect to GC level, differences in GC proportion may be over scale of 100 kb to megabases. These long segments are called GC-content domains or isochores [1], [2]. Many studies propose that the differences of GC proportion appear as an outcome from a selection process [3]. It is well-known that an average GC proportion in chromatin organization and, hence, gene regulation is significant [4]. So GC proportion has been revealed to correlate with genomic properties such as DNA bendability and regulated replication.

In the last years, this topic has been investigated by many researchers [5], [6], [7]. This stimulates the elaboration of computational techniques that are applied to large-scale biological experimental data. Positive relationships have been discovered between GC level and recombination in humans, birds, and plants [8], [9], [10], [11]. Spencer *et al.* [5] have discovered that recombination proportions are too fast-evolving to have permanent meanings on base composition.

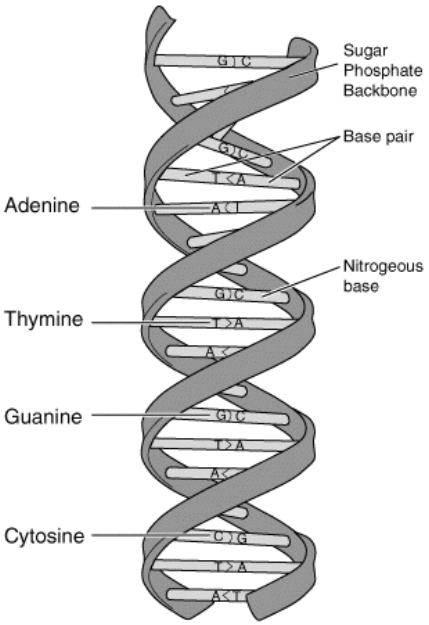


Fig. 1. The DNA structure. http://www.niherst.gov.tt/scipop/sci-bits/genetics-and-epigentics.htm

Positions in a DNA sequence at which nucleotides C or G are situated can be represented by a 1, and locations with T or A are situated can be represented by a 0. More formally, a sequence $a = \{a_1, \ldots, a_L\}$ of length L is given, where $a_m \in \{A, C, G, T\}$. The sequence may be transformed to a binary sequence $b = \{b_1, \ldots, b_L\}$ in which

$$b_m = \begin{cases} 1 & \text{if } a_m \in \{C, G\}, \\ 0 & \text{if } a_m \in \{A, T\}. \end{cases}$$

This work was carried out when the first author was at the Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, NO 7491, Trondheim, Norway.

978-83-60810-58-3/$25.00

From mathematical point of view we can designate a bound of segments with different GC ratio as a break-point or a change-point. Biological applications of the change-point problem, in particular, to DNA sequences, have been extensively considered in literature (see, for example, [12], [13], [14], [15], [16]). Note that the multiple change-point problem is a flexible model, which can be applied in many areas such as economics, finance, environmental control [17], [18], [19], signal detection, quality control [20], health and surveillance [21], [22]. Various techniques to the change-point problem with independent observations have been developed [16], [23], [24], [25], including stochastic optimization methods [13], [26], [27], [28], [29], [30], [31] and Markov chain Monte Carlo (MCMC) algorithms [14], [32], [33], [34], [35]. The Cross-Entropy (CE) method for independent case was developed in [13].

We can formulate a more general change-point problem for a sequence of dependent observations. The case of the Markov dependence in biological sequences was investigated in different articles. Polansky [36] considered cases with known and unknown number of change-points. The author applied the likelihood ratio, the bootstrap for estimation p-values for these cases, the Bayesian information criterion (BIC) and the Akaike information criterion (AIC) with unknown number of change-points. Zhang and Siegmund [37] proposed a new penalty component in the modified BIC. Avery and Herderson [12] investigated a problem of prediction of the occurrence of the definite sequence in DNA. For this purpose they considered the first-order, the second-order and the higher-order Markov chain models. Then the authors [38] developed a nonparametric method based on the approach of Pettitt [39]. Krauth [15], [40] used the exact Fisher test and the finite conditional tests for the multiple change-point problem in binary first-order Markov sequences. In this paper, we develop the CE method for identifying change-points in the first-order Markov dependence in binary sequences for artificial and real data.

We use the genome of the Bacteriophage *lambda*, a virus of the intestinal bacterium *Escherichia coli*, and a part of the Human Major Histocompatibility Region. Consideration of individual chromosomes is one of the most common approaches in the literature [41], [42]. Particularly it is very important for the analysis of the cancer genome [43].

The paper is structured as follows. Section 2 provides a statement of the multiple change-point problem in mathematical terms. In Section, 3 we describe a general framework of Cross-Entropy method. Section 4 contains developing the Cross-Entropy algorithm for the multiple change-point problem in dependent case. In Section 5, we discuss the results of numerical experiments.

II. The Multiple Change-Point Problem in Binary Markov Sequences

In mathematical terms we can describe the general multiple change-point problem as follows. A binary sequence $b = (b_1, \dots, b_L)$ of length L is given. A segmentation of the sequence is specified by giving the positions of the change-points $c = (c_1, \dots, c_N)$ and the number of change-points N, where $1 = c_0 < c_1 < \cdots < c_N < c_{N+1} = L$. This means that a change-point is a boundary between two neighbouring regions, and the value c_n is the sequence position of the rightmost character of the segment to the left of the n-th change-point.

In this model we assume that characters within each region are generated by Bernoulli trials with first-order Markov dependence. The probability distribution, which depends on the segment, can be represented by a transition matrix

$$\begin{pmatrix} \theta_0 & 1-\theta_0 \\ \theta_1 & 1-\theta_1 \end{pmatrix},$$

where $\theta_0 = P(X_{m+1} = 0 \mid X_m = 0)$, $1-\theta_0 = P(X_{m+1} = 1 \mid X_m = 0)$, $\theta_1 = P(X_{m+1} = 0 \mid X_m = 1)$, $1-\theta_1 = P(X_{m+1} = 1 \mid X_m = 1)$.

Thus, the likelihood function of N, $c = (c_1, \dots, c_N)$, and

$$\theta = (\theta_{00}, \theta_{10}, \dots, \theta_{0n}, \theta_{1n}, \dots, \theta_{0N}, \theta_{1N}),$$

is given by

$$\begin{aligned} f(N, c, \theta) &= P(X_1 = b_1) \\ &\times \prod_{n=0}^{N} \theta_{0n}^{\mathbf{I}_{00}(c_n, c_{n+1})} (1-\theta_{0n})^{\mathbf{I}_{01}(c_n, c_{n+1})} \\ &\times \theta_{1n}^{\mathbf{I}_{10}(c_n, c_{n+1})} (1-\theta_{1n})^{\mathbf{I}_{11}(c_n, c_{n+1})}, \end{aligned}$$

where $\mathbf{I}_{ij}(c_n, c_{n+1})$ is the number of times i ($i = 0, 1$), is followed by j ($j = 0, 1$) in the segment bounded by sequence positions $c_n + 1$ and c_{n+1}.

In order to simplify optimization, we consider the log-likelihood function at point $x = (N, c, \theta)$, having observed $b_1, \dots, b_L$,

$$\begin{aligned} \pi(x) &= \ln P(X_1 = b_1) \\ &+ \sum_{n=0}^{N} \Big(\mathbf{I}_{00}(c_n, c_{n+1}) \ln \theta_{0n} \\ &+ \mathbf{I}_{01}(c_n, c_{n+1}) \ln(1-\theta_{0n}) \\ &+ \mathbf{I}_{10}(c_n, c_{n+1}) \ln \theta_{1n} + \mathbf{I}_{11}(c_n, c_{n+1}) \ln(1-\theta_{1n}) \Big). \end{aligned} \tag{1}$$

III. The Cross-Entropy Method

From mathematical point of view the multiple change-point detection problem can be interpreted as a maximization problem of the log-likelihood function defined in (1).

Let F be a real valued performance function on $\mathcal{X}$, where $\mathcal{X}$ is a finite set of states. We want to find the optimum of F over $\mathcal{X}$, and the state corresponding to this value (which is a vector of positions of change-points). We can apply stochastic optimization methods for this optimization problem, in particular, the CE method.

The CE method is a technique for the estimation of rare event probabilities [44], [45], [46]. This estimation problem can be reformulated as an optimization problem. Thus we

define a set of indicator functions $\{I_{\{S(x)\geq\gamma\}}\}$ on $\mathcal{X}$ for different levels $\gamma \in R$. Let $\{f(\cdot,u)\}$ be a family of probability density functions (pdfs) on $\mathcal{X}$ with a real-valued parameter u. Following [45], we associate the optimization problem with the problem of estimating

$$l(\gamma) = \mathbf{P}_u(S(X) \geq \gamma) = \sum_x I_{\{S(x)\geq\gamma\}} f(x,u) = \mathbf{E}_u I_{\{S(X)\geq\gamma\}},$$

where γ is a known or unknown parameter and $\mathbf{P}_u$ is the probability measure under which the random state X has the pdf $f(\cdot,u)$.

The problem of estimating l is not trivial. Adaptive changes to the pdf are based on the Kullback-Leibler (or the CE) distance. Thus it allows to create a sequence $f(\cdot,u_0), f(\cdot,u_1), \ldots, f(\cdot,u^*)$. The final pdf $f(\cdot,u^*)$ corresponds to the density at an optimal point. This means that the CE method creates a sequence of pairs $\{(\gamma_t,u_t)\}$, which converges quickly to a close neighbourhood of the optimal tuple (γ^*,u^*). More specifically, we should set up u_0 and simulation parameters, and then we carry out the following procedure [45]:

1) **Adaptive updating of** γ_t. For a fixed u_{t-1}, let γ_t be a $(1-\rho)$-quantile of $\widehat{S}(X)$ under u_{t-1}. A simple estimator $\widehat{\gamma}_t$ of γ_t is
$$\widehat{\gamma}_t = \widehat{S}_{(\lceil(1-\rho)N_2\rceil)},$$
where, for a random sample $X_1,\ldots,X_{N_2}$ from $f(\cdot,u_{t-1})$, $\widehat{S}_{(i)}$ is the i-th order statistic of the performances $\widehat{S}(X_1),\ldots,\widehat{S}(X_{N_2})$.
2) **Adaptive updating of** u_t. For fixed γ_t and u_{t-1}, derive u_t from the solution of the CE program
$$\max_u D(u) = \max_u \mathbf{E}_{u_{t-1}} I_{\{\widehat{S}(X)\geq\gamma_t\}} \ln f(X,u).$$

IV. The Cross-Entropy Method for the Multiple Change-Point Problem

Let N be the number of change-points and c be a set of the change-points, which is a nondecreasing N-dimensional vector.

We apply the CE algorithm that uses normal distributions to simulate the change-point positions. The CE method updates the parameters in each step and updating is continued until a convergence state is achieved. A variance-based stopping criterion is used to estimate the fit of the combinations of change-points in each step.

Our study differs from previous [13] in the following aspects. Firstly, we consider a change-point problem for a sequence of dependent observations. Secondly, we apply the BIC (Bayesian information criterion) [47], [48] in order to estimate the number of change-points, which is usually unknown. The combination that minimizes F (our performance function) under the corresponding N is considered as the optimal solution. Therefore, we replace the problem of maximization of log-likelihood function with minimization problem of the BIC.

TABLE I
Parameters θ in Example 1

positions	θ_1	θ_2
1–2000	0.9	0.5
2001–4000	0.4	0.15
4001–6000	0.1	0.6
6001–8000	0.6	0.9
8001–10000	0.2	0.4
10001–12000	0.4	0.2
12001–14000	0.2	0.7
14001–16000	0.6	0.5
16001–18000	0.4	0.9
18001–20000	0.2	0.2
20001–22000	0.7	0.5

For each change-point vector c in the sample, we obtain the maximum likelihood estimate of parameters with respect to the each of the segments and evaluate the performance function F. The performance function, the BIC, which we minimize is

$$F = -2\pi(x) + k\ln(L), \tag{2}$$

where $\pi(x)$ is the log-likelihood as in (1) of the sequence. We use the standard penalty

$$k\ln(L) = (3N+2)\ln(L).$$

In each iteration an *elite* sample is defined as the best performing combinations of change-points with respect to the performance function score. The process is carried out until a specific stopping criterion is achieved.

In each step, the simulation parameters are updated accordingly. The main steps of our algorithm are described in Algorithm 1.

We should specify N_1, ρ, ε , the parameters of the algorithm as well as the initial values for the simulation parameters μ and σ^2. Note that we choose the parameters under the conditions which guarantee convergence of the algorithm [49].

V. Numerical Results

In this section, we include results of numerical experiments that illustrate the performance of the CE method. In the first example, we consider a synthetic sequence with a known distribution, which allows us to provide direct comparison of estimated and true profiles in terms of the Root Mean Squared Error (RMSE). The second and the third examples use real DNA sequences and we do not have any information about the structure of dependence. We apply a test of independence for these examples.

A. Example 1: Artificial data

Let $(b_1, b_2, ..., b_{22000})$ be a sequence of random variables generated with the parameters from Table I.

At first, we assume that we do not know the number of change-point and apply our algorithm for different N. We run our algorithm with the following simulation parameters: the elite proportion value $\rho = 0.1$ and the sample size $N_1 = 1500$.

Algorithm 1 Algorithm for change-point detection

1: Choose initial sets for

$$\mu^{(0)} = \left(\mu_1^{(0)}, \mu_2^{(0)}, \ldots, \mu_N^{(0)}\right)$$

and

$$(\sigma^2)^{(0)} = \left((\sigma_1^2)^{(0)}, (\sigma_2^2)^{(0)}, \ldots, (\sigma_N^2)^{(0)}\right).$$

The length of both vectors is N. Set $t = 1$.

2: Generate a random sample $c^{(1)}, c^{(2)}, \ldots, c^{(N_1)}$ from the normal distributions with parameters $\left(\mu^{(t-1)}, (\sigma^2)^{(t-1)}\right)$, where $c^{(i)} = \left(c_1^{(i)}, c_2^{(i)}, \ldots, c_N^{(i)}\right)$, $i = 1, 2, \ldots, N_1$, is a change-point vector.

3: For $i = 1, 2, \ldots, N_1$ order $\left(c_1^{(i)}, c_2^{(i)}, \ldots, c_N^{(i)}\right)$ from smallest to biggest.

4: Evaluate the performance of each $c^{(1)}, c^{(2)}, \ldots, c^{(N_1)}$ based on (2).

5: Define the elite sample, which is the best performing combinations of the change-points.

6: Let $N_{elite} = \rho N_1$ be the size of the elite sample.

7: For all $j = 1, 2, \ldots, N$, estimate the parameters $\mu_j^{(t)}$ and $(\sigma_j^2)^{(t)}$ using the elite sample and update the current parameter sets as follows:

$$\mu_j^{(t)} = \frac{\sum\limits_{i \in I} c_j^{(i)}}{N_{elite}}, \qquad \left(\sigma_j^2\right)^{(t)} = \frac{\sum\limits_{i \in I} \left(c_j^{(i)} - \mu_j^{(t)}\right)^2}{N_{elite}},$$

where I is the set of indices of the best performing samples.

8: Stopping criterion is $\max_j (\sigma_j^2)^{(t)} < \varepsilon$.

9: **if** Stopping criterion is met **then**

10: stop the process and identify the combination of the positions of change points $c^{(i)}$ that minimizes the BIC

11: **else**

12: $t = t + 1$;

13: and iterate from step 2.

14: **end if**

Then we obtain the best solution for different models in each of the N situations which minimize the BIC. We can see from Figure 2 that the minimum value of the BIC at $N = 10$, which corresponds to the number of change-points in Table I.

The true profiles of this sequence as well as the estimated profile can be seen in Figures 3, 4. We can see that the estimated and the true plots are very similar to each other. This indicates that the CE method works very well and it properly captures the segments in the binary sequence.

To test the efficiency of the CE method, we have applied this algorithm with various values for the parameter ρ, which is used to obtain the elite sample. We calculate the RMSE for the different algorithms when applied to the synthetic sequence

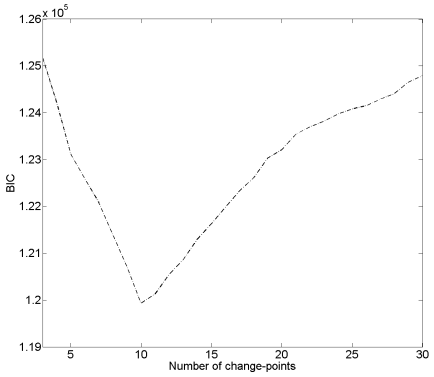


Fig. 2. The scores of the BIC for different N

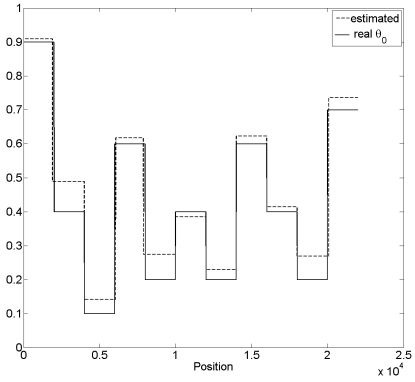


Fig. 3. The profile of θ_0 obtained by the CE algorithm

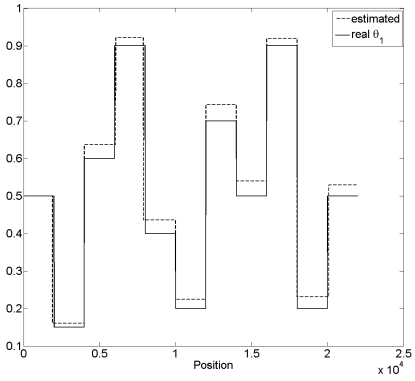


Fig. 4. The profile of θ_1 obtained by the CE algorithm

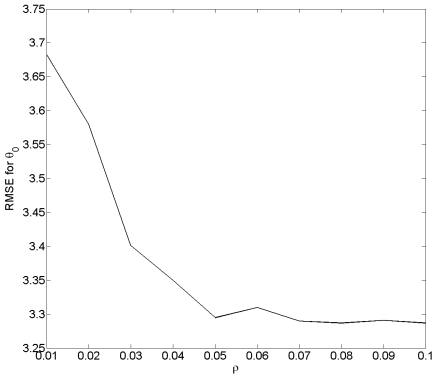


Fig. 5. The values of the RMSE for θ_0 depending on ρ

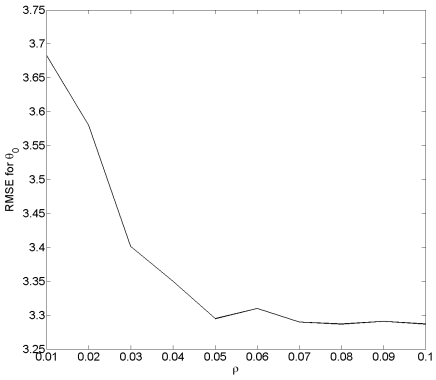


Fig. 6. The values of the RMSE for θ_1 depending on ρ

of 22000 characters

$$\text{RMSE} = \sqrt{\sum_{i=1}^{22000} (t(i) - e(i))^2},$$

where $e(i)$ is estimated value at position i and $t(i)$ is the true parameter value.

The RMSE and CPU time are obtained for ρ values from 0.01 to 0.1 with step of 0.01 for the model when number of change-points is 10. We have obtained the average results based on 50 simulations under each of the ρ values. We can see from Figures 5, 6 that the plots are slowly decreasing, at the same time the plot on Figure 7 is increasing. In this study, we focus on the RMSE, though it would be possible to choose ρ in such a way that will balance the trade-off between the RMSE and the CPU time.

B. Example 2: Real data (Bacteriophage lambda)

We apply the CE with the same parameter specification as above to the genome of the Bacteriophage *lambda*, a virus of the intestinal bacterium *Escherichia coli*. The length of the sequence is 48,502 bases. Boys and Henderson [50] studied this sequence with 4 multinomial outcomes (each base is one of either A, C, G, T) for the comparison of different algorithms

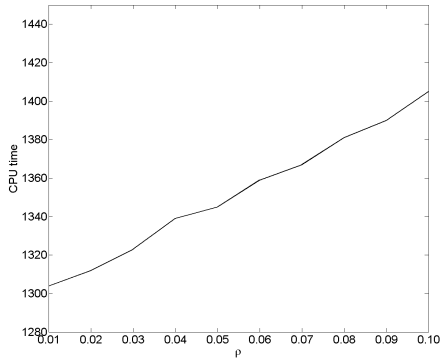


Fig. 7. CPU time for different ρ

TABLE III
OBSERVED FREQUENCIES OF THE 4 POSSIBLE PAIRS OF BASES FOR EXAMPLE 2. EXPECTED FREQUENCIES ASSUMING INDEPENDENCE OF SUCCESSIVE BASES ARE GIVEN IN PARENTHESES

First base	**Second base**	**Second base**	**Total**
	0	1	
0	12544	11776	24320
	(12194.85)	(12125.15)	
1	11776	12405	24181
	(12125.15)	(12055.85)	
Total	24320	24181	48501

under the independence assumption. Table II presents a brief summary of the results obtained in [51].

Table III shows the observed frequencies and the expected frequencies for the Pearson χ^2-test of independence. It can be calculated from the table that the value of the test statistic is 40.22. On comparing with a χ^2-distribution with 1 degree of freedom, we conclude that the hypothesis about independence should be rejected ($p < 10^{-6}$). Therefore, we consider a case with the first-order Markov dependence.

We can calculate the BIC for different number of change-points. From Table II we can see that the authors found 8 change-points based on the use of 4-symbol alphabet. According to our approach we found that 6236 was the minimum value of the BIC at $N = 9$. Next, we check each change-point using the Fisher exact test. We calculate p-values and conclude that there are evidences for change-points at 5806 ($p_1 = 7.82 \cdot 10^{-4}$), 19503 ($p_4 = 0.018$), 22109 ($p_5 = 1.25 \cdot 10^{-11}$), 27660 ($p_6 = 6.18 \cdot 10^{-6}$), 38018 ($p_8 = 0.0045$), and 46259 ($p_9 = 5.19 \cdot 10^{-4}$).

Note that our main objective is to identify change-points in GC ratio, not in the model parameters θ_0, θ_1. Therefore, we present our conclusions without profiles of θ_0 and θ_1 and locations of false change-points. The GC profile can be seen on Figure 8. The discordance can be explained by the fact that the results in Table II were obtained using a different model with 4-character alphabet, whereas we used a binary representation. From this comparison we can see that both methods identify

TABLE II
ESTIMATED SEGMENTS AND ESTIMATED PROPORTIONS OF A, C, G, AND T FOR EACH SEGMENT

	A	C	G	T	G+C	A+T
0 – 20091	0.23	0.25	0.32	0.20	0.57	0.43
20092 – 20919	0.29	0.29	0.30	0.11	0.59	0.41
20902 – 22544	0.26	0.24	0.27	0.23	0.51	0.49
22545 – 24117	0.29	0.14	0.16	0.40	0.30	0.70
24118 – 27829	0.29	0.20	0.18	0.33	0.38	0.62
27830 – 33082	0.23	0.26	0.22	0.29	0.48	0.52
33083 – 38029	0.27	0.22	0.21	0.31	0.43	0.57
38030 – 46528	0.30	0.23	0.26	0.22	0.49	0.51
46529 – 48502	0.27	0.18	0.22	0.33	0.40	0.60

TABLE IV
OBSERVED FREQUENCIES OF THE 4 POSSIBLE PAIRS OF BASES FOR EXAMPLE 3

First base	Second base	Second base	Total
	0	1	
0	5344	5345	10689
	(5713.56)	(4975.49)	
1	5346	3964	9310
	(4976.44)	(4333.56)	
Total	10690	9309	19999

the most significant change-points and the proposed method provides a smoother profile of GC ratio.

C. Example 3: Real data (MHC Region)

This example uses a part of the Human Major Histocompatibility Region (MHC) (for further detail, see [52]). Due to this being real DNA, we do not know the true profile (as well as in Example 2). Instead we look for agreement between the CE and two well-known methods: IsoFinder [16], [23], [24] and the BAIS [34], [35]. At first, we repeat the Pearson test of independence. The value of the test statistic from Table IV is 51.35. This means that the hypothesis about independence should also be rejected ($p < 10^{-6}$).

We use the same algorithm parameters as before. We found a change-point vector and checked each position using the exact Fisher test. There are 6 significant change-points in this part of MHC sequence: 953 ($p_1 = 3.67 \cdot 10^{-4}$), 7257 ($p_4 = 0.0078$), 9132 ($p_5 = 7.80 \cdot 10^{-6}$), 13041 ($p_6 = 6.28 \cdot 10^{-12}$), 16114 ($p_7 = 3.19 \cdot 10^{-11}$), and 18954 ($p_8 = 1.05 \cdot 10^{-30}$).

Figure 9 shows the GC profiles for the CE algorithm, the BAIS and the IsoFinder. We use the following simulation parameters: the BAIS algorithm for 1000 iterations and $K = 50$ parallel chains, and IsoFinder with a 0.95 significance level and tract size of 1,000. It is clear that all algorithms can detect the major regions within the MHC sequence. IsoFinder identifies seven major regions while the other methods all identify several smaller regions within these major regions. The agreement between these methods allows for a great deal of confidence in the exactness of the CE method as both the BAIS method and IsoFinder are well established.

VI. CONCLUSION

In this paper, we have developed the Cross-Entropy method for identifying change-points in binary Markov sequences. In order to identify the correct number of change-points we propose to use the BIC. This approach is easy to implement and can also be extended to more general multiple change-point models. We have demonstrated the effectiveness of this technique in examples using both real and synthetic sequences. The method has been shown to be highly effective on synthetic data and real DNA sequences and compete well with existing approaches.

The proposed approach gives results similar to previous outcomes but it is not sufficient for understanding of dependence mechanism in DNA sequences. Our future research will include consideration of Markov dependence of a higher order (the second or more). The proposed method can be implemented using parallel computing, which will significantly decrease the CPU time. For the independent case, this feature was realized in R-package *breakpoint* [53].

ACKNOWLEDGMENT

The first author was supported by ERCIM programme. This work was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme at the Norwegian University of Science and Technology, Trondheim, Norway. This programme is supported by the Marie Curie Co-funding of Regional, National and International Programmes (COFUND) of the European Commission.

REFERENCES

[1] G. Bernardi, *Structural and evolutionary genomics. Natural Selection in Genome Evolution*. Amsterdam: Elsevier, 2004.
[2] M. Costantini, O. Clay, F. Auletta, and G. Bernardi, "An isochore map of human chromosomes," *Genome Res.*, vol. 16, pp. 536–541, 2006, http://dx.doi.org/10.1101/gr.4910606
[3] L. Ren, G. Gao, D. Zhao, M. Ding, J. Luo, and H. Deng, "Developmental stage related patterns of codon usage and genomic GC content: searching for evolutionary fingerprint by models of stem cell differentiation," *Genome Biol.*, vol. 8, p. R35, 2007, http://dx.doi.org/10.1186/gb-2007-8-3-r35
[4] M. Semon, D. Mouchiroud, and L. Duret, "Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance," *Hum. Mol. Genet.*, vol. 14, pp. 421–427, 2005, http://dx.doi.org/10.1093/hmg/ddi038

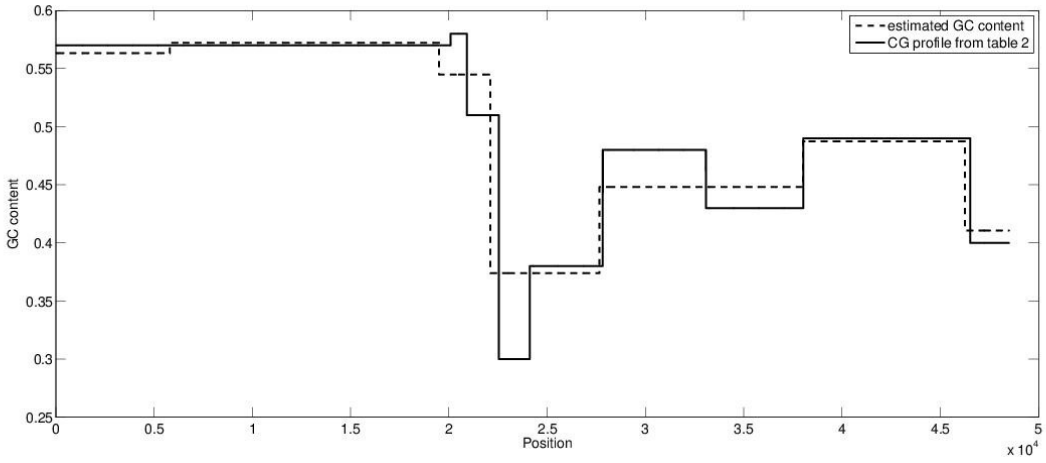


Fig. 8. Bacteriophage *lambda*.

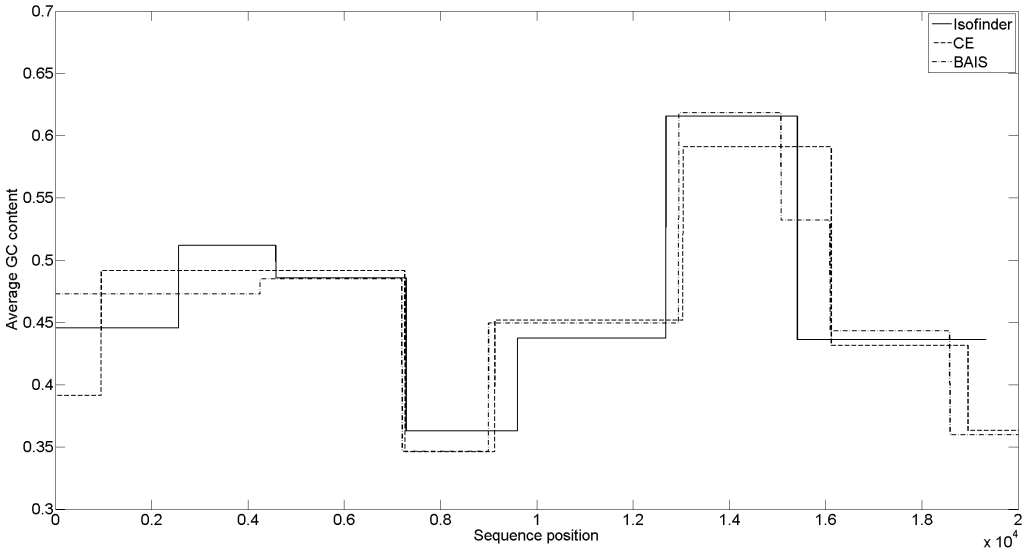


Fig. 9. GC profiles for the CE, IsoFinder and BAIS methods on the MHC sequence.

[5] C. Spencer, P. Deloukas, S. Hunt, J. Mullikin, S. Myers, B. Silverman, P. Donnelly, D. Bentley, and G. McVean, "The influence of recombination on human genetic diversity," *PLoS Genet*, vol. 2, p. e148, 2006, http://dx.doi.org/10.1371/journal.pgen.0020148

[6] A. Vinogradov, "Isochores and tissue-specificity," *Nucleic Acids Res.*, vol. 31, pp. 5212–5220, 2003, http://dx.doi.org/10.1093/nar/gkg699

[7] A. Vinogradov, "Dualism of gene GC content and CpG pattern in regard to expression in the human genome: magnitude versus breadth," *Trends Genet*, vol. 21, pp. 639–643, 2005, http://dx.doi.org/10.1016/j.tig.2005.09.002

[8] L. Hurst, C. Brunton, and N. Smith, "Small introns tend to occur in GC-rich regions in some but not all vertebrates," *Trends Genet*, vol. 15, pp. 437–439, 1999, http://dx.doi.org/10.1016/S0168-9525(99)01832-6

[9] T. Ikemura and K. Wada, "Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers; relation between nucleotide sequence data and cytogenetic data," *Nucleic Acids Res.*, vol. 19, pp. 4333–4339, 1991.

[10] T. Takano-Shimizu, "Local changes in GC/AT substitution biases and in crossover frequencies on drosophila chromosomes," *Mol. Biol. Evol.*, vol. 18, pp. 606–619, 2001, http://dx.doi.org/10.1093/oxfordjournals.molbev.a003841

[11] E. Willams and L. Hurst, "The proteins of linked genes evolve at similar rates," *Nature*, vol. 407, pp. 900–903, 2000, http://dx.doi.org/10.1038/35038066

[12] P. Avery and D. Henderson, "Fitting Markov chain models to discrete state series such as DNA sequences," *Appl. Statist.*, vol. 48, no. 1, pp. 53–61, 1999, http://dx.doi.org/10.1111/1467-9876.00139

[13] G. E. Evans, G. Y. Sofronov, J. M. Keith, and D. P. Kroese, "Estimating change-points in biological sequences via the cross-entropy method," *Ann. Oper. Res.*, vol. 189, no. 1, pp. 155–165, 2011, http://dx.doi.org/10.1007/s10479-010-0687-0

[14] J. M. Keith, "Segmenting eukaryotic genomes with the generalized Gibbs sampler," *J. Comp. Biol.*, vol. 13, no. 7, pp. 1369–1383, 2006, http://dx.doi.org/10.1089/cmb.2006.13.1369

[15] J. Krauth, "Multiple change points and alternating segments in binary trials with dependence," in *Innovations in Classification, Data Science,*

and Information Systems, D. Baier and K. Wernecke, Eds. Springer, Berlin, 2004, pp. 154–164, http://dx.doi.org/10.1007/3-540-26981-9_19

[16] J. Oliver, P. Bernaola-Galvan, P. Carpena, and R. Roman-Roldan, "Isochore chromosome maps of eukaryotic genomes," *Gene*, vol. 276, pp. 47–56, 2001, http://dx.doi.org/10.1016/S0378-1119(01)00641-2

[17] I. López, M. Gámez, J. Garay, T. Standovár, and Z. Varga, "Application of change-point problem to the detection of plant patches," *Acta Biotheoretica*, vol. 58, pp. 51–63, 2010, http://dx.doi.org/10.1007/s10441-009-9093-x

[18] J. R. Thomson, W. J. Kimmerer, L. R. Brown, K. B. Newman, R. Mac Nally, W. A. Bennett, F. Feyrer, and E. Fleishman, "Bayesian change point analysis of abundance trends for pelagic fishes in the upper San Francisco estuary," *Ecological Applications*, vol. 20, no. 5, pp. 1431–1448, 2010, http://dx.doi.org/10.1890/09-0998.1

[19] M. Priyadarshana, G. Sofronov, "A modified cross-entropy method for detecting change-points in the Sri-Lankan stock market," In: *The IASTED International Conference on Engineering and Applied Science (EAS2012)*, Chen, B. M.; Khan, M. T. and Tan, K-K. (Eds.), 2012, pp. 321-326, http://dx.doi.org/10.2316/P.2012.785-041

[20] G. Sofronov, T. Polushina, and M. Priyadarshana, "Sequential change-point detection via the cross-entropy method," in *The 11th Symposium on Neural Network Applications in Electrical Engineering (NEUREL2012)*, B. Reljin and S. Stankovic, (Eds.), 2012, pp. 185–188, http://dx.doi.org/10.1109/NEUREL.2012.6420004

[21] C. Sonesson and D. Bock, "A review and discussion of prospective statistical surveillance in public health," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, vol. 166, no. 1, pp. 5–21, 2003, http://dx.doi.org/10.1111/1467-985X.00256

[22] J. Whittaker and S. Frühwirth-Schnatter, "A dynamic changepoint model for detecting the onset of growth in bacteriological infections," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 43, no. 4, pp. 625–640, 1994, http://dx.doi.org/10.2307/2986261

[23] J. Oliver, P. Carpena, R. Roman-Roldan, T. Mata-Balaguer, A. Mejias-Romero, M. Hackenberg, and P. Bernaola-Galvan, "Isochore chromosome maps of the human genome," *Gene*, vol. 300, pp. 117–127, 2002, http://dx.doi.org/10.1016/S0378-1119(02)01034-X

[24] J. Oliver, R. Roman-Roldan, J. Perez, and P. Bernaola-Galvan, "Segment: identifying compositional domains in DNA sequences," *Bioinformatics*, vol. 15, pp. 974–979, 1999, http://dx.doi.org/10.1093/bioinformatics/15.12.974

[25] G. Y. Sofronov, G. E.Evans, J. M. Keith, and D. P. Kroese, "Identifying change-points in biological sequences via sequential importance sampling," *Environmental Modeling and Assessment*, vol. 14, no. 5, pp. 577–584, 2009, http://dx.doi.org/10.1007/s10666-008-9160-8

[26] T. Polushina and G. Sofronov, "Change-point detection in biological sequences via genetic algorithm," in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC'2011)*, 2011, pp. 1966–1971, http://dx.doi.org/10.1109/CEC.2011.5949856

[27] M. Priyadarshana and G. Sofronov, "The Cross-Entropy method and multiple change-points detection in zero-inflated DNA read count data," in *The 4th International Conference on Computational Methods (ICCM2012)*, Y. T. Gu, S. C. Saha (Eds.), 2012 , pp. 1–8.

[28] M. Priyadarshana and G. Sofronov, "GAMLSS and Extended Cross-Entropy Method to Detect Multiple Change-Points in DNA Read Count Data," in *Proceedings of the 28th International Workshop on Statistical Modelling*, Muggeo VMR, Capursi V, Boscaino G, Lovison G (Eds.), 2013, vol.1, pp. 453–457.

[29] T. V. Polushina and G. Y. Sofronov, "A hybrid genetic algorithm for change-point detection in binary biomolecular sequences," in *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2013)*, 2013, pp. 1–8, http://dx.doi.org/10.2316/P.2013.793-026

[30] M. Priyadarshana, T. Polushina, and G. Sofronov, "A hybrid algorithm for multiple change-point detection in continuous measurements," in *International Symposium on Computational Models for Life Sciences, AIP Conference Proceedings*, vol. 1559, 2013, pp. 108–117, http://dx.doi.org/10.1063/1.4825002

[31] M. Priyadarshana, T. Polushina, and G. Sofronov, "Hybrid algorithms for multiple change-point detection in biological sequences," in Signal and Image Analysis for Biomedical and Life Sciences (Advances in Experimental Medicine and Biology), Sun, C., Bednarz, T., Pham, T. D., Vallotton, P., Wang, D. (Eds.), Springer, 2014, in press.

[32] J. M. Keith, P. Adams, S. Stephen, and J. S. Mattick, "Delineating slowly and rapidly evolving fractions of the drosophila genome," *J. Comp. Biol.*, vol. 15, no. 4, pp. 407–430, 2008, http://dx.doi.org/10.1089/cmb.2007.0173

[33] J. M. Keith, D. P. Kroese, and D. Bryant, "A generalized Markov sampler," *Methodology and Computing in Applied Probability*, vol. 6, no. 1, pp. 29–53, 2004, http://dx.doi.org/10.1023/B:MCAP.0000012414.14405.15

[34] J. Keith, D. Kroese, and G. Sofronov, "Adaptive independence samplers," *Statistics and Computing*, vol. 18, no. 4, pp. 409–420, 2008, http://dx.doi.org/10.1007/s11222-008-9070-2

[35] G. Sofronov, "Change-point modelling in biological sequences via the bayesian adaptive independent sampler," *International Proceedings of Computer Science and Information Technology*, vol. 5, pp. 122–126, 2011.

[36] A. Polansky, "Detecting change-points in Markov chains," *Computational Statistics and Data Analysis*, vol. 51, pp. 6013–6026, 2007, http://dx.doi.org/10.1016/j.csda.2006.11.040

[37] N. Zhang and D. Siegmund, "A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data," *Biometrics*, vol. 3, pp. 22–32, 2007, http://dx.doi.org/10.1111/j.1541-0420.2006.00662.x

[38] P. Avery and D. Henderson, "Detecting a changed segment in DNA sequences," *Appl. Statist.*, vol. 48, no. 4, pp. 489–503, 1999, http://dx.doi.org/10.1111/1467-9876.00167

[39] A. Pettitt, "A non-parametric approach to the change-point problem," *Appl. Statist.*, vol. 28, pp. 126–135, 1979, http://dx.doi.org/10.2307/2346729

[40] J. Krauth, "Tests for multiple change points in binary Markov sequences," in *From Data and Information Analysis to Knowledge Engineering*. Springer, Berlin, 2006, pp. 670–677, http://dx.doi.org/10.1007/3-540-31314-1_82

[41] R. Thurman, N. Day, W. Noble, and J. Stamatoyannopoulos, "Identification of higher-order functional domains in the human ENCODE regions," *Genome Res.*, vol. 17, pp. 917–927, 2007, http://dx.doi.org/10.1101/gr.6081407

[42] H. Xu, C. Wei, F. Lin, and W. Sung, "An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data," *Bioinformatics*, vol. 24, pp. 2344–2349, 2008, http://dx.doi.org/10.1093/bioinformatics/btn402

[43] B. Zeitouni, V. Boeva, I. Janoueix-Lerosey, O. Delattre, A. Nicolas, and E. Barillot, "SVDetect - a bioinformatic tool to identify genomic structural variations from paired-end next-generation sequencing data," *Bioinformatics*, vol. 26, pp. 1895–1896, 2010, http://dx.doi.org/10.1093/bioinformatics/btq293

[44] Z. I. Botev, D. Kroese, and T. Taimre, "Generalized cross-entropy methods with applications to rare-event simulation and optimization," *Simulation*, vol. 83, no. 11, pp. 785–806, 2007, http://dx.doi.org/10.1177/0037549707087067

[45] R. Rubinstein and D. Kroese, *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. New York: Springer-Verlag, 2004.

[46] R. Rubinstein and D. Kroese, *Simulation and the Monte Carlo Method*. John Wiley & Sons, 2007.

[47] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978, http://dx.doi.org/10.1214/aos/1176344136

[48] Y. Yao, "Estimating the number of change-points via Schwarz criterion," *Statistics and Probability Letters*, vol. 6, pp. 181–189, 1988, http://dx.doi.org/10.1016/0167-7152(88)90118-6

[49] A. Costa, O. Jones, and D. Kroese, "Convergence properties of the cross- entropy method for discrete optimization," *Operations Research Letters*, vol. 35, no. 5, pp. 573–580, 2007, http://dx.doi.org/10.1016/j.orl.2006.11.005

[50] R. Boys and D. Henderson, "A Bayesian approach to DNA sequence segmentation," *Biometrics*, vol. 60, pp. 573–588, 2004, http://dx.doi.org/10.1111/j.0006-341X.2005.040701_1.x

[51] J. Braun, R. Braun, and H. Müller, "Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation," *Biometrika*, vol. 87, no. 2, pp. 301–314, 2000, http://dx.doi.org/10.1093/biomet/87.2.301

[52] The MHC Sequencing Consortium, "Complete sequence and gene map of a human major histocompatibility complex," *Nature*, vol. 401, no. 6756, pp. 921–923, 1999, http://dx.doi.org/10.1038/44853

[53] M. Priyadarshana, and G. Sofronov, "Breakpoint (R-package);" software available at http://cran.r-project.org/web/packages/breakpoint.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 479–486
DOI: 10.15439/2014F297
ACSIS, Vol. 2

An efficient algorithm for the density Turán problem of some unicyclic graphs

Halina Bielak
Institute of Mathematics
Maria Curie Skłodowska University
Pl. M. Curie-Skłodowskiej 5
20-031 Lublin, Poland
Email: hbiel@hektor.umcs.lublin.pl

Kamil Powroźnik
Institute of Mathematics
Maria Curie Skłodowska University
Pl. M. Curie-Skłodowskiej 5
20-031 Lublin, Poland
Email: kamil.pawel.powroznik@gmail.com

***Abstract*—Let $H = (V(H), E(H))$ be a simple connected graph of order n with the vertex set $V(H)$ and the edge set $E(H)$. We consider a blow-up graph $G[H]$.**

We are interested in the following problem. We have to decide whether there exists a blow-up graph $G[H]$, with edge densities satisfying special conditions (homogeneous or inhomogeneous), such that the graph H does not appear in a blow-up graph as a transversal.

We study this problem for unicyclic graphs H with the cycle C_3. We show an efficient algorithm to decide whether a given set of edge densities ensures the existence of H in the blow-up graph $G[H]$.

***Index Terms*—blow-up graph; density; Turán density problem; unicyclic graph.**

I. Introduction

TURÁN [10] stated the first results in extremal graph theory. Then many authors extended this subject and formulated similar and new Turán density problems. [1], [3], [4], [6], [8], [9] and [11] obtained interesting results for some families of graphs.

In this paper we present an algorithm for testing whether a unicyclic graph with a given set of edge densities is a factor (transversal) of a blow-up graph. Our algorithm has the time complexity at most $\mathcal{O}(n^2)$, where n is the number of vertices of the unicyclic graph.

Csikvári and Nagy [5] discovered some interesting algorithm for testing whether a tree with a given set of edge densities is a factor of a blow-up graph. We extend their algorithm to the family of unicyclic graphs with the cycle C_3.

Now we define some notions and notations. Other definitions one can find in [2] and [7].

Let $H = (V(H), E(H))$ be a simple connected graph of order n with the vertex set $V(H)$ and the edge set $E(H)$. By P_k we denote the path with k vertices. By C_k we denote the cycle with k vertices. The set $S \subset V(H)$ is called *an independent vertex set* if the subgraph of H induced by S has empty set of edges.

Let

$$N_H(v) = \{x \in V(H) \mid \{v, x\} \in E(H)\}$$

be *the neighbourhood* of the vertex $v \in V(H)$ in the graph H. $|N_H(v)|$ is called *the degree* of v in $V(H)$. Each vertex of degree 1 in a graph H is called *a leaf* of the graph H.

We say that the graph H is *r-regular* if each vertex of H has degree r. The set $M \subset E(H)$ is called *the matching* (or *independent edge set*) in the graph H if the subgraph of H induced by M is 1-regular.

For a connected graph H we define *a blow-up graph* $G[H]$ of the graph H as follows. First we replace each vertex $i \in V(H)$ by an independent set of vertices A_i. Throughout this paper A_i is called *a cluster*. Next we connect vertices between the clusters A_i and A_j if i and j are adjacent in H, $i, j \in V(H)$. The graph induced by $A_i \cup A_j$ in $G[H]$ is a subgraph of a complete bipartite graph. See Fig. 2 and Fig. 3 which present examples of a blow-up graphs $G[H]$ of the graph H presented in Fig. 1.

For any two clusters we define *the density* between them by the following formula

$$d(A_i, A_j) = \frac{e(A_i, A_j)}{|A_i||A_j|},$$

where $e(A_i, A_j)$ denotes the number of edges between the clusters A_i and A_j.

The graph H is *a transversal* of $G[H]$ if H is a subgraph of $G[H]$ such that we have a homomorphism

$$\phi : V(H) \to V(G[H])$$

for which $\phi(i) \in A_i$ for all $i \in V(H)$. Other terminology: H is *a factor* of $G[H]$. An edge $e = \{i, j\}$ of the graph H we denote by $e = ij$.

The density Turán problem can be defined as follows. Let us determine the critical edge density, denoted by d_{crit}, which ensures the existence of the subgraph H of $G[H]$ as a transversal. Precisely, assume that all edges $e = \{i, j\}$ in the graph H satisfy the condition

$$d(A_i, Aj) > d_{crit},$$

where $i, j \in V(H)$. Then, no matter how we construct the blow-up graph $G[H]$, it contains the graph H as a transversal.

978-83-60810-58-3/$25.00

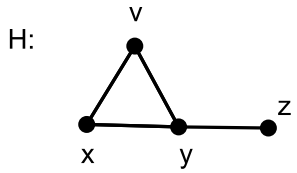


Fig. 1. The graph $H \in \mathcal{U}_{n,3}$ with the vertex set $V(H) = \{x, y, z, v\}$.

On the other words, for any value $d < d_{crit}$ there exists a blow-up graph $G[H]$ such that

$$d(A_i, A_j) > d$$

for all edges $ij \in E(H)$ and it does not contain H as a transversal. This problem was studied in [9].

By [5] we know that it is useful to consider more general problem. Let us assume that for every edge $e \in E(H)$ a density γ_e is given. Now our task is to decide if the set of densities $\{\gamma_e\}_{e \in E(H)}$ ensure the existence of the graph H as a transversal or we can construct a blow-up graph $G[H]$ such that

$$d(A_i, A_j) \geq \gamma_{ij},$$

but it does not induce the graph H as a transversal. This more general setting allows to use inductive proofs (see the proof of Theorem 7). We call this general case as *the inhomogeneous condition* on the edge densities, while the above condition of having a common lower bound $d_{crit}(H)$ for densities is called *the homogeneous case*.

Let $\mathcal{U}_{n,p}$ be a family of unicyclic graphs of order n with the cycle C_p. The path P_2 and the cycle C_3 are trivial unicyclic graphs for further considerations. In this paper we study the inhomogeneous density Turán problem for unicyclic graphs in the family $\mathcal{U}_{n,3}$, i.e. with the unique cycle C_3 (see Fig. 1).

Fig. 2 and Fig. 3 present two blow-up graphs $G_1[H]$ and $G_2[H]$ of the graph H presented in Fig. 1. In both cases we have the following values of the densities between clusters

$$d(A_x, A_y) = d(A_y, A_z) = \frac{3}{20},$$

$$d(A_x, A_v) = \frac{3}{16},$$

$$d(A_y, A_v) = \frac{1}{10}.$$

Let us recall the definition of *the multivariate matching polynomial* of the graph. The polynomial is the useful tool for the proof of our results.

Definition 1. *Let H be a graph and let $\underline{x_e}$ be the vector of variables x_e, $e \in E(H)$. We define the multivariate matching polynomial F_H of the graph H as follows*

$$F_H(\underline{x_e}, t) = \sum_{M \in \mathcal{M}} \left(\prod_{e \in M} x_e \right) (-t)^{|M|},$$

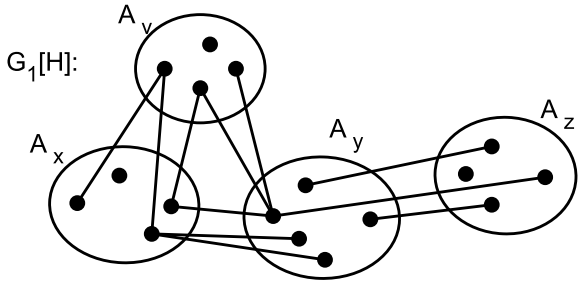


Fig. 2. An example of the blow-up graph $G[H]$ of the graph H presented in Fig. 1 with a transversal H.

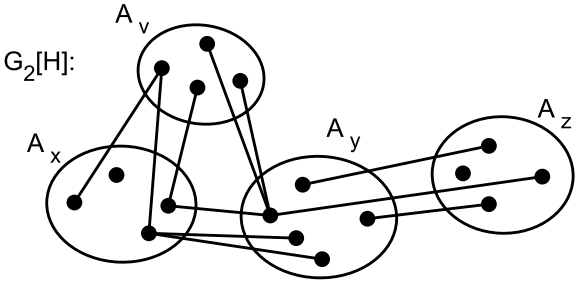


Fig. 3. An example of the blow-up graph $G[H]$ of the graph H presented in Fig. 1 without a transversal H.

where the summation goes over all matchings of the graph H, including the empty matching.

Fig. 4 and Fig. 5 present the paths P_2, P_4 and the unicycle graph $H \in \mathcal{U}_{6,3}$ with variables x_e assigned to each edge.

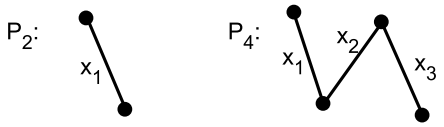


Fig. 4. Paths P_2 and P_4 with variables x_e assigned to each edge.

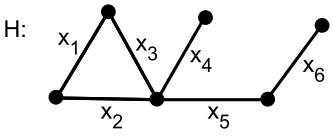


Fig. 5. Graph $H \in \mathcal{U}_{6,3}$ with variables x_e assigned to each edge.

By definition of the multivariate matching polynomial we have

$$F_{P_2}(\underline{x_e}, s) = 1 - sx_1,$$

$$F_{P_4}(\underline{x_e}, s) = 1 - s(x_1 + x_2 + x_3) + s^2 x_1 x_3,$$

$$F_H(\underline{x_e}, s) = 1 - s(x_1 + x_2 + x_3 + x_4 + x_5 + x_6) +$$

$$s^2(x_1x_4 + x_1x_5 + x_1x_6 + x_2x_6 + x_3x_6 + x_4x_6) - s^3 x_1 x_4 x_6.$$

II. Some results for the homogeneous case

For the completness of this paper we present some results for the homogeneous Turán density problem in this section. For this case Nagy [9] presented the following lower and upper bounds for the critical density d_{crit}.

Theorem 1 (Nagy [9]). *For a graph H we have*

$$\left(1-\frac{1}{\Delta(H)}\right) \leq d_{crit}(H) \leq \left(1-\frac{1}{\Delta^2(H)}\right),$$

where $\Delta(H)$ is the maximal degree of H.

Then Csikvári and Nagy [5] improved the upper bound.

Theorem 2 (Csikvári and Nagy [5]). *Let $\Delta(H)$ be the largest degree of the graph H. Then we have*

$$d_{crit}(H) \leq 1-\frac{1}{e(2\Delta(H)-1)},$$

where e is the base of the natural logarithm.

Now let us recall the definition of *the matching polynomial* of the graph.

Definition 2. *Let H be a weighted graph with constant weight function $w(e)=1$ for all edges $e \in E(H)$. Then the matching polynomial is defined as*

$$M(H,t)=\sum_{k=0}^{n/2}(-1)^k m_k(H) t^{n-2k},$$

where $m_k(H)$ denotes the number of k independent edges in the graph H.

Using this polynomial Csikvári and Nagy [5] stated the upper bound for the critical density as in Theorem 3.

Theorem 3 (Csikvári and Nagy [5]). *Let $\Delta(H)$ be the largest vertex degree in the graph H and let $t(H)$ be the largest root of the matching polynomial. Then we have*

$$d_{crit}(H) \leq 1-\frac{1}{(t(H))^2}.$$

In particular,

$$d_{crit}(H) < 1-\frac{1}{4(\Delta(H)-1)}.$$

What is more Nagy [9] showed the exact value of the critical density for trees.

Theorem 4 (Nagy [9]). *Let T be a tree. Then we have*

$$d_{crit}(T)=1-\frac{1}{\lambda^2_{max}(T)},$$

where $\lambda_{max}(T)$ is the maximum eigenvalue of the adjacency matrix of the tree.

Furthermore, Nagy [9] showed that for the cycle of order n and for the path of order $n+1$ the critical densities are equal.

Theorem 5 (Nagy [9]). *Let C_n be a cycle on n vertices and P_{n+1} be a path on $n+1$ vertices. Then we have*

$$d_{crit}(C_n)=d_{crit}(P_{n+1})=1-\frac{1}{4\cos^2\frac{\pi}{n+2}}.$$

We formulate the followig open problem.

Open problem: count the critical density $d_{crit}(H)$ for $H \in \mathcal{U}_{n,p}$, $p \geq 3$.

III. Inhomogeneous case: unicyclic graphs with the cycle C_3

In this section we study the inhomogeneous case when graph $H \in \mathcal{U}_{n,3}$, e.i. H is unicyclic with the cycle C_3 and for each edge $e \in E(H)$ the edge density γ_e is given. We extend some results presented in [5], where authors studied the inhomogeneous case for trees and proved the following theorem.

Theorem 6. *(Csikvári, Nagy [5]) Let T be a tree of order n and let v be a leaf of T. Assume that for each edge of T a density $\gamma_e = 1-r_e$ is given. Let $T^{'}$ be a tree obtained from T by deleting the leaf v and the edge uv, where u is the unique neighbour of v. Let the edge densities $\gamma_e^{'}$ in $T^{'}$ be defined as follows*

$$\gamma_e^{'}=\begin{cases}\gamma_e=1-r_e, & \text{if } e \text{ is not incident to } u,\\ 1-\frac{r_e}{1-r_{uv}}, & \text{if } e \text{ is incident to } u.\end{cases}$$

Then the set of densities $\{\gamma_e\}_{e\in E(T)}$ ensures the existence of the factor T if and only if all $\gamma_e^{'} \in (0,1]$ and the set of densities $\{\gamma_e^{'}\}_{e\in E(T')}$ ensures the existence of the factor $T^{'}$.

Theorem 6 provides authors of [5] with an efficient algorithm to decide whether a given set of edge densities in tree ensures the existence of a transversal or does not ensure. Their algorithm is presented below as **Algorithm** T for the completeness of our paper.

We extend the algorithm (**Algorithm** T) to the family of unicyclic graphs with the cycle C_3. The new algorithm (**Algorithm** $\mathcal{U}_{n,3}$) is based on the following Theorem 7 proved below by an extension of the method discovered in [5].

Theorem 7. *Let $H \in \mathcal{U}_{n,3}$ be a unicyclic graph of order n with the cycle C_3 and assume that for each edge $e \in E(H)$ a density $\gamma_e = 1-r_e$ is given. If the order of H is greater then 3, let v be a leaf of H and u be the unique neighbour of v, then let $H^{'}$ be a graph obtained from H by deleting the leaf v and an edge uv. Let the densities $\gamma_e^{'}$ in $H^{'}$ be defined as follows*

$$\gamma_e^{'}=\begin{cases}\gamma_e=1-r_e, & \text{if } e \text{ is not incident to } u,\\ 1-\frac{r_e}{1-r_{uv}}, & \text{if } e \text{ is incident to } u.\end{cases}$$

If the order of H is equal to 3 (i.e., H is isomorphic to C_3 with $V(H)=\{a,b,c\}$), then let $H^{'}$ be a graph obtained from H by deleting the vertex a and edges ab and ac. $H^{'}$ is a path P_{bc}. Let the density $\gamma_{bc}^{'}$ in $H^{'}$ be defined as follows

$$\gamma_{bc}^{'}=1-\frac{r_{bc}}{(1-r_{ab})(1-r_{ac})}.$$

Algorithm T

Step 0.
Let there be given a tree T^0 and edge densities γ_e^0. Set $T := T^0$ and $r_e = 1 - \gamma_e^0$.

Step 1.
Consider (T, r_e).

- **if** $|V(T)| = 2$ *and* $0 \leq r_e < 1$ **then**

  STOP: the densities γ_e^0 ensure the existence of a factor T^0.

- **if** $|V(T)| \geq 2$ *and there exists an edge for which* $r_e \geq 1$ **then**

  STOP: the densities γ_e^0 do not ensure the existence of a factor T^0.

Step 2.
if $|V(T)| \geq 3$ *and* $0 \leq r_e < 1$ *for all edges* $e \in E(T)$ **then**

DO pick a vertex v of degree 1 and let u be its unique neighbour. Let $T^{'} := T - v$ and

$$r_e^{'} = \begin{cases} r_e, & \text{if } e \text{ is not incident to } u, \\ \frac{r_e}{1-r_{uv}}, & \text{if } e \text{ is incident to } u. \end{cases}$$

Jump to *Step 1* with $(T, r_e) := (T^{'}, r_e^{'})$.

Then the set of densities $\{\gamma_e\}_{e \in E(H)}$ *ensures the existence of the factor* H *if and only if all* $\gamma_e^{'} \in (0, 1]$ *and the set of densities* $\{\gamma_e^{'}\}_{e \in E(H')}$ *ensures the existence of the factor* $H^{'}$.

Proof. Let $H \in \mathcal{U}_{n,3}$ and let the set of densities $\gamma_e = 1 - r_e$ be given for each $e \in E(H)$. First we prove the following statement: if all $\gamma_e^{'}$ are indeed densities and they ensure the existence of a factor $H^{'}$, then the original densities γ_e ensure the existence of a factor H.

Let $G[H]$ be a blow-up of the graph H such that the density between A_i and A_j is at least γ_{ij}, where A_i is a cluster of the vertex $i \in V(H)$. We show that it contains a factor H.

Let us consider a graph $H \in \mathcal{U}_{n,3}$ with $n > 3$ vertices.

Let $v, u \in V(H)$, where v is a leaf of H and $u \in N_H(v)$.

Define $R_{v,u}$ as the subset of A_u in the following way (see Fig. 6).

$$R_{v,u} = \{x \in A_u \mid x \text{ is incident to some edge between } A_u \text{ and } A_v\}.$$

Note that

$$|R_{v,u}||A_v| \geq e(R_{v,u}, A_v) = \gamma_{uv}|A_u||A_v|.$$

Hence

$$|R_{v,u}| \geq \gamma_{uv}|A_u|.$$

Now we show the lower bound for the number of edges incident to $R_{v,u}$. Let $k \in N_H(u)$. By the inclusion - exclusion formula we count the lower bound for the number of edges between $R_{v,u}$ and A_k as follows.

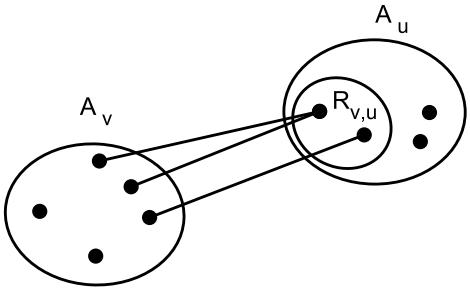


Fig. 6. Clusters A_v and A_u with the set $R_{v,u}$.

$$e(R_{v,u}, A_k) \geq e(A_u, A_k) - (|A_u| - |R_{v,u}|)|A_k| =$$
$$|R_{v,u}||A_k| + (\gamma_{ku} - 1)|A_k||A_u| \geq$$
$$|R_{v,u}||A_k| + (\gamma_{ku} - 1)\frac{1}{\gamma_{uv}}|R_{v,u}||A_k| =$$
$$\left(1 - \frac{r_{ku}}{1 - r_{uv}}\right)|R_{v,u}||A_k| = \gamma'_{ku}|R_{v,u}||A_k|.$$

Now, by deleting the vertex set A_v and $A_u \backslash R_{v,u}$ from $G[H]$, we obtain a graph which is a blow-up of $H^{'}$ with edge densities ensuring the existence of the factor $H^{'}$.

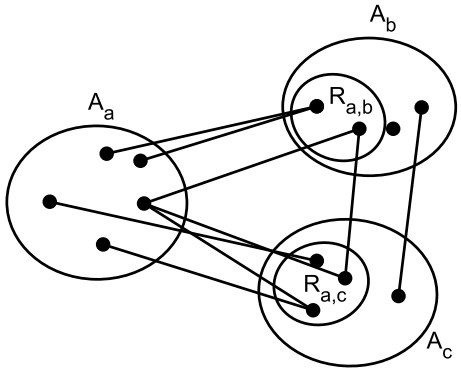


Fig. 7. Clusters A_a, A_b and A_c with the sets $R_{a,b}$ and $R_{a,c}$.

Moreover, by the definition of $R_{v,u}$ the factor $H^{'}$ can be extended to a factor H.

Now consider the situation when $n = 3$ and graph H is a cycle C_3 with the vertex set $\{a, b, c\}$. Let A_a be a cluster of vertex a. Define sets $R_{a,b}$ and $R_{a,c}$ in the following way (see Fig. 7).

$$R_{a,b} = \{x \in A_b \mid x \text{ is incident to some edge between } A_b \text{ and } A_a\},$$

$$R_{a,c} = \{x \in A_c \mid x \text{ is incident to some edge between } A_c \text{ and } A_a\}.$$

Note that

$$|R_b||A_a| \geq e(R_b, A_a) = \gamma_{ab}|A_a||A_b|,$$
$$|R_c||A_a| \geq e(R_c, A_a) = \gamma_{ac}|A_a||A_c|.$$

Hence we have the following lower bounds for the cardinalities of $R_{a,b}$ and $R_{a,c}$

$$|R_{a,b}| \geq \gamma_{ab}|A_b|$$

and

$$|R_{a,c}| \geq \gamma_{ac}|A_c|.$$

Next we show how many edges are incident to $R_{a,b}$ and $R_{a,c}$. Using the inclusion - exclusion formula we count the lower bound for the number of edges between $R_{a,b}$ and $R_{a,c}$

$$e(R_{a,b}, R_{a,c}) \geq e(A_b, A_a) - (|A_b| - |R_{a,b}|)|A_c| -$$
$$(|A_c| - |R_{a,c}|)|A_b| + (|A_b| - |R_{a,b}|)(|A_c| - |R_{a,c}|) =$$
$$|R_{a,b}||R_{a,c}| + (\gamma_{bc} - 1)|A_b||A_a| \geq$$
$$|R_{a,b}||R_{a,c}| + (\gamma_{bc} - 1)\frac{1}{\gamma_{ab}}\frac{1}{\gamma_{ac}}|R_{a,b}||R_{a,c}| =$$
$$\left(1 - \frac{r_{bc}}{(1 - r_{ab})(1 - r_{ac})}\right)|R_{a,b}||R_{a,c}| = \gamma^{'}_{bc}|R_{a,b}||R_{a,c}|.$$

Now, by deleting the vertex sets A_a, $A_b \backslash R_{a,b}$ and $A_c \backslash R_{a,c}$ from $G[C_3]$, we obtain a graph which is a blow-up of $C^{'}_3 = P_2$, $V(P_2) = \{b, c\}$, with edge densities ensuring the existence of the factor P_2. Moreover, by the definition of $R_{a,b}$ and $R_{a,c}$ the factor P_2 can be extended to a factor C_3.

Note that if

$$\gamma^{'}_{ku} < 0,$$

then

$$\gamma_{ku} + \gamma_{uv} < 1.$$

So there exists a construction which does not induce the path P_3 with the consecutive vertices k, u, v, where $i \in A_i$ ($i \in \{k, u, v\}$) in this case. Therefore, if some $\gamma^{'}_{ku} < 0$ then there exists a construction for a blow-up graph of H without a factor of H.

Next assume that all the $\gamma^{'}_e$ are proper densities, but there is a construction of a blow-up graph, say $G^{'}[H^{'}]$, with edge densities at least $\gamma^{'}_e$, but which does not induce a factor $H^{'}$. Thus we construct a blow-up $G[H]$ of the graph H not inducing H. We consider two possible cases. First let the picked vertex v be a leaf in H and $H^{'} = H - v$. Then set $A_u = \{u^*\} \cup A^{'}_u$ and $A_v = \{v\}$. We connect v to all elements of $A^{'}_u$ but do not connect to u^* without changing densities in $G^{'}[H^{'}]$ and with density γ_{vu} (see Fig. 8).

Now let $H^{'} = H - c$, where c is a vertex of C_3, $V(C_3) = \{a, b, c\}$. Then set $A_a = \{a^*\} \cup A^{'}_a$, $A_b = \{b^*\} \cup A^{'}_b$ and $A_c = \{c\}$. We connect c to all elements of $A^{'}_a$ and $A^{'}_b$ but do not connect to a^* and b^* without changing densities in $G^{'}[H^{'}]$ and with densities γ_{ca} and γ_{cb} (see Fig. 9). □

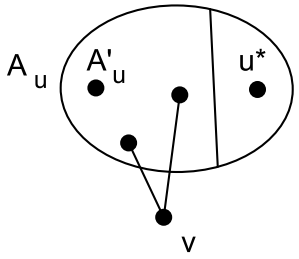


Fig. 8. We assume that $G^{'}[H^{'}]$ is without transversal $H^{'}$. The construction of the blow-up graph $G[H]$ without transversal H for the case where v is a leaf in H and $H^{'} = H - v$. The cluster $A^{'}_u$ is in $G^{'}[H^{'}]$. Let $A_u = A^{'}_u \cup u^*$ and $A_v = \{v\}$ be clusters in $G[H]$.

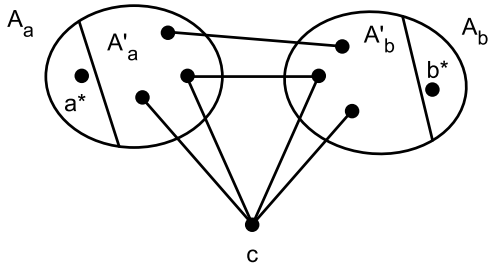


Fig. 9. We assume that $G^{'}[H^{'}]$ is without transversal $H^{'}$. The construction of the blow-up graph $G[H]$ without transversal H for the case where c is a vertex of C_3, $V(C_3) = \{a, b, c\}$ in H and $H^{'} = H - c$. The clusters $A^{'}_a$ and $A^{'}_b$ are in $G^{'}[H^{'}]$. Let $A_a = \{a^*\} \cup A^{'}_a$, $A_b = \{b^*\} \cup A^{'}_b$ and $A_c = \{c\}$ be clusters in $G[H]$.

Theorem 7 provides us with the algorithm (**Algorithm** $\mathcal{U}_{n,3}$) to decide whether a given set of edge densities ensures the existence of a transversal H in a blow-up graph $G[H]$ or does not ensure, where $H \in \mathcal{U}_{n,3}$.

For further considerations recall some results presented in papers [3] and [5]. First lemma gives condition on edge densities in the triangle C_3 which allows us to check if these densities ensure existing of C_3 in a blow-up graph $G[C_3]$. The second results gives condition on existing graph H as a factor in a blow-up graph $G[H]$ in terms of the multivariate matching polynomial F_H.

Lemma 1. *(Bondy, et al. [3]) Let α, β, γ be the edge densities between the clusters of a blow-up graph of the triangle - a cycle C_3. If*

$$\alpha\beta + \gamma > 1, \beta\gamma + \alpha > 1, \gamma\alpha + \beta > 1,$$

then the blow-up graph contains a triangle as a transversal.

Theorem 8. *(Csikvári, Nagy [5]) Assume that for the graph H we have*

$$F_H(\underline{r_e}, t) > 0$$

for all $t \in [0,1]$ and some vector $\underline{r_e}$ of weights, where $r_e \in [0,1]$ for each edge $e \in E(H)$. Then the densities $\gamma_e = 1 - r_e$ ensure the existence H as a transversal.

Let $H := C_3$ with vertices a, b, c and edge densities $\gamma_e = 1 - r_e$, where $e \in \{ab, bc, ac\}$. Assume that all $r_e \in [0,1)$ and run the **Algorithm** $\mathcal{U}_{n,3}$ by deleting vertex a from the graph C_3 with edges incident to it (means ab and ac). As a result we get a graph $H^{'}$ as a path $P_2 = bc$ with edge density

$$\gamma'_{bc} = 1 - r'_{bc} = 1 - \frac{r_{bc}}{(1 - r_{ab})(1 - r_{ac})}.$$

For $H^{'}$ we have

$$F_{H'}(\underline{r_e}, t) = 1 - tr'_{bc}.$$

By Theorem 8 we need

$$F_{H'}(\underline{r_e}, t) > 0$$

for $t \in [0,1]$.

Hence

$$\frac{1}{r'_{bc}} > 1,$$

$$(1 - r_{ab})(1 - r_{ac}) - r_{bc} > 0$$

and

$$\gamma_{ab}\gamma_{ac} + \gamma_{bc} > 1.$$

Similar inequalities are received when, instead of a vertex a, we delete in **Algorithm** $\mathcal{U}_{n,3}$ vertex b or c. As we can see we have a result presented in Proposition 1 consensual with Lemma 1.

Proposition 1. *Let a, b, c be vertices in a triangle C_3. Assume that $\gamma_e = 1 - r_e$ be an edge density assigned to each edge $e \in E(C_3)$, where $E(C_3) = \{ab, ac, bc\}$. If*

$$\frac{r_{ab}}{(1 - r_{ac})(1 - r_{bc})} < 1, \ \frac{r_{ac}}{(1 - r_{ab})(1 - r_{bc})} < 1$$

and $\dfrac{r_{bc}}{(1 - r_{ab})(1 - r_{ac})} < 1,$

then the set of densities $\{\gamma_e\}_{e \in E(C_3)}$ ensures existence of a transversal C_3 in a blow-up graph $G[C_3]$.

By running **Algorithm** $\mathcal{U}_{n,3}$ on some unicyclic graph $H \in \mathcal{U}_{n,3}$ with $\gamma_e = 1 - tr_e$ and using the multivariate matching polynomial $F_H(\underline{r_e}, s)$ we can prove the following lemma.

Lemma 2. *Let H be a weighted unicyclic graph of order $n > 2$ with the cycle C_3. Let $\gamma_e = 1 - tr_e$ be densities assigned to each edge $e \in E(H)$, where $r_e \in [0,1)$. Assume that after running Algorithm $\mathcal{U}_{n,3}$ we get a cycle C_3 with*

$$F_{C_3}(\underline{r_e}, t) = 0,$$

then t is a root of the multivariate matching polynomial $F_H(\underline{r_e}, s)$ of the graph H.

Proposition 2. *Let H be a weighted unicyclic graph of order $n > 2$ with the cycle C_3. Let $\gamma_e = 1 - tr_e$ be the densities assigned to each edge $e \in E(H)$. Assume that after running Algorithm $\mathcal{U}_{n,3}$ we get a cycle C_3 with the vertex set $V(C_3) = \{a, b, c\}$ and with $F_{C_3}(\underline{r_e}, t) = 0$ and, after restart Algorithm $\mathcal{U}_{n,3}$, we get a path P_2 (by deleting the vertex a and edges ab, ac), then*

$$F_{P_2}(\underline{r_e^{'}}, s) = \frac{t^2 r_{ab} r_{ac} + t r_{bc}}{(1 - tr_{ab})(1 - tr_{ac})} - s\frac{tr_{bc}}{(1 - tr_{ab})(1 - tr_{ac})}.$$

Proof. Assume that after running **Algorithm** $\mathcal{U}_{n,3}$ we get a cycle C_3 with edge densities $\gamma_e = 1 - tr_e$. Let $V(C_3) = \{a, b, c\}$ and $r_{ab}, r_{ac}, r_{bc} \in [0,1)$. The multivariate matching polynomial

$$F_{C_3}(\underline{r_e}, s) = 1 - s(r_{ab} + r_{ac} + r_{bc})$$

has exactly one root

$$t = \frac{1}{(r_{ab} + r_{ac} + r_{bc})}.$$

By deleting vertex a from the cycle C_3 with the edges $e_{ab} = ab$ and $e_{ac} = ac$ we obtain a path $P_2 = bc$. By Theorem 7 we get that

$$F_{P_2}(\underline{r_e^{'}}, s) = 1 - sr_{bc}^{'} = 1 - s\frac{tr_{bc}}{(1 - tr_{ab})(1 - tr_{ac})}.$$

By multiplying both sides by

$$(1 - tr_{ab})(1 - tr_{ac})$$

we have

$$(1 - tr_{ab})(1 - tr_{ac})F_{P_2}(\underline{r_e^{'}}, s) =$$

$$(1 - tr_{ab})(1 - tr_{ac}) - str_{bc} =$$

$$1 - t(r_{ab} + r_{ac} + r_{bc}) + tr_b c + t^2 r_{ab} r_{ac} - str_{bc}.$$

So

$$(1 - tr_{ab})(1 - tr_{ac})F_{P_2}(\underline{r_e^{'}}, s) - t^2 r_{ab} r_{ac} - tr_{bc} + str_{bc} =$$

$$F_{C_3}(\underline{r_e}, t) = 0.$$

Algorithm $U_{n,3}$

Input: a unicyclic graph $H \in \mathcal{U}_{n,3}$ with the set of edge densities $\{\gamma_e\}_{e\in E(H)}$.

Output: a boolean value

$$D = \begin{cases} TRUE, & \text{the densities } \gamma_e \text{ ensure the existence of a factor } H, \\ FALSE, & \text{the densities } \gamma_e \text{ does not ensure the existence of a factor } H. \end{cases}$$

Consider a weighted graph (H, r_e), where $r_e = 1 - \gamma_e$.

Step 1.

- **if** $|V(H)| = 2$ *(means H is a path P_2) and* $0 \le r_e < 1$ **then**

  STOP: D :=TRUE.

- **if** $|V(H)| \ge 2$ *and there exists an edge for which* $r_e \ge 1$ **then**

  STOP: D :=FALSE.

Step 2.

- **if** $|V(H)| = 3$ *(means H is a cycle C_3) and* $0 \le r_e < 1$ *for all edges* $e \in E(H)$ **then**

  pick a vertex c of the graph H and let a, b be its neighbours. Let $H' := H - c$ and

$$r'_{ab} = \frac{r_{ab}}{(1 - r_{ac})(1 - r_{bc})}.$$

- **if** $|V(H)| > 3$ *and* $0 \le r_e < 1$ *for all edges* $e \in E(H)$ **then**

  pick a vertex v of degree 1 and let u be its unique neighbour. Let $H' := H - v$ and

$$r'_e = \begin{cases} r_e, & \text{if } e \text{ is not incident to } u, \\ \frac{r_e}{1 - r_{uv}}, & \text{if } e \text{ is incident to } u. \end{cases}$$

Go to *Step 1* with $(H, r_e) := (H', r'_e)$.

Hence

$$F_{P_2}(\underline{r'_e}, s) = \frac{t^2 r_{ab} r_{ac} + t r_{bc}}{(1 - t r_{ab})(1 - t r_{ac})} - s \frac{t r_{bc}}{(1 - t r_{ab})(1 - t r_{ac})}.$$

By the definition of $F_{P_2}(\underline{r'_e}, s)$ we have

$$\frac{t^2 r_{ab} r_{ac} + t r_{bc}}{(1 - t r_{ab})(1 - t r_{ac})} = 1$$

and

$$t(r_{ac} + r_{ab} + r_{bc}) = 1.$$

Note that if

$$\gamma'_{bc} = 1 - \frac{t r_{bc}}{(1 - t r_{ab})(1 - t r_{ac})} = 0,$$

then

$$t r_{bc} = (1 - t r_{ab})(1 - t r_{ac})$$

and

$$\frac{t^2 r_{ab} r_{ac}}{(1 - t r_{ab})(1 - t r_{ac})} = 0,$$

$$t r_{ab} t r_{ac} = 0.$$

Therefore,

$$t(r_{ac} + r_{ab} + r_{bc}) = 1 + t r_{ab} t r_{ac} = 1.$$

So t is the root of $F_{C_3}(\underline{r_e}, t)$. □

From above consideration we deduce that **Algorithm** $\mathcal{U}_{n,3}$ works correctly with time complexity at most $\mathcal{O}(n^2)$. **Algorithm** $\mathcal{U}_{n,3}$ can be implemented in such a way that a vertex of the subgraph C_3 be considered (picked) in the last step of the algorithm.

IV. Conclusion

We have presented **Algorithm** $\mathcal{U}_{n,3}$ for testing whether the unicyclic graph $H \in \mathcal{U}_{n,3}$ with the set of edge densities $\{\gamma_e\}_{e\in E(H)}$ is a factor of a blow-up graph $G[H]$. Precisely, we have the answer whether the edge densities ensure the existence of the factor or do not ensure. In future work we will study the density Turán problem for an arbitrary graph of the family $\mathcal{U}_{n,p}$, $p \geq 4$, and for other families of graphs. Moreover, we wish to construct efficient algorithms for testing the existence of blow-up graphs with factors of the families.

Open problem: Look for the density Turán problem algorithm for families of connected graphs with blocks (i.e., 2-connected components) isomorphic to cycles and/or P_2 .

References

[1] R. Baber, J.R. Johnson and J. Talbot, The minimal density of triangles in tripartite graphs, *LMS J. Comput. Math.*, 13 (2010), 388–413, http://dx.doi.org/10.1112/S1461157009000436.

[2] B. Bollobás, Extremal Graph Theory, *Academic Press* (1978).

[3] A. Bondy, J. Shen, S. Thomassé and C. Thomassen, Density Conditions for triangles in multipartite graphs, *Combinatorica*, 26 (2006), http://dx.doi.org/10.1007/s00493-006-0009-y.

[4] W.G Brown, P. Erdös and M. Simonovits, Extremal problems for directed graphs, *Journal of Combinatorial Theory, Series B* 15 (1) (1973), 77–93, http://dx.doi.org/10.1016/0095-8956(73)90034-8.

[5] P. Csikvári and Z. L. Nagy, The density Turán Problem, *Combinatorics, Probability and Computing*, 21 (2012), 531–553, http://dx.doi.org/10.1017/S0963548312000016.

[6] Z. Füredi, Turán type problems, *Survey in Combinatorics* Vol. 166 of *London Math. Soc. Lecture Notes* (A.D. Keedwell, ed.) (1991), 253–300, http://dx.doi.org/10.1017/cbo9780511666216.010.

[7] C.D. Godsil and G. Royle, Algebraic Graph Theory, *Springer* (2001), http://dx.doi.org/10.1007/978-1-4613-0163-9.

[8] G. Jin, Complete subgraphs of r-partite graphs, *Combin. Probab. Comput.*, 1 (1992), 241–250, http://dx.doi.org/10.1017/s0963548300000274.

[9] Z.L. Nagy, A multipartite version of the Turán problem - density conditions and eigenvalues, *The Electronic Journal of Combinatorics*, 18 (2011), # P46.

[10] P. Turán, On an extremal problem in graph theory, *Mat. Fiz. Lapok*, 48 (1941), 436–452.

[11] R. Yuster, Independent transversal in r-partite graphs, *Discrete Math.*, 176 (1997), 255–261, http://dx.doi.org/10.1016/s0012-365x(96)00300-7.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 487–492
DOI: 10.15439/2014F52
ACSIS, Vol. 2

Routing on Dynamic Networks: GRASP versus Genetic

Benoit Bernay
Université Blaise Pascal
LIMOS CNRS Laboratory,
LABEX IMOBS3
Clermont-Ferrand 63000, France
Email: bernay@isima.fr

Samuel Deleplanque
Université Blaise Pascal
LIMOS CNRS Laboratory
LABEX IMOBS3
Clermont-Ferrand 63000, France
Email: deleplanque@isima.fr

Alain Quilliot
Université Blaise Pascal
LIMOS CNRS Lab.
LABEX IMOBS3
Clermont-Ferrand, France
Email : alain.quilliot@isima.fr

***Abstract*—We address here a large scale routing and scheduling transportation problem, through introduction of a flow model designed on a dynamic network. We deal with this model while using a master/slave decomposition scheme, and testing the behavior on this scheme of both a GRASP algorithm and a Genetic algorithm.**

I. Introduction

WE ALREADY introduced (see [9]), in the context of a partnership with an industrial player, a flow/multi-commodity flow model **FMS** which aimed at optimizing the management of a urban shuttle fleet. This model involved a dynamic network (see [2, 10]), that is a network with time indexed vertices, which made easy expressing temporal constraints. At this time we designed a GRASP algorithmic scheme, which allowed us handling a kind of large scale pre-emptive *Pick Up and Delivery* problem (see [10]), while using an ad hoc aggregation mechanism and performing random negative circuit cancelling.

We consider here the same model **FMS**, close to **CFA** (*Capacitated Flow Assignment*) models (see [1]) used in telecommunications, but we deal with it in a simpler way, while using an auxiliary cost vector as the master variable of a master/slave decomposition scheme. This scheme induces the design of resolution heuristics which mainly rely on simple shortest path procedures instead of complex negative circuit cancelling procedure, and whose generic features makes implementation easier. While next section II is devoted to a rough description of the **FMS** model, our main contribution is about the description in Section III of this master/slave decomposition scheme, from which we derive (sections IV and V) both a GRASP (*Greedy Random Adaptive Search Procedure*, see [5, 6]) algorithm, and a *genetic* algorithm (see [6, 7, 8]). We detail the way those algorithms are implemented, and test (Section VI) their respective behaviors.

II. The FMS Model

A. Main Notations and Definitions

A network G, with vertex set X and arc set E, is denoted by $G = (X, E)$. A *flow* vector is an arc indexed vector f with rational or integral values such that, for every vertex x, we have $\Sigma_{e \text{ enter into } x} f_e = \Sigma_{e \text{ comes out } x} f_e$ (*Kirchhoff Law*). The *arc support* of f is the arc subset $Arc\text{-}Supp(f) \subseteq E$, which contains all arcs $e \in E$ such that $f_e \neq 0$. A *multi-commodity flow* vector f is a flow vector collection $f = \{f(k), k \in K\}$. $Sum(f)$ is the *Aggregated Flow* $Sum(f) = \Sigma_{k \in K} f(k)$.

B. The Shuttle Problem (see [16])

We consider a *Urban Area* network $H = (Z, U)$: nodes of H mean either production sites $y_1,\ldots,y_m$ ($m = 7$ in the original application), or *residential* areas, and arcs mean elementary connections. A demand D_k, $k \in K$, is a 4-uple (o_k, d_k: *origin/destination nodes*, L_k: *Load*, t_k: *deadline*): L_k users have to be transported from o_k and to d_k (at least one of both nodes being an industrial node) while starting (arriving) after (before) time st_k (at_k). *Quality of Service* (QoS) requires this trip not to last more than T_k time units. Users alternatively walk and use a *shuttle* system; so, every arc e of H is endowed with a *walking* length $l_p(e)$ and with a *vehicle* length $l_v(e)$. Vehicles start from and end into a *Depot* node. Our goal is to route the shuttles while meeting the demands and minimizing both the number of vehicles (*Fixed Investment Cost*) and their running times (*Running Cost*). ***Route preemption* is allowed**: several vehicles may be involved in meeting a given demand.

C. The Dynamic Network H-Dyn.

We derive it from H by associating (see 2, 8, 10), with any node x of Z, (NP+1) copies of x, indexed from 0 to NP, which represent the states of x at the instants 0, δ,…, NP.δ;

δ is an elementary time unit, chosen between 3 mn and 6 mn in our application; *NP* is a parameter which fixes the planning period (between 2 and 3 h). We add 2 fictitious vertices *DP, DP** and set $X = \{x_r, x \in Z, r \in 0,\ldots, NP\} \cup \{DP, DP^*\}$. As for the arc set *E*, we round modulo δ the *vehicle* and *walking* lengths of any arc *u* in *U* by setting: $l^*_p(u) = \lceil l_p(u)/\delta \rceil$, $l^*_v(u) = \lceil l_v(u)/\delta \rceil$; then we define the labeled arc family *E* as containing:

- *wait* arcs (x_r, x_{r+1}), $x \in Z$, $r \in 0,\ldots, NP$-1 : such an arc is considered twice, with *walk* and *vehicle* labels;
- arcs $(DP, Depot_r)$, $(Depot_r, DP^*)$, $r \in 0,\ldots, NP$, with *vehicle* labels.
- arcs $(x_r, z_{r+l^*v(u)})$, $u = (x, z) \in U$, r such that $0 \le r \le NP - l^*_v(u)$, with *vehicle* label;
- *walk* arcs $(x_r, z_{r+l^*p(u)})$, $u = (x, z) \in U$, r such that $0 \le r \le NP - l^*_p(u)$, with *walk* label;
- a *backward* arc (DP^*, DP).

We denote by *A* the subset of *E* defined by the *vehicle* arcs. We provide, in a natural way, any arc *e* with an *Economical Cost* c_e and a *QoS Cost* p_e.

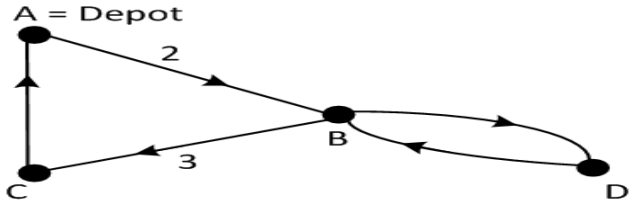


Fig. 1: Urban Transit Network *H* = (*Z*, *U*)

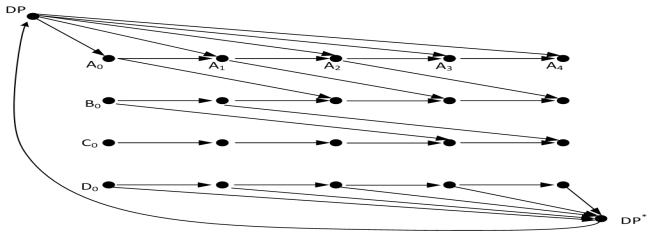


Fig. 2: Dynamic Network *H-Dyn* = (*X, E*).

Remark 1: *Size of H-Dyn*. We consider that (x, y) is an arc of *H* if a vehicle may move from *x* to *y* during time unit δ. If *H* contains 200 nodes, *H-Dyn* may contain up to 10^5 arcs.

D. The Flow/Multi-commodity Shuttle Model (FMS)

We want to route both vehicles and users. Aggregating vehicle routes yields, on the dynamic network *H-Dyn* = (*X, E*), some **integral flow vector *F***, and that user's routes may be represented as a **rational multi-commodity flow** $f = \{f(k), k \in K\} \ge 0$. Measuring *f* in such a way the capacity of any vehicle becomes equal to 1 yields the following **FMS**: *Flow/Multi-Commodity Flow Shuttle* model:

FMS: *Flow/Multi-Commodity Flow Shuttle* Model.
Input: The *Urban Area* network *H* = (*Z*, *U*), and the discrete time space $\{0,.., N\delta\}$;
Output: Compute, on the dynamic network *H-Dyn* an integral flow vector *F*, and a **rational multi-commodity flow** $f = \{f(k), k \in K\} \ge 0$, such that:

- *F* is null on the *walk* arcs ;
- For any $k \in K$, $f(k)$ routes L_k users from o_k to d_k between time st_k and time at_k;
- $g = Sum(f)_e = \Sigma_{k,e} f(k)_e \le F_e$, for any arc of *H-Dyn* with *vehicle* label;
- $Cost(F) + QoS(f) = c.F + p.Sum(f) = \Sigma_{e \text{ in } A} c_e.F_e + \Sigma_{e \text{ in } E} p_e.Sum(f)_e$ is minimal

Conversely, if *F* and *f* satisfy **FMS** constraints and if *Route Preemption* is allowed, then they yield a feasible solution of the *Shuttle* Problem. Our model casts *temporal constraints* into the construction of the network *H-Dyn*. We denote by $\mathbf{FMS}_g$ the time-polynomial min cost integral flow problem which derives from **FMS** by fixing $g = Sum(f)$.

Remark 2: ***FMS Model Size***. If the number of demands is 250, then the size of *f* may be up to $25\ 10^6$: the resulting **FMS** model is a large scale NP-Hard MIP problem.

III. FUNDAMENTAL TOOLS

Before describing algorithm, we need to specify which objects and procedures they will involve.

A.A Master/Slave Encoding of a FMS Solution

The quality of a **FMS** solution relies on its ability to make users share vehicles. While the size *H-Dyn* may eventually be very large, the number of arcs which are going to support non null *F* and $g = Sum(f)$ is comparatively small. So, a key object in our model should be the *arc support set* $A = Arc\text{-}Supp(F) = \{e \in E$ such that $F_e \ne 0\}$ of *F*. The following theoretical result, whose proof can be obtained through standard mathematical programming techniques, will help us in dealing with this *arc support set*:

***Dualization* Theorem**: *Let (F, f) an optimal solution of the **FMS** model. Then there exists a price vector $\mu \ge 0$, with indexation on the arc set of H-Dyn, such that:*

- *$\mu_e = 0$ for any arc e which is a walk arc or a wait arc and which is not in A;*
- *$\mu_e \ge c_e$ for any arc e in A; $\mu_e = +\infty$ for any vehicle arc e which is not in A;*
- *Every flow f(k) is an optimal solution of the min cost flow problem defined by: (E1)*
  - *for every $k \in K$, f(k) routes load L_k f(k) from o_k to d_k between time st_k and time at_k;*
  - *$\Sigma_e (\mu_e + p_e).f(k)_e$ is the smallest possible.*

So, the knowledge of both *arc support set* A and cost vector μ allow us to derive, through shortest path procedures, the aggregated flow $g = Sum(f)$. Flow vector F is computed as a solution of $\mathbf{FMS}_g$. We impose every vector $f(k)$, $k \in K$, to be routed along a single path. So, a well-fitted representation of a **FMS** solution is given by:

- the *set Arc-Supp*$(F) = \{e \in E$ such that $F_e \neq 0\}$;
- the related cost vector $\mu = \mu_e$, $e \in A$.

Those objects define the *Master* part of such a solution, whose *Slave* part is defined by F and the collection $\Gamma(k)$ of the paths followed by the flow vectors $f(k)$, k in K.

B. Dealing with the FMS_g Problem

We deal with $\mathbf{FMS}_g$ through *column generation*, while using an arc/path formulation of $\mathbf{FMS}_g$:

> $\mathbf{FMS}_g$:{Λ denotes the set of paths from *DP* to *DP**;
> Compute a vector $G = (G_\gamma, \gamma \in \Lambda) \geq 0$, with rational values, such that:
> - for any arc e of *H-Dyn* with *vehicle* label, $\Sigma_{\gamma \text{ such that } e \in \gamma} G_\gamma \geq \lceil g_e \rceil$;
> - $\Sigma_\gamma Cost(\gamma).G_\gamma$ is minimal}

If Λ_0 is some *active* subset of Λ, and if $\lambda = (\lambda_e, e$ in the arc subset of *H-Dyn* with *vehicle* label$) \geq 0$ is a dual solution of the restriction of $\mathbf{FMS}_g$ to Λ_0, then the related *Pricing* (search for the new entering column) sub-problem is as a largest path problem, handled by Bellman algorithm. So, when dealing with the $\mathbf{FMS}_g$ problem, we do in such a way that we are always provided with some current *active* path subset Λ_0 of Λ, which evolves in an incremental way.

C. Deriving the paths $\Gamma(k)$ from A and μ.

Dualization Theorem tells us that support set A and cost vector μ should identify the arcs along which users are going to share a same vehicle. If A and μ were conveniently chosen, paths $\Gamma(k)$, $k \in K$, should be shortest paths for cost vector $(p + \mu)$. So, all throughout the execution of our processes, we derive paths $\Gamma(k)$, $k \in K$, as shortest paths for the following cost vector $C^{A,\mu}$:

- If e is a *walk* other *wait* arc which is not in A, then $C^A_e = p_e$;
- If e is in A, then $C^A_e = \mu_e + p_e$;
- Else $C^A_e = p_e + c\mu^*$, where $\mu^* = \text{Max}_{e \in A} \mu_e/c_e$. (E2)

As a matter of fact, for a given *vehicle* arc $e \notin A$, we apply (E2), which means that we want to keep paths $\Gamma(k)$, $k \in K$ from using vehicle arcs which are not in A, only when no path $\Gamma(k)$, $k \in K$ involves e. Else, we use μ^* defined by:

$\mu^* = \text{Mean Value}_{e \in A} \mu_e$.

D. A Randomized Initialization

This initialization procedure **FMS-INIT** works through successive insertions of demands D_k, $k \in K$, into a current aggregated flow vector g:

> **FMS-INIT Procedure**:
> - g: current aggregated flow vector; K_0: set of inserted demands;
> - F and λ: primal and dual solutions of $\mathbf{FMS}_g$;
> - Λ_0 = set of *active vehicle* paths;
>
> While $K - K_0$ is not empty do
>
> Randomly Pick up $k \in K - K_0$ and *Insert* it into K_0: route demand k according to some path $\Gamma(k)$ in *H-Dyn*, in such a way that: (I1)
> - $\Gamma(k)$ connects o_k to d_k, while satisfying related temporal constraints;
> - the induced increase in the cost $\lambda.\lceil g \rceil + p.g$ is the smallest possible ;
>
> Update F, λ and Λ_0.
>
> Set $A = Arc\ Support$ of F; For every arc e in A, set:
> $\mu_e = \lambda_e.\lceil g_e \rceil$. (I2)

The above *Insert* instruction (I1) is handled by a *shortest path* Bellman-like Algorithm.

E. Local Transformation and Mutation Operators

The **FMS-INIT** previous process gives rise in a generic way to a local transformation operator *TRANS*, which acts on a current solution A, μ, F, $\Gamma = \{\Gamma(k), k \in K\}$ as follows:

> **Local Operator *TRANS***(K_0: K_0 subset of K)
> - Randomly select $K_0 \subseteq K$ and withdraws paths-flows $\{f(k), k \in K_0\}$ from g; Update flow vector F;
> - Reinsert demands D_k, $k \in K_0$, according to the III.D, while starting from current partial solution (F, g);
> - Consequently update A and μ.

Operator **TRANS** will be used here in both GRASP scheme, according to a *Descent* strategy and in a genetic meta-heuristic scheme, as a *mutation* operator.

F. Crossover Operator

Given two feasible **FMS** solutions A_1, μ_1, F_1, $\Gamma_1 = \{\Gamma_1(k), k \in K\}$ and A_2, μ_2, F_2, $\Gamma_2 = \{\Gamma_2(k), k \in K\}$. *SON-CREATE* derives children (A, μ) and (A', μ') as follows:

Crossover Operator SON-CREATE:

For every arc e in $(A_1 \cap A_2)$, insert e into both A and A' and randomly assign related value μ_e or μ'_e with one of both values $(\mu_{1,e} + \mu_{2,e})/2$ and $(3.\mu_{1,e} - \mu_{2,e})/2$;
For every arc e in $(A_1 - A_2) \cup (A_2 - A_1)$, randomly insert e into either A or A' and randomly assign related value μ_e or μ'_e with one of both values $(\mu_{1,e} + \mu_{2,e})/2$ or $(3.\mu_{1,e} - \mu_{2,e})/2$;
Compute path collections $\Gamma = \{\Gamma(k), k \in K\}$ and $\Gamma' = \{\Gamma'(k), k \in K\}$ as in III.C; Compute F and F' flow vectors as in III.B, together with dual vectors λ and λ';
Update cost vectors μ and μ' according to (I2).

IV. A GRASP ALGORITHM FMS-GRASP FOR FMS

A GRASP: *Greedy Random Adaptive Search Procedure* (see [5, 6]) algorithmic scheme works by performing first a greedy randomized initialization process, and next a descent loop. It may be run according to several replications, either in a sequential or in a parallel mode. Here, we get:

FMS-GRASP(*R*: *Replication Number*, *Q*: *Subset Size*, *Loop*: *Loop Length Bound*);
For i = 1..*R* do
  Initialize A, μ, F, $\Gamma = \{\Gamma(k), k \in K\}$ through **FMS-INIT**; *Possible*;
  While *Possible* do (I3: *Descent* loop)
    Modify A, μ, F, $\Gamma = \{\Gamma(k), k \in K\}$ in such a way cost $c.F + p.g$ is improved;
    If *Failure*(*Modify*) then Not *Possible*;
The result of **FMS-GRASP** is the best A, μ, F, $\Gamma = \{\Gamma(k), k \in K\}$ ever obtained.

(I3) involves the *TRANS* operator as follows:

Possible;
While *Possible* do
  Trial-Number <- 1; *Success* <- *False*;
  Do Until *Success* or *Trial-Number* > *Loop*
    Generate $K_0 \subseteq K$ with cardinality Q; Save current A, μ, F, $\Gamma = \{\Gamma(k), k \in K\}$; (I4)
    Apply $TRANS(K_0)$ to A, μ, F, Γ; If $c.F + p.g$ is improved then *Success* Else
      Restore A, μ, F, Γ;
      Trial-Number <- *Trial-Number* + 1;
  Possible <- *Success*;

Choosing K_0 in the (I4) Instruction: it is defined by the paths $\{\Gamma(k), k \in K\}$ which contain the arcs e with the highest $(\mu_e + p_e)$ values.

A *Random Walk* Variant of FMS-GRASP: Because of the computing costs induced by Instruction (I4), we also implement a *Random Walk* strategy:

FMS-GRASP-1(*R*: *Replication Number*; *RW*: *Loop Length Bound; Q: Subset Size*);
For i = 1..*R* do
  Initialize A, μ, F, $\Gamma = \{\Gamma(k), k \in K\}$ through **FMS-INIT**;
  For *Counter* = 1..*RW* do (I4.1: *Random Walk* loop)
    Generate $K_0 \subseteq K$ with cardinality Q;
    Apply $TRANS(K_0)$ to A, μ, F, Γ;
The result is the best (A, μ, F, Γ) ever obtained.

V. A GENETIC ALGORITHM FMS-GEN FOR FMS

The main components of a *Genetic* algorithm are (see [6, 7, 8]): its *Encoding* scheme (*Chromosome* Representation); the *Initialization* Procedure which yields the initial *population* Σ; its *Mutation* operator; its *Crossover* operator.

Clearly, the *Encoding* scheme is the encoding scheme of Section III.A whose *master* objects are:
- the *arc support set Arc-Supp*$(F) = \{e \in E$ such that $F_e \neq 0\}$ of F;
- the related cost vector $\mu = \mu_e$, $e \in A$;

and the *slave* objects are the flow vector F and the path collection $\Gamma = \{\Gamma(k), k \in K\}$.
Initialization is performed through Card(Σ) successive applications of **FMS-INIT**.
Mutation results from application of the operator *TRANS*, with parameter K_0 generated with a given cardinality Q, Q becoming a parameter of the global process:

FMS-Mutation(A, μ, F, $\Gamma = \{\Gamma(k), k \in K\}$, Q);
  Generate some subset K_0 of K with cardinality Q;
  Apply $TRANS(K_0)$ to A, μ, F, Γ;

The *FMS-Crossover crossover* operator is the **SON-CREATE** operator of Section III.F.

What remains to be discussed here is the *Fitness* Criterion, and the way *FMS-Crossover* is applied:
- Given $(A_1, \mu_1, F_1, \Gamma_1 = \{\Gamma_1(k), k \in K\})$ and $(A_2, \mu_2, F_2, \Gamma_2 = \{\Gamma_2(k), k \in K\})$ in current population Σ, *Fitness* is related here to the cardinality of the difference set $(A_1 - A_2) \cup (A_2 - A_1)$: the smallest is it, the largest is the *Fitness* measurement;
- Best-fitted pairs are selected, in order to avoid *cloning*, with the constraint that no solution σ belongs to more than 2 pairs. It is done in a heuristic way.

The main parameters of the deriving Genetic Algorithm **FMS-GEN** are the *population* size P, the number LG of iterations of *mutation/crossover* process and the *size* Q.

VI. NUMERICAL EXPERIMENTS

Experiments are performed on a LINUX server CentOS 5.4, Processor Intel Xeon 3.6 GHZ, with help of the CPLEX 12 library.

A. Instance Generation

An instance is defined by: the *Urban Area* network $H = (Z, U)$, with n vertices and m arcs; *demands* D_k, $k \in K$; *walking* lengths $l_p(e)$, and *vehicle* lengths $l_v(e)$, $e \in U$; *Vehicle* cost vector c and *User* cost vector p; the size NP of the time-space; the arc number NA of *H-Dyn*. We generate our own small and large instances: nodes of H are points of the 2D Euclidean space, with adjacency related to distance thresholds; *demands* D_k, $k \in K$ are randomly generated through uniform distribution.

B. Evaluation of FMS-INIT.

We first consider small instances, for which we get an exact optimal value through the CPLEX.12 Library, and next consider larger size instances, with focus on the large scale issue. In both cases:

- n, m are respectively the node and arc numbers of H, NP is the period number, NOD is the number of demands; L is the mean value of loads L_k, $k \in K$, and α is the mean ration p_e/c_e, $e \in E$.
- R is the replication number of **FMS-INIT**.

Small instances: We focus on precision of **FMS-INIT**, and test packages of 10 instances. *GAP-MEAN* is the mean error $GAP = (VAL - OPT)/OPT$: VAL = cost value computed by **FMS-INIT**, OPT = optimal value computed by CPLEX.12. *GAP-VAR* is the variance of *GAP*. We get:

R	*GAP-MEAN*	*GAP-VAR*
1	22.5	25.8
5	17.3	20.7
10	13.4	17.9
20	11.9	15.4
50	10.5	13.6

Table 1: **FMS-INIT** Evaluation, Small Instances, 10 instances/packages; **Group-Instance***: n = 10, m = 30, NP = 10; NOD = 10; α = 0.5; L = 0.2;* Impact of R.

Analysis: Parameter R plays a key role. *GAP-VAR* is usually large: a single **FMS-INIT** run may yield poor solutions.

Large instances: We focus here on CPU times and on the sensitivity to parameter R: $V(R)$ is the mean value $(Max(R) - Min(R))/Min(R)$, where $Max(R)$ and $Min(R)$ are respectively the worse and best values obtained through **FMS-INIT**(R), while *Var-V*(R) is the related variance. *Mean-CPU* is the mean running time, while *Var-CPU* is the related variance.

R	*V*(*R*)	*Var-V*(*R*)	*Mean-CPU (in s)*	*Var-CPU*
1	0.0	0.0	24.8	12.5
5	0.08	0.05	98.5	67.2
10	0.14	0.04	177	101.0
20	0.18	0.04	329	151.4
50	0.22	0.03	745	256.1

Table 2: **FMS-INIT** Evaluation, Large Instances, 10 instances/packages; **Group-Instance***: NA = 66256; NOD = 100; α = 0.5; L = 0.2;* Impact of R.

Analysis: The replication mechanism is crucial.

C. Evaluation of FMS-GRASP.

We focus on the respective ability of the standard Descent loop with parameter *TH* and of the random walk with parameter *RW* to improve the initial solution.

Small instances, 10 instances/packages with n = 10, m = 30, NP = 10, NOD = 10, L = 0.2; α = 0.5: *GAP-MEAN* is the mean error $GAP = (VAL - OPT)/OPT$, where VAL is computed by **FMS-GRASP**, and *GAP-VAR* is the variance of *GAP*. We use R = 10, Q = 3.

TH	*GAP-MEAN*	*GAP-VAR*
1	13.1	17.8
4	11.3	11.0
8	9.4	9.5
15	7.2	8.3

Table 3: **FMS-GRASP/Descent**: Impact of *TH*

RW	*GAP-MEAN*	*GAP-VAR*
1	13.3	17.8
4	11.5	13.1
10	9.3	11.8
40	6.8	8.8
80	3.8	5.1

Table 4: **FMS-GRASP/Random Walk**: Impact of *RW*

Analysis: *Random Walk* is more efficient than *Descent*.

Large instances, 10 instances/packages, with NA = 66256; NOD = 100; α = 0.5; L = 0.2. $IMPROVE = (Min(R) - Val(R, RW))/Val(R,W)$, where $Min(R)$ is computed by **FMS-INIT**(R) and $Val(R, W)$ is computed by **FMS-GRASP**(R, RW). $IMPROVE(R, RW)$ is the mean *IMPROVE* Value.

R	*RW*	*IMPROVE*(*R*, *RW*)	*Mean-CPU*
1	5	2.3	39
1	100	9.8	249
5	5	1.7	151
5	100	9.1	1012
10	5	1.4	257
10	100	6.7	1618

Table 5: **FMS-GRASP** Evaluation, Large Instances: Impact of R and RW.

Analysis: Computing times remain under control. Improvement margin induced by the *Random Walk* loop are close to the values obtained for small instances.

D. Evaluation of FMS-GEN.

We use the same tests as in Section VI.C. *P* is the size of the population, *LG* is the length of the main loop of the process. The population Σ is initialized by **FMS-INIT**(*P*).

Small instances: 10 instance packages with $n = 10$, $m = 30$, $NP = 10$; $NOD = 10$; $L = 0.2$; $\alpha = 0.5$; *GAP-MEAN* is the mean error GAP = (*VAL* – *OPT*)/*OPT*, where *VAL* is the cost value of the solution which is computed by **FMS-GEN**, and *OPT* is the optimal result computed by CPLEX.12 *GAP-VAR* is the variance of *GAP*. We focus on difficult instances, and deal with rather small populations (no more than 30) and small *LG* values. We use $P = 10$, $Q = 3$; $\Pi = 1$;

P	*GAP-MEAN*	*GAP-VAR*
4	10.3	14.7
10	6.2	8.0
20	4.2	5.4
30	3.8	4.5

Table 6: **FMS-GEN** Evaluation, Impact of *P*.

LG	*GAP-MEAN*	*GAP-VAR*
10	9.3	11.7
20	9.1	11.4
50	6.2	7.0
100	2.9	4.1

Table 7: **FMS-GEN** Evaluation, Impact of *LG*.

Large instances: For any instance, we evaluate the improvement ratio *IMPROVE* = (*Min*(*P*) - *Val*(*P, LG*))/*Val*(*P,LG*), where *Min*(*P*) is the value obtained while running **FMS-INIT**(*P*) and *Val*(*P, LG*) is the value obtained while running **FMS-GEN**(*P*, *LG*). *IMPROVE*(*P*, *LG*) is the mean *IMPROVE* value on 5 instance package defined by parameter values: *NA* = 66256; *NOD* = 100; $\alpha = 0.5$; $L = 0.2$. We use $Q = 15$ and $\Pi = 1$.

P	*LG*	*IMPROVE*(*P*, *LG*)	*Mean-CPU*
4	10	4.6	451
4	20	7.2	828
4	50	9.6	1830
10	10	3.3	1296
10	20	5.8	2265
10	50	7.3	4520

Table 8: **FMS-GEN** Evaluation, Large Instances, Impact of *P* and *LG*.

General comment: The GRASP scheme is less accurate the the GA scheme, but it is more flexible and tackles more easily large scale instances. When it comes to practical applications, accuracy is not such an issue. So it comes that we may consider here that, from this point of view, GRASP performs better.

VII. CONCLUSION

Reformulating the **FMS** model through through implicit representations allows us to design efficient GRASP and genetic algorithms. Still, we notice that since those algorithms rely on sophisticated LP techniques, we should now study the way to efficiently involve recently emerging generic framework, like ILP software SCIP/CPLEX, in such a way development and maintenance costs be minimized.

REFERENCES

1. AHUJA. R.K, MAGNANTI. T.L, ORLIN. J.B, REDDY. M.R: *Applications of network optimization*; Chap. 1 **Network Models, Handbook O.R & Manag. Sci.** 7, p 1-83, ISBN 013617549X, (1995).
2. ARONSON. J.E: *A survey on dynamic network flows* ; **Ann. Op. Res.** 20, p 1-66, DOI 10.1007/BF02216922, (1989).
3. CORDEAU. J.P, TOTH. P, VIGO. D: *A survey of optimization models for train routing and scheduling*; **Transportation Science** 32, p 380-404, DOI 10.1287/trsc.32.4.380, (1998).
4. CRAINIC. T, .GENDREAU. M, FARVOLDEN. M: *A simplex based Tabu search method for network design*; **INFORMS Journal on Computing** 12, p 223-236, DOI 10.1287/ijoc.12.3.223.12638, (2000).
5. RESENDE. M, RIBEIRO. C: *Greedy Random Adaptive Procedure*, Handbook of Metaheuristics, Int. Series on O.R and Manageent Sciences, 146, p 283-319, DOI 10.1007/978-1-4419-1665-5_10, (2002).
6. EL GHAZALI. T: **Metaheuristics from Design to Implementation**, *Wiley Interscienc*e, ISBN 978-0-470-49690-9 (2009).
7. REEVES C.R: *Genetic algorithms for the operations researcher*; **INFORMS Journal of Computing** 9, 3, p 231-250, DOI 10.1007/0-306-48056-5_3, (1997).
8. ANGELOVA. M, ATANASSOV. K, PENCHEVA. T: *Purposeful model parameter genesis in simple genetic algorithms*; **Computer and Mathematics with Applications** 64, p 221-228, DOI 10.1016/j.camva.2012.01.047, (2012)
9. QUILLIOT. A, LIBERALINO. H, BERNAY.B:: *Large Scale Multi-Commodity Flow Handling on Dynamic Networks*, **Proc. LSSC 2013**, Szozopol, Bulgaria, to appear in LNCS 8353, Springer, (2013).
10. BORNDORFER. R, GROTSCHEL. M, LOBEL. A: *Optimization of transportation systems*, **Konrad-Zuse-Centrum Information Technik Berlin**, Report 98-09, (1998).

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 493–500
DOI: 10.15439/2014F50
ACSIS, Vol. 2

Exact and Approximation Algorithms for Linear Arrangement Problems

Alain Quilliot
LIMOS CNRS UMR 6158
LABEX IMOBS3
Université Blaise Pascal
Bat ISIMA, BP 10125
Campus des Cézaux,
63173 Aubière, France
Email: alain.quilliot@isima.fr

Djamal Rebaine
UQAC
Département d'Informatique
Chicoutimi, Saguenay, Quebec
Canada
Email: Djamal.Rebaine@uqac.ca

***Abstract*—We present here new results and algorithms for the *Linear Arrangement* Problem (LAP). We first propose a new lower bound, which links LAP with the *Max Cut* Problem, and derive a LIP model as well as a branch/bound algorithm for the general case. Then we focus on the case of interval graphs: we first show that our lower bound is tight for unit interval graphs, and derive an efficient polynomial time approximation algorithm for general interval graphs.**

I. Introduction

LET G = (X, E) be a non oriented graph where X and E respectively denote the vertices and edges of G. The *Linear Arrangement* Problem (LAP) consists of finding a one-to-one mapping ϕ from X to $\{1, \ldots, |X|\}$ that minimizes:

$$f(G,\varphi)=\sum_{(x,y)\in E}|\varphi(y)-\varphi(x)|$$

The LAP problem has applications (see [3, 11]) in Information Retrieval and Industrial Storage, and may also appear as a sub-problem of some Network Design models (see [4]). It is, even in practice, a very difficult combinatorial optimization problem. The corresponding decision LAP was first shown to be *NP*-complete for arbitrary graphs (see for example [8, 9]) and next, for interval graphs [5] and bipartite graphs [9]. However, polynomial time algorithms were designed for trees [4], unit interval graphs [6], paths, cycles, complete bipartite graphs, grid graphs [11] and restricted series-parallel graphs [1]. A survey is available in [3].

Since LAP is *NP*-hard, even for graphs which usually turn most difficult problems into time-polynomial ones, the heuristic approach is therefore justified to deal with it. So the goal of this theory oriented study is to provide tools for the design of exact and approximation LAP algorithms with some focus on interval graphs. The paper is organized as follows. Section 2 introduces an linear ordering based reformulation of LAP. In Section 3, we propose a general lower bound, which links LAP with the well-known *Max Cut* Problem, and next derive, in Section 4 and 5, an ILP model together with a branch/bound algorithm. Finally, in section 6 we restrict our study to the case of interval graphs and propose an approximation algorithm, whose efficiency is briefly tested in Section 7.

II. Notations, definitions, LAP reformulation

A simple (non oriented) graph with no loop is denoted by G = (X, E): X (E) is the node (edge) set of G. We denote by (x, y) = (y, x) an edge with end-nodes x and y in X. If $A \subseteq X$, then G_A is the sub-graph induced by A from G. If $x \in X$, then $\Gamma_G(x) = \{y \in X$ such that $(x, y) \in E\}$ is the *neighbour* set of x. The complementary graph $G^c = (X, E^c)$ of G is defined by: $E^c = \{(x, y)$ such that (x, y) ¿ E and $x \neq y\}$. A *triangle* of G is a clique with 3 nodes. An *anti-edge* is a pair $e = (x, y) = (y, x)$, $x \neq y$, such that $(x, y) \notin E$. A *fork* with *root* x is any (non oriented) triple $f = \{x, y, z\} = \{x, z, y\}$ such that $(x, y), (x, z) \in E$, and $(y, z) \notin E$. An *anti-fork* with *root* z is any triple $f = \{x, y, z\} = \{y, x, z\}$ such that $(x, y) \in E$ and $(x, z), (y, z) \notin E$.

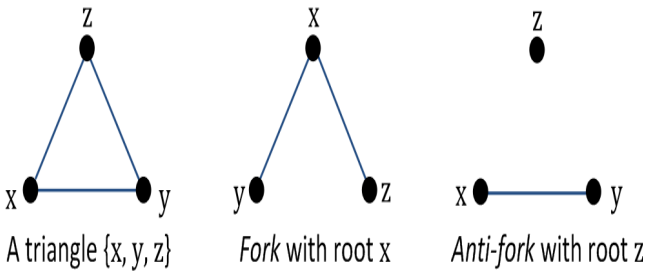


Figure 1: Triangles, Forks and Anti-Forks

LAP Reformulation: A *linear ordering* of a set X is a binary order relation σ such that, for any pair x, y in X, $x \neq y$, we have either $x \sigma y$ or $y \sigma x$. Given a graph G = (X, E) and a linear ordering σ of X. For any edge e = (x, y), we set BE(e, z, σ) = (*elementary break* of e by z according to σ) = 1 if $x \sigma z \sigma y$ or $y \sigma z \sigma x$, and 0 otherwise. We set BG(G, σ) = (*Global Break* of G according to σ) = $\Sigma_{e,z}$ BE(e, z, σ). If $\phi(\sigma)$ is the one-to-one mapping from X into $\{1, \ldots, |X|\}$ which derives from σ, then: $f(G, \phi(\sigma)) = \Sigma_{(x,y) \in E} |\phi(y) - \phi(x)| = BG(G, \sigma) + |E| = \Sigma_{e,z}$ BE(e, z, σ) + |E|.

So, solving LAP means seeking σ that minimizes the *Global Break* $BG(G, \sigma)$: we denote by LAP(G) the related optimal value $Inf_{\sigma} BE(G, \sigma)$.

III. Linking Max Cut and LAP: A General Lower Bound

Computing a good linear ordering σ of the vertices of a graph $G = (X, E)$ means efficiently deciding, for any $x \in X$, which vertices of the *neighbour* set $\Gamma_G(x)$ are located before x according to σ. This local decision process may be linked to the well-known *Max Cut* Problem [2,7,10]:

> ***Max Cut* Problem**: Let $H = (Z, F)$ be a simple graph. We denote by $Z = A \cup^{Ex} B$ any partition of Z into 2 disjoint subsets, and *Cut*(A, B) the number of edges of H with one extremity into A and the other into B. Solving the *Max Cut* Problem means seeking a *partition* $Z = A \cup^{Ex} B$ that maximizes *Cut*(A, B). We denote by *Max-Cut*(H) the related optimal value.

Let us consider now some graph $G = (X, E)$. We denote by *Tr*(G) the number of triangles of G. For any vertex x in X, we set:

- $H(x)$ = the complementary graph of the sub-graph of G which is induced by $\Gamma_G(x)$. Note that x is not a node of the graph H(x), since G has no loop;
- $m(x)$ = the number of edges of H(x); $V(x) = m(x) - Max\text{-}Cut(H(x))$.

Theorem 1: *For any graph G = (X, E), we have:* $LAP(G) \geq Tr(G) + \Sigma_x V(x)$.

Proof: Let us consider a linear ordering σ of G, and set:

- $Fk(G, \sigma)$ = number of *forks* f = {x, y, z}, f with root x, of G, such that $((x \sigma y) \wedge (x \sigma z)) \vee ((y \sigma x \wedge z \sigma x))$.
- $AFk(G, \sigma)$ = number of *anti-forks* f = {x, y, z} of G, f with root z, such that $(x \sigma z \sigma y) \vee (y \sigma z \sigma x)$.

Let us first check that: $BG(G, \sigma) = Tr(G) + Fk(G, \sigma) + AFk(G, \sigma)$. (E1)

In order to do so, we consider an edge e = (x, y), and a node z, different from x and y. While counting $BE(e, z, \sigma)$, we consider three cases:

Case 1: x, y and z define a triangle. Then $BE(e, z, \sigma) = 1$ if either $x \sigma z \sigma y$ or $y \sigma z \sigma x$. In such a case no quantity $BE((x, z), y, \sigma)$, $BE((z, y), x, \sigma)$ is equal to 1. So, if x, y, z define a triangle, there exists exactly one node t in{x, y, z} such that $BE(e(t), t, \sigma) = 1$, where e(t) is the edge which is defined by {x, y, z} – t. We get $\Sigma_{e = (x, y), z \text{ such that } (x, y, z) \text{ is a triangle}} BE(e, z, \sigma) = Tr(G)$.

Case 2: f = {x, y, z} is a *fork* with root x. Then $BE((x, y), z, \sigma) = 1$ if either $x \sigma z \sigma y$ or $y \sigma z \sigma x$, and then $BE((x, z), y, \sigma) = 0$. Conversely, $BE((x, z), y, \sigma) = 1$ if either $x \sigma y \sigma z$ or $z \sigma y \sigma x$, and then $BE((x, y), z, \sigma) = 0$. So, (x, y, z) yields an elementary break iff y and z are located on the same side of x according to σ. Then:

$$\Sigma_{e = (x, y), z \text{ adjacent to exactly 1 extremity of } e} BE(e, z, \sigma) = \Sigma_x \Sigma_{y,z \in \Gamma G(x), (y, x) \notin E, y, z \text{ located the same way with respect to x}, \sigma} 1 = Fk(G, \sigma).$$

Case 3: f = {x, y, z} is an *Anti-Fork* with root z. Therefore, we have:

$\Sigma_{e = (x, y), z \text{ such that } (x, z) \notin E \text{ and } (y, z) \notin E} BE(e, z, \sigma) = AFk(G, \sigma)$.

We get (E1) from the relation: $\Sigma_{e, z} BE(e, z, \sigma) = \Sigma_{e = (x, y), z \text{ such that } (x, y, z) \text{ is a triangle}} + \Sigma_{e = (x, y), z \text{ adjacent to 1 extremity of } e} BE(e, z, \sigma) + \Sigma_{e = (x, y), z \text{ adjacent to no extremity of } e} BE(e, z, \sigma)$.

For any $x \in X$, a feasible solution $A(x,\sigma) \cup^{Ex} B(x, \sigma)$ of *Max Cut* is defined on H(x), by setting: $A(x,\sigma) = \{y \in \Gamma_G(x), \text{ such that } y \sigma x\}$; $B(x, \sigma) = \{y \in \Gamma_G(x), \text{ such that } x \sigma y\}$. Its value, in *Max Cut* sense, is: $m(x) - \Sigma_{y,z \in \Gamma G(x), (y, x) \notin E, y, z \text{ located the same way with respect to x}, \sigma} \mathbf{1} \leq Max\text{-}Cut(H(x))$. It follows that, for any $x \in X$, $\Sigma_{y,z \in \Gamma G(x), (y, x) \notin E, y \text{ and } z, \text{ located the same way with respect to x}, \sigma} \mathbf{1} \geq V(x)$. Then we get: $Fk(G, \sigma) = \Sigma_x \Sigma_{y,z \in \Gamma G(x), (y, x) \notin E, y, z \text{ located the same way with respect to x}, \sigma} \mathbf{1} \geq \Sigma_x V(x)$. We conclude. □

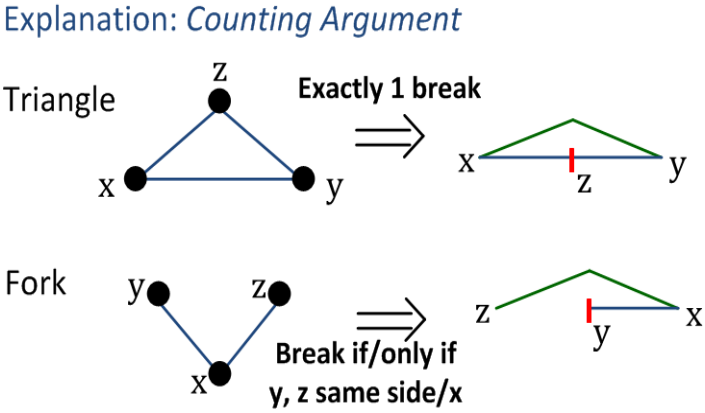


Figure 2: Theorem 1 Counting Argument

Remarks: Small experiments make appear that this bound is often tight, specifically in the case of *chordal* graphs. Still, the *Max Cut* Problem, which has been extensively studied, is *NP*-Hard [9]. So, one may ask about the practical use of the above lower bound. The answer is two-sided:

1. Even though *Max Cut* is NP-Hard, it may be considered as easier to handle than LAP: it admits a natural quadratic {0,1} formulation ([4, 11]) and instances related to the H(x), $x \in X$, are smaller than for original LAP.
2. We shall see in Section 6 that, in the case of interval graphs, our bounding scheme gives rise to an efficient polynomial time approximation scheme.

IV. Quadratic and linear models for LAP

The counting argument of Theorem 1 may also be used in order to derive a quadratic model for LAP. For every pair of nodes (x, y), we need to decide whether y is located right or left in relation to x according to the linear ordering σ. We express this information through a {0, 1} valued $W = (W_{x,y}, x \neq y)$ whose semantics is that: $W_{x,y} = 1$ (0) $\Leftrightarrow$ y located on the right (left) side of x according to linear ordering σ. Counting argument of Theorem 1 tells us that the number of elementary breaks induced by some linear ordering σ is a sum of:

- the number of triangles, which does not depend on linear ordering σ;
- the number of fork {x, y, z}, which are such that y and z are located on the same side with respect to root x according to σ ; (Q1)
- the number of anti-fork {x, y, z}, which are such that x and y are located on different side with respect to root z according to σ. (Q2)

In order to deal with quantity Q1, we introduce a {0, 1} valued vector $U = (U_{x,y}, (x, y) \text{ edge} \in E)$, whose semantics are: $U_{x,y} = 1$ (0) $\Leftrightarrow$ y located on the right (left) side of x. In order to deal with quantity Q2, we introduce a {0, 1} valued vector $V = (V_{x,y}, (x, y) \text{ anti-edge} \notin E)$, whose semantics are: $V_{x,y} = 1$ (0) $\Leftrightarrow$ y located on the right (left) side of x. While it is easy to state the constraints which must be satisfied by U, V, W in order to make them define a consistent linear ordering σ, we see that Q1 becomes equal to:

(Number of *forks* with root x) –

$$\Sigma_x \Sigma_{z,\ y \text{ such that } (y,z) \in E^c} U_{x,y}.\ (1 - U_{x,z}).$$

Also Q2 becomes equal to:

$$\Sigma_z \Sigma_{x,\ y \text{ such that } (x,y) \in E} V_{z,x}.(1 - V_{z,y}).$$

We deduce the following quadratic LAP model:

A Quadratic Linear Formulation of LAP.

- **Variables**
  - $U_{x,y}$, x, y such that (x, y) edge $\in$ E: $U_{x,y} = 1$ (0) $\Leftrightarrow$ y located to the right (left) of x
  - $V_{x,y}$, x, y such that (x, y) anti-edge $\in E^c$: $V_{x,y} = 1$ (0) $\Leftrightarrow$ y located to the right (left) of x
  - $W_{x,y}$, $x \neq y$: $W_{x,y} = 1$ (0) $\Leftrightarrow$ y located to the right (left) of x

- **Constraints** *(Consistency)*
  - For any x, y, $W_{x,y} + W_{y,x} = 1$
  - For any edge (x, y) $\in$ E, $U_{x,y} = W_{x,y}$
  - For any anti-edge (x, y) $\in E^c$, $V_{x,y} = W_{x,y}$
  - For any x, y, z, all distincts, $W_{x,y} + W_{y,z} \geq W_{x,z}$
  - For any x, y, z, all distincts, $(1 - W_{x,y}) + (1 - W_{y,z}) \geq (1 - W_{x,z})$

- **Minimize**

$$\Sigma_z \Sigma_{(x,y) \in E} V_{z,y}.(1 - V_{z,x}) - \Sigma_x \Sigma_{(y,z) \in E^c} U_{x,z}.\ (1 - U_{x,y})$$

If we denote by *Fork*(G) the number of forks of the graph G, we easily get:

Theorem 2 : *The optimal value of this quadratic {0,1} program is equal to LAP(G) – Tr(G) – Fork(G).*

This quadratic {0, 1} model may be easily turned into a linear one by introducing additional vectors S and T as follows:

- $S = (S_f, f = (x, y, z)$, fork with root x) subject to: $S_f \leq (1 - U_{x,z})$ and $S_f \leq U_{x,y}$; we consider here that forks are oriented, that means that (x, y, z) and (x, z, y) define 2 distinct forks with root x;

- $T = (T_f, g = (x, y, z)$, anti-fork with root z) T_g subject to: $T_g \geq (1 - V_{z,x})$ and $T_g \geq V_{z,y}$; we consider here that anti-forks are oriented, that means that (x, y, z) and (y, x, z) define 2 distinct anti-forks with root z.

Then minimizing the quadratic quantity $\Sigma_z \Sigma_{(x,y) \in E} V_{z,x}.(1 - V_{z,x}) - \Sigma_x \Sigma_{(y,z) \in E^c} U_{x,y}.\ (1 - U_{x,z})$ means minimizing the linear quantity $\Sigma_g T_g - \Sigma_f S_f$.

V. A BRANCH/BOUND ALGORITHM FOR LAP

We may derive from previous section an exact Branch/Bound method for LAP:

- *Branching* is performed by picking up some pair (x, y) of nodes and considering the two alternatives x σ y and y σ x according to linear ordering σ; Any sequence of such decisions may be extended through transitivity into a partial ordering of the node set X;

- *Bounding* is performed through integer linear programming, while extending Theorem 1 in a natural way: if σ is a partial ordering of the node set X obtained as above, we may set, for any node x:

  - $Max\text{-}Cut_\sigma(H(x))$ = Optimal value of the *Max-Cut* instance which is defined on the graph H(x), augmented with the following constraints:
    - If y such that (x, y) $\in$ E is also such that y σ x , then y must be on the subset A of the partition $\Gamma_G(x) = A \cup^{Ex} B$;
    - If z such that (x, z) $\in$ E is also such that x σ z , then z must be on the subset B of the partition $\Gamma_G(x) = A \cup^{Ex} B$;

  - $V_\sigma(x) = m(x) - Max\text{-}Cut_\sigma(H(x))$;

  - $W_\sigma(x) = \text{Inf}_{\text{partitions } A \cup Ex\ B \text{ of } X - \{x\} - \Gamma G(x)}$ Card({(x, y) $\in$ E, with x, y such that x $\in$ A, y $\in$ B})

  Then we see that the quantity $Tr(G) + \Sigma_x V_\sigma(x) + \Sigma_x W_\sigma(x)$ provides us with a lower bound for the best (in LAP sense) linear extension of σ.

- Branching strategy comes in a natural way: we give priority to pairs (x, y) which define edges of the graph G, and choose them in such a way the difference between the best alternative and the worst one is the largest possible.

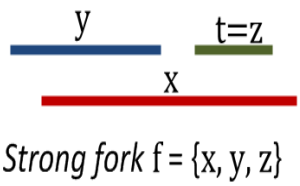
VI. THE CASE OF INTERVAL GRAPHS: A RESTRICTED VERSION OF LAP

We first introduce additional definitions related to interval graphs:

- A simple graph with no loop G = (X, E) is an *interval graph* if it is the intersection graph of a set [o(x), d(x)], x $\in$ X, of closed intervals of the real line. Those intervals may be chosen such that points o(x), d(x), x $\in$ X, are distinct. We assume this hypothesis to be always satisfied. We set:
  - x $\subset$ y if o(x) < o(y) and d(y) < d(x);
  - x << y if d(x) < o(y);
  - x *Ov* y if o(x) < o(y) < d(x) < d(x).

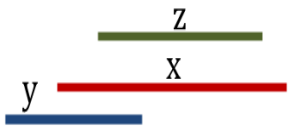
- In case X is an interval family with distinct endpoints, we say that a linear ordering σ of X is (*Ov*, <<)-*consistent* if it is consistent with both orderings *Ov* and <<. We denote by σ-*can* the *canonical linear ordering,* which is defined as follows: x σ-*can* y if, and only if, o(x) < o(y).

- Then, we say that a *fork* f = {x, y, z} with root x of such an interval graph G = (X, E) is a *strong fork* if there exists t $\in$ {y, z} such that t $\subset$ x, and that a triangle (x, y, z) is a *strong triangle* if at least some node is contained into another one (for instance z $\subset$ x).



Strong fork f = {x, y, z}

Figure 3: *Strong fork* f = {x, y, z}



Strong triangle = {x, y, z}

Figure 4: *Strong triangle* = {x, y, z}

- We say that G is a *Unit Interval graph* if intervals [o(x), d(x)], x $\in$ X may be chosen in such way that no pair x, y exists such that x $\subset$ y.

- We finally say that a subset Y of X is a *Left-(Ov, <<)-Section* (*Right-(Ov, <<)-Section*) if, for any x, y $\in$ X such that x $\in$ Y and (y Ov x) $\vee$ (y << x), then we also have y $\in$ Y (x $\in$ Y).

A. A Direct Application of Theorem 1 to Unit Interval Graphs

In the case of unit interval graphs, Theorem 1 allows us to state:

Theorem 2: *If G = (X, E) is a unit interval graph, then σ-can is an optimal solution of LAP.*

Proof: Let us suppose that an *elementary break* (e = (x,y), z, σ-*can*) exists, and that x *Ov* y, which implies that x σ-*can* y. If x << z then y σ-*can* z and z does not break e. Similarly, if z << x then z << x and z does not break e. It comes that x $\cap$ z is not empty. By the same way, y $\cap$ z is not empty and{x, y, z} is a *triangle*. So, there is a one-to-one correspondence between *triangles* and *elementary breaks*. So, BG(G, σ-*can*) = *Tr*(G), and we conclude. □

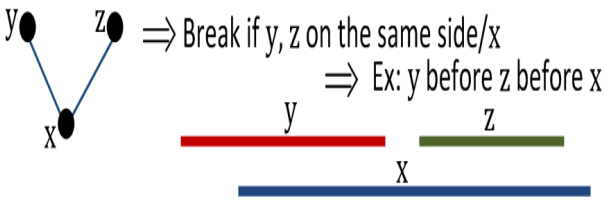


Figure 5: Theorem 2 Argument

B. An Approximation Result.

In the case of general interval graphs, σ-*can* may not be optimal. As a matter of fact, optimal solution may even not be (*Ov*, <<)-*consistent*:

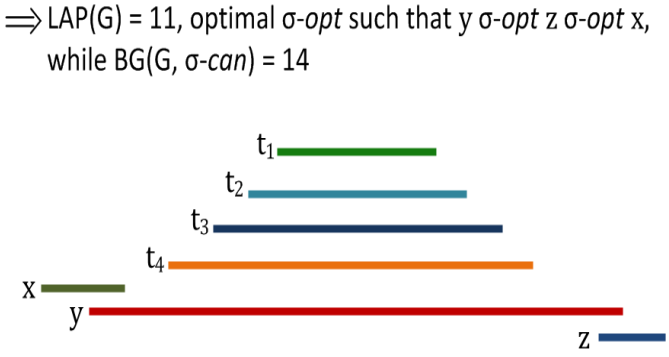


Figure 6: non consistency of (*Ov*, <<)

Still, what can be easily checked is that σ-*can* produces a 2-approximation if we refer to the standard definition of LAP:

Theorem 3: *Given an interval graph G = (X, E) with m edges. Then the following inequality holds:*

$BG(G, \sigma\text{-}can) \leq 2LAP(G)+m.$

Proof: A *Global Break* oriented proof comes by induction on the cardinality of X. Let x_0 be the first (smallest) element of X according to σ-*can*, and σ-*opt* some optimal solution of LAP. Induction tells us that: (E2)

$$\Sigma_{e,\, z \neq x0}\, BE(e, z, \sigma\text{-}can) \leq$$
$$m - |\Gamma_G(x_0)| + 2.\, \Sigma_{e,\, z \neq x0}\, BE(e, z, \sigma\text{-}opt).$$

Since all vertices of $\Gamma_G(x_0) \cup \{x_0\}$ are consecutive according to σ-*can*, we get that:

$$\Sigma_{x \in \Gamma G(x0),\, z \in X}\, BE((x_0, x), z, \sigma\text{-}can) =$$
$$\Sigma_{x,\, z \in \Gamma G(x0)}\, BE((x_0, x), z, \sigma\text{-}can)$$
$$= |\Gamma_G(x_0)| \times (|\Gamma_G(x_0)| - 1)/2.$$

On the other hand, if we refer to σ-*opt*, we get: (E3)

$$\Sigma_{x \in \Gamma G(x0),\, z \in X}\, BE((x_0,x), z, \sigma\text{-}opt) \geq$$
$$\Sigma_{x,\, z \in \Gamma G(x0)}\, BE((x_0,x), z, \sigma\text{-}opt) \geq$$
$$\lfloor |\Gamma_G(x_0)|/2 \rfloor (\lfloor \Gamma_G(x_0) \rfloor - 1)/2 + \lceil |\Gamma_G(x_0)|/2 \rceil (\lceil |\Gamma_G(x_0)|/2 \rceil - 1)/2.$$

We derive the result by combining (E2) and (E3). □

C. A Restricted Version CLAP of LAP.

However, experiments will show that best linear orderings are most often (*Ov*, <<)-*consistent.* So we are going to study the following restriction CLAP of LAP:

(*Ov*, <<)-*Consistent Linear Arrangement* Problem (*CLAP*): {Compute a (*Ov*, <<)-consistent linear ordering σ which minimizes BG(G, σ)}.

The following lemma bridges CLAP with Theorem 1.

Lemma 1: *In the case the linear ordering σ is (Ov, <<)-consistent, we have that: BG(G, σ) = Tr(G) + SFk(G, σ), where SFk(G, σ) is the number of strong forks f = (x, y, z), x = Root(f) such that ((x σ y) ∧ (x σ z)) ∨ (y σ x) ∧ (z σ x)).*

Proof: left to the reader (same proof as for Theorem 1). □

Extending Theorem 1 to CLAP leads us to introduce a specific version of *Max-Cut*:

(Ov, <<)-Consistent Unit Cost Max-Cut Problem (C-Max-Cut): Given a graph H = (Z, F), which is the complementary graph of an interval graph H^c = (Z, F^c), and two disjoint subsets A_0 and B_0 of Z, such that:

- A_0 (B_0) is a *Left-(Ov, <<)-Section* (*Right-(Ov, <<)-Section*) of H^c;
- Both A_0 and B_0 define complete sub-graphs of H^c = (Z, F^c).

Compute a partition $Z = A \cup^{Ex} B$, such that:

1. A contains A_0 and is a *Left-(Ov, <<)-Section* of H^c;
2. B contains B_0 and is a *Right-(Ov, <<)-Section* of H^c;
3. the number of edges of H which connect A and B = $|\{(x, y) \in E, x \in A, y \in B\}|$ is the largest possible;
4. A is maximal for the set inclusion order, provided 1, 2, 3 are satisfied.

We denote by *C-Max-Cut*(H, A_0, B_0) the related optimal value. Then we set, for the interval graph G = (X, E) and for any vertex x in X:

- $\Gamma^{Ov,\subset}{}_G(x)$ = (*0v*, ⊂)-*neighbour* set of x = {y ∈ $\Gamma_G(x)$, y ≠ x, such that (y ⊂ x) or (y *Ov* x) or (x Ov y)};
- H(x) = complementary graph of the sub-graph induced by $\Gamma^{Ov,\subset}{}_G(x)$;
- $A_0(x)$ = {y ∈ Z such that y *Ov* x}; $B_0(x)$ = {y ∈ Z such that x *Ov* y};

- m(x) = number of edges of H(x); CV(x) = m(x) - *C-Max-Cut*(H(x), $A_0(x)$, $B_0(x)$).

Then we get:

Lemma 2: *CLAP(G)* $\geq$ *Tr(G)* + Σ_x *CV(x).*

Proof: For every x $\in$ X, we set E*(x) = {non oriented pairs (y, z) such that:

- y $\in \Gamma^{Ov,\subset}{}_G(x)$, z $\in \Gamma^{Ov,\subset}{}_G(x)$; (y, z) $\notin$ E ;
- at least one of both relations y $\subset$ x or z $\subset$ x holds;
- relation ((x σ y) $\wedge$ (x σ z)) $\vee$ ((y σ x) $\wedge$ (z σ x)) holds}

SFk(G, σ) may be written as $\Sigma_{x \in X}$ |E*(x)|. Since σ is (*Ov*, <<)-consistent, we may relax the "*at least … y* $\subset$ *x or z* $\subset$ *x holds*" constraint which characterizes E*(x).

So, for any x $\in$ X:

|E*(x)| $\geq$

$\Sigma_{y,z \in \Gamma G(x),\ (y,x) \notin E,\ y,z \text{ are located the same way with respect to x, } \sigma}$ **1**.

For any x $\in$ X, we get a feasible solution A(x,σ) $\cup^{Ex}$ B(x, σ) of the *C-Max-Cut* instance defined by H(x), $A_0(x)$, $B_0(x)$, by setting: A(x,σ) = {y $\in \Gamma^{Ov,\subset}{}_G(x)$, such that y σ x}; B(x, σ) = {y $\in \Gamma^{Ov,\subset}{}_G(x)$, such that x σ y}. Its value is:

m(x) –

$\Sigma_{x,y \in \Gamma Ov,\subset G(x),\ (y,x) \notin E,\ y,z \text{ located the same way with respect to x, } \sigma}$ **1**

$\leq$ *C-Max-Cut*(H(x), $A_0(x)$, $B_0(x)$).

It follows that, for any x $\in$ X:

|E*(x)| $\geq$

$\Sigma_{x,y \in \Gamma Ov,\subset G(x),\ (y,x) \notin E,\ y,z \text{ located the same way with respect to x, } \sigma}$ **1**

$\geq$ CV(x).

Then, we get that SFk(G, σ) = Σ_x |E*(x)| $\geq$ Σ_x CV(x). We conclude. □

D.Solving C-Max-Cut and Evaluating CV(x)

The complexity of the *Max Cut* problem in the case of the complementary graph of an interval graph is still an open issue. However, things are easier with *C-Max-Cut*:

Theorem 4: *Given = (Z, F), A_0 and B_0 as in the definition of C-Max-Cut. Let us set, for every vertex z* $\in$ *Z* – (A_0 $\cup$ B_0)*:*

- $d_H^-(A_0, z)$ = |{*t* $\in$ *Z* – A_0 *such that t* << *z* }| + |{*t* $\in$ A_0 *such that t* << *z*}|;
- $d_H^+(B_0, z)$ = |{*t* $\in$ *Z* – B_0 *such that z* << *t* }| + |{*t* $\in$ B_0 *such that z* << *t* }|.

Then we solve C-Max-Cut by setting:

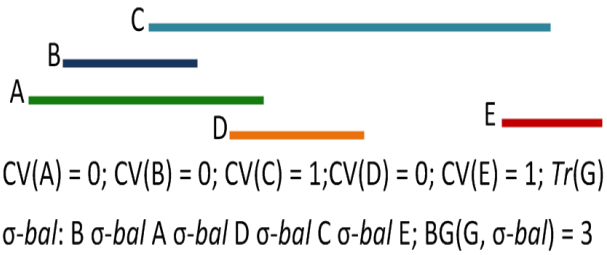
- *A* = {*z* $\in$ *Z* – (A_0 $\cup$ B_0) *such that* $d_H^-(A_0, z) \geq d_H^+(B_0, z)$} $\cup A_0$;
- *B* = {*z* $\in$ *Z* – (A_0 $\cup$ B_0) *such that* $d_H^-(A_0, z) < d_H^+(B_0, z)$} $\cup B_0$.

Proof: Left to the reader. □

E. An Exact Solution σ*-bal for CLAP.*

We construct this solution σ-*bal*, by setting, for any pair x, y in X, x σ-*bal* y if, and only if, one among the following options holds:

- (x << y) or (x *Ov* y);
- (x $\subset$ y) and $d_{H(y)}^-(A_0(y), x) \leq d_{H(y)}^+(B_0(y), x)$; (E5)
- (y $\subset$ x) and $d_{H(x)}^-(B_0(x), y) < d_{H(x)}^+(A_0(x), y)$. (E6)



CV(A) = 0; CV(B) = 0; CV(C) = 1;CV(D) = 0; CV(E) = 1; *Tr*(G) = 2;

σ-*bal*: B σ-*bal* A σ-*bal* D σ-*bal* C σ-*bal* E; BG(G, σ-*bal*) = 3

Figure 7: A σ-*bal* construction

Lemma 3: *The* σ*-bal relation is transitive.*

Proof: left to the reader. □

We are now ready to state the optimality of σ-*bal*.

Theorem 5: *The relation* σ*-bal is an optimal solution of CLAP, which satisfies:*

1. *BG(G,* σ*-bal)* $\leq$ *BG(G,* σ*-can).*
2. *Tr(G)* $\leq$ *BG(G,* σ*-bal) = Tr(G) +* Σ_x *CV(x)* $\leq$ *Tr(G) + Strong-Fork/2, where Strong-Fork is the number of strong forks of the interval graph G.*

Sketch of the Proof: From Lemma 3, we have that σ-*bal* is a (<<, *Ov*)-consistent linear ordering. Then the optimality of σ-*bal* (and so, the fact that BG(G, σ-*bal*) $\leq$ BG(G, σ-*can*)) derives, through a simple computation, from the fact that since it locally achieves, for any node x, the lower bound CV(x), then it also globally achieves the lower bound of Lemma 2. □

We easily deduce that this result has an algorithmic interpretation:

Corollary 3 (left to the reader): *Computing σ-bal may be done in O(Arc-⊂) time, where Arc-⊂ is the number of arcs of the digraph induced on X by the ⊂ ordering.*

VII. NUMERICAL EXPERIMENTS

We implemented both the Branch/Bound algorithm of Section 5 and the Approximation algorithm σ-*bal*. We did it on a LINUX server CentOS 5.4, Processor Intel Xeon 3.6 GHZ, while using the CPLEX 12 library when dealing with integer linear programs. For interval graphs, our Branch/Bound scheme could solve, in no more than few minutes, instances with up to 40 nodes. This allowed us to perform a comparative analysis of the precision of the Lower Bound LB(G) = $Tr(G) + \Sigma_x V(x)$ provided by Theorem 1, and of both approximation algorithms σ-*can* and σ-*bal*.

We use 10 instance groups related to Card(X) = 10, 20, 30 and 40, and for every instance group, compute:

- the mean gap LB-GAP = (LAP(G) –LB(G))/LB(G) between the optimal LAP value and the lower bound LB;
- the mean gap CLAP-GAP = (BG(G, σ-*bal*) -LAP(G))/LAP(G) between the optimal value of CLAP, computed by σ-*bal*, and the optimal value;
- the mean gap CAN-GAP = (BG(G, σ-*can*) -LAP(G))/LAP(G) between the value defined by the canonical ordering σ-*can*, and the optimal value.

We get results which are described in the following Table 1.

N	**10**	**20**	**30**	**40**
LB-GAP	9.8%	12.5%	10.7%	14.5%
CLAP-GAP	3.2%	4.5%	4.1%	5.0%
CAN-GAP	9.6%	12.3%	11.5%	10.9%

Table 1: Comparative precision of lower bound LB and approximation solutions σ-*can* and σ-*bal*.

Table 2 provides now the specific results related to n = 10, which by the way, gives an estimation of the LAP(G) values which may derive from interval graph of this size:

INSTANCE NUMBER	**LB(G)**	**LAP(G)**	**CLAP(G)**	**CAN(G)**
1	10	12	12	13
2	8	8	8	8
3	13	16	17	19
4	12	13	14	15
5	12	13	13	14
6	10	11	11	11
7	15	15	17	18
8	13	15	15	17
9	11	11	11	12
10	9	10	10	10

Table 2: Values LAP(G), LB(G), CLAP(G) and CAN(G) related with a 10 instances group with Card(X) = 10.

VIII. CONCLUSION

This paper, with theoretical focus, proposes approximation results for the *Linear Arrangement* problem, in the case of interval graphs. Further research should be about the extension of our approaches to chordal graphs and circular graphs, as well as about the design of efficient exact algorithms.

REFERENCES

1. Achouri S., Bossart T., Munier-Kordon A. (2009): A polynomial algorithm for MINDSC on a subclass of series parallel graphs, *RAIRO Operations Research*, pp. 145-156, DOI: 10.1051/ro/2009009
2. Barahona F., Mahjoub A.R (1986): On the cut polytope, *Math. Prog.* 36, pp. 157-173, DOI: 10.1007/BF02592023
3. Charon I., Hudry O. (2010): An updated survey on the linear ordering problem for weighted or unweighted tournaments, *Annals of Operations Research,* 175, pp. 107-158, DOI: 10.1007/010479-009-0648-7
4. Chung FRK. (1984): On optimal linear arrangement of trees. *Comp. & Maths/Appl.*, 11, pp. 43-60, DOI: 10.1145/73833.738333.73866

5. Cohen J., Fomin F., Heggernes P., Kratsch D., Kucherov G. (2006): Optimal linear arrangement of interval graphs, *Proc. MFCS'06*, pp 267-279, Springer-Verlag, DOI: 10.1007/1182069_24
6. Corneil DG., Kim H., Natarajan S., Olarin S., Sprague AP. (1995): A simple linear time algorithm of unit interval graphs, *Information Processing Letters* 55, pp. 99-104, DOI: 10.1016/0020-0190(95)00046-F
7. Chvatal V., Ebenegger C. (1990): A note on line digraphs and the directed Max-Cut problem, *Discrete Applied Maths* 29, pp 165-170, DOI: 10.1016/0166-218X(90)90141-X
8. Even S., Shiloach Y. (1975): NP-Completeness of Several Arrangement Problems, *Technical Report #43*, *Computer Science Department*, *The Technion*, Haifa, Israel, DOI: 10.1007/11821069_24
9. Garey MR., Johnson DS. (1979): *Computers and intractability: a guide to the theory of NP-completeness*, Computer Press, ISBN-13: 978-0716710455.
10. Grotschel, M. (ed.) (2004): *The Sharpest Cut*, MPS-SIAM Series on Optimization, ISBN-13: 978-0898715521
11. Horton SB. (1997): *The optimal linear arrangement problem: algorithms and approximation*, Phd thesis, Georgia Institute of Technology.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 501–509
DOI: 10.15439/2014F208
ACSIS, Vol. 2

An Application of Developmental Genetic Programming for Automatic Creation of Supervisors of Multi-task Real-Time Object-Oriented Systems

Krzysztof Sapiecha
Department of Computer Engineering
Cracow University of Technology
Cracow, Poland
Email: krzysztof.sapiecha@gmail.com

Leszek Ciopiński
Department of Computer Science
Kielce University of Technology
Kielce, Poland
Email: l.ciopinski@tu.kielce.pl

Stanisław Deniziak
Department of Computer Science
Kielce University of Technology
Kielce, Poland
Email: s.deniziak@tu.kielce.pl

***Abstract*—A concept of artificial supervisor of multi-task real-time object-oriented system is introduced. Next, a procedure for automatic creation of artificial supervisors is presented. The procedure is based on developmental genetic programming. As an input data, UML diagrams are used. A representative example of creation of a supervisor of building a house illustrates the procedure. The efficiency of the procedure from various points of view and comparison considerations are given.**

I. Introduction

A SYSTEM must meet user requirements and constraints. Besides specified functionalities, cost effectiveness and high performance are among the most important ones. Real-time (RT) systems are present in all areas of human life. Punctuality is an extra requirement for them. Sometimes punctuality requires very high performance. However, the higher performance the higher cost of the system. Usually the cost is limited.

Going into details, one can find RT systems in civil engineering, in traveling, in computer engineering, in banking, and so on. In the first case, a building enterprise is such a system. It owns resources, such as workers and building machinery, necessary to build a house according to the requirements of a client. These usually comprise functionalities of the house, its cost and a deadline. In the second case a human being is an RT system. He knows which means of transportation may use to meet his requirements. From among flights, trains, buses, rented cars, and even walking he selects a set, so that to reach a target on time and at affordable cost. An embedded computer system may be another example of RT system. A designer of such a system has to decide what of the tasks of the system is to allocate to what of its processing components, so that to get maximum of the performance and not to surpass the cost. In a bank an account may be operated in different ways. However, an owner of the account usually wants to get maximum profit with an acceptable risk in a specific time period. Summarizing, an RT system uses some hardware or software objects of its resources so that a specific goal is achieved, on time.

Multi-task system (MS) is a system where more than one task can be processed at the same time. A home computer is a familiar example of the MS. Common tasks are word processing, printing, communicating, and playing games. The system contains objects: hard drive, a monitor, a printer, a network adapter and an optical drive. Some of these objects are required for a subset of the tasks while others are required for all these tasks. The monitor and hard drive will always be in use whereas the printer is used only for printing, the network adapter is used for communicating, and the optical drive is used for reading stored materials. The enterprise is an example of RT MS, while a traveler is not.

Usually, RT systems should be optimized for cost vs. speed of operation (speed of reaching a goal or a target). Therefore, a building enterprise, and a traveler, and hardware/software system designer, and other RT systems have to be endowed with optimization engine. We will call these engines: artificial supervisors (AS) or artificial managers (AM) of resources. An AS should find an optimum use of supervised resources, taking into account the requirements and the constraints. This means that the AS decides what functionalities should be allocated to what resources and in which order these functionalities should be executed. Actually, it has to find a solution for a specific case of the well-known Resource-Constrained Project Scheduling Problem (RCPSP) which consists in rescheduling the project tasks (RT system tasks) efficiently using limited renewable resources (components/objects of the RT system) minimizing the maximal completion time of all activities [1].

The RCPSP is an NP-complete problem which is computationally very hard [2] [3]. Möhring [4] states that it is one of the hardest problems of Operational Research. Therefore, a skilled specialist with an assistance of the planner (Computer Aided System Engineering in case of the enterprise) might play a role of such an AS only for small systems containing a limited number of tasks and a moderate number of resources. No doubt, in case of real life systems, particularly RT MS, the AS must be a very powerful optimization engine.

The general RCPSP model cannot cover all situations that occur in practice. Therefore, many researchers have developed many variants and extensions of project scheduling problems, often using the standard RCPSP as a starting point [5].

978-83-60810-58-3/$25.00

Constructing an efficient AS for a given class of RCPSP problems is very difficult and time consuming. Moreover, a scheduling strategy that is optimal for one problem may not be efficient for others. Hence, instead of developing a general AS for all RT MS, in this paper, we propose a method that automatically generates a dedicated AS for the specific RT MS. The method is based on an idea derived from developmental genetic programming (DGP). It is universal and can be applied to optimization of RT MS of any kind. Our methodology is illustrated and evaluated with the help of a representative example.

Genetic programming (GP) is an extension of the genetic algorithm [6], in which the population consists of computer programs. In the DGP [7] [8], strategies that create solutions evolve, instead of computer programs. In this approach a genotype and a phenotype are distinguished. The genotype is a procedure that constructs a solution of the problem. It is composed of genes representing elementary functions, constructing the solution. The phenotype represents a target solution. During evolution, only genotypes are evolved, while genotype-to-phenotype mapping is used in the fitness computation, which is required for the genotype selection process. Next, all genotypes are rated according to an estimated quality of the corresponding phenotypes. The goal of the optimization is to find the procedure constructing the best solution. The idea is based on the theory from the molecular biology, concerning protein synthesis that produces proteins (phenotype) from the DNA (genotype). In our approach the AS corresponds to the genotype while the phenotype is the solution i.e. a makespan. First, the DGP is used to find the optimal solution, and the genotype constructing this solution is saved as the AS.

The method is universal, but an AS must be well-fitted to a particular RT MS. The RT MS is a micro-world with its own functionalities and resources. Therefore, a formal specification of the RT MS is an input data to the method. RT MS, where a number of resources have punctually to execute a number of tasks are good micro-worlds for object-oriented modeling. Objects may play a role of resources that execute tasks in real time for some costs. A widely accepted standard for modeling object-oriented systems (OOSs) is the Unified Modeling Language (UML) [9]. It shows how to write a system's blueprints, including conceptual things such as business processes and system functions. It encompasses OOSs of any kind, particularly real-time multi-task OOSs (RT MOOSs). Using the UML for modeling RT OOS has been a subject of many publications [10] [11] [12]. Hence, this will be applied here.

Related work is briefly described in section II. In section III the problem is stated. Section IV briefly shows how early UML models should be used as input data for the method, and section V explains how DGP can create the supervisors and the initial solutions. In section VI a computational experiment evaluating our approach is described. The experiment explains of how a supervisor of a simple RT MOOS (of building a house) is created. Finally, section VII contains conclusions.

II. Related Work

Genetic approach was proved as very efficient for solving RCPSP problems. Ones of the most efficient genetic algorithms for RCPSP are presented in [13] [14]. In [15] the method of improving the genetic algorithm for optimization of multi-task project scheduling was proposed. It was showed that the method is competitive in comparison with 11 other heuristic approaches. A method of solving a large scale RCPSP is presented in [16]. In this solution, a genetic algorithm is used and a method of encoding classical RCPSP problem in the chromosome is described. Results achieved by authors of [16] give a slight improvement, in comparison with other existing heuristics.

For the first time Developmental Genetic Programming was proposed by Koza, Bennet, Andre and Keane [17], to create electrical circuits. This methodology evolves circuit-construction tree, in which nodes correspond to functions defining the developmental process. The initial circuit consists of an embryo and a test fixture. The sample embryo is at least one modifiable wire while fixture is one or more unmodifiable wires or electrical components. The circuit is developed by progressively applying functions in the circuit-construction tree to the modifiable parts (wires and electronic components) of the embryonic circuit.

A similar methodology was used by Deniziak and Górski in the co-synthesis of embedded systems described by task graphs [18]. The system-construction tree is based on a task graph. Each node of the tree specifies an implementation of the corresponding task. The embryo is an allocation of the first task. First (initial) population is created randomly. Then after evolution, using crossover, mutation and reproduction, an optimal (or suboptimal) solution is found.

In [19] a list of 36 instances of human-competitive results produced by the GP is presented. A lot of them concern of synthesis of an analog electrical circuits, developing quantum algorithms, designing controllers. According to our best knowledge, there is no approach concerning optimization of object-oriented real-time multi-task systems using DGP methods.

III. Problem Statement

Let us assume that information about the functionalities of RT MOOS and resources available for implementation of these functionalities are specified with the help of UML early diagrams: use case, activity, and sequence. This is typical while designing OOS of any kind. To start the system a supervisor is needed, which allocates the functionalities into the resources in such a way that the cost will be minimal while all real-time constraints will be satisfied. Both, number of the functionalities and the number of resources, are large enough to exclude a human being as the supervisor. Therefore, an engine which optimizes supervising the system should be worked out. The engine will be named an artificial supervisor (AS), since it does what the supervisor should do.

An AS should work as follows:

1) it should work out a schedule for RT MOOS which would be optimal under current operational conditions, and
2) adjust the schedule, to keep its optimality, when the conditions have changed (some of the resources had failed, for example).

The goal of the research is to introduce a method of automatic generation of ASs from the diagrams. An approach based on an idea derived from developmental genetic programming is used.

The procedure consists of two steps. In the first one information included in the UML diagrams are transformed into a task graph and a library of objects working for the system. These are input data to the second step. In this step (Section V) the AS is created. To this end a universal method of evolution of a genotype of the AS is applied. Decision options, which may be contained in the genes of the AS, are defined and then the genotype is created developmentally, using DGP-like approach.

An example of the generation of a supervisor in a building enterprise is used to illustrate the method. The enterprise is an RT MOOS because its resources may be dealt with as objects of different kinds, human or technical, which work in real time and in multi-task mode of operation. A user of the RT MOOS specifies the functionalities of the house, a deadline of the implementation and cost constraints. In the case of a small building enterprise, a contractor assisted by a CASE tool (Computer Aided Software Engineering) can elaborate optimal or semi-optimal schedule of building the house. However, big consortia own a large number of resources and implement many different constructions. Hence, this duty must be waived from the contractor and placed onto an AS. Not the contractor, but the AS, which is engaged in building the house, should generate an optimal schedule for management of enterprise resources.

Summarizing, for each of the implementations an AS should be generated. The AS elaborates optimal schedule of the implementation. A procedure of generation of ASs is universal. However, to generate a specific AS (for building a housing estate or LNG terminal, or a bridge, or managing a bank account, and so on) it should be supplemented with data describing what should be supervised. This is done with the help of UML diagrams.

An AS should react to events that make the schedule non-optimal, such as failures of the resources, unexpected delays of task executions and so on. In such situations, any break in work could generate huge costs. Thus, changes in the schedule should be done in real-time, and as soon as possible.

IV. From UML Early Models to Library of Resources

The first step of the method consists in the generation of input data for DGP, which in turn will create an adequate supervisor. To this end UML early models of RT MOOS, which will be under optimization, are used. In case of building a house the models describe all activities of the supervisor that controls the whole process of the construction.

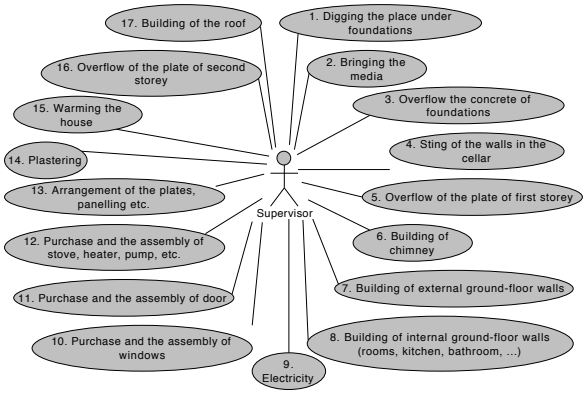


Fig. 1. Use case diagram for building a house

A. Functionalities and sequential constraints

Functionalities are described with the help of a UML use case diagram where a Supervisor is the main actor. It owns resources (objects) performing tasks in real-time and may face orders of task executions. Use case diagram describing building a house is given on Fig. 1. Actually, it maintains the enterprise which works as a real RT MOOS.

A task is an activity performed by a specific user of an RT MOOS. In the diagram, each of the use cases corresponds to one of such tasks, since a use case is an action performed by an object (objects) which aims to yield an observable goal for the user. Thus, each of the tasks has a use case that explains what the task is, and how it should function. Moreover, a use case may include statements about pre-conditions (required before the task began), post-conditions (valid when the task was successfully completed) and, if needed, exceptions.

The diagram on Fig. 1 contains 17 use cases (stages of a house building; numbered from 1 to 17 on Fig. 2) which should be scheduled for enterprise resources. Therefore, assignment of the use cases, to the resources, is a subject of optimization. However, the diagram may not say anything, that one of the cases must be used before another one. Digging foundations must precede their laying, and plastering must be done before warming a house, for example[1]. In general, an RT MOOS as an example of a multi-task system may have sequential constraints. Tasks should not be executed in arbitrary orders because some of the tasks need to be executed before others.

The Supervisor knows use case sequential constraints. This can be specified with the help of an extension and of an inclusion associations («extend» or «include» stereotypes [9]) and on pre- and post-conditions defined for the use cases (sequential dependencies [20]). As the summary, a UML

[1]Numerical prefixes are introduced to identify the use cases and will be used later on.

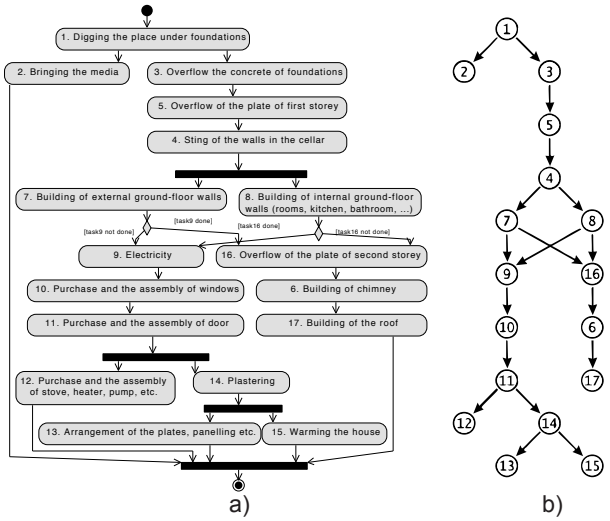


Fig. 2. Activity Diagram (a) and Task Graph (b) of the system.

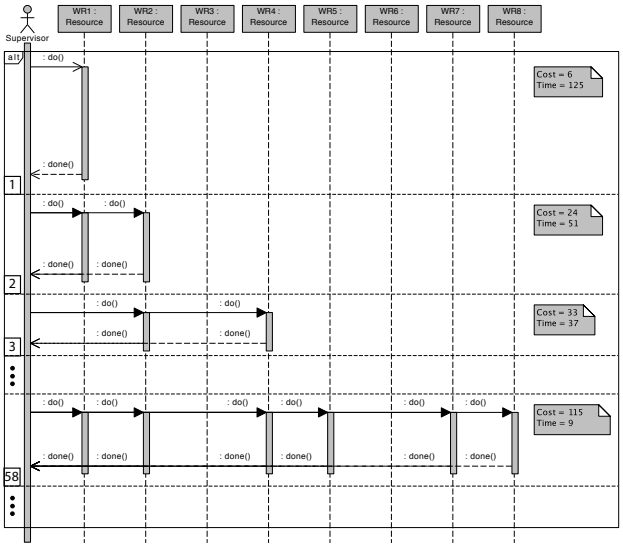


Fig. 3. Sequence diagram of Bringing the media use case.

activity diagram [9] for an RT MOOS can be defined, and then transferred into a task graph (TG) showing an execution of the tasks in a real-time. The constraints for the example are shown on Fig. 2a whereas their corresponding TG is shown on Fig. 2b[2].

B. Resources

The resource is an object required to execute a task. In general, it could be a human, a tool or any other object, which is reusable or renewable. If there are many resources of the same type, each of them should be presented as a separate resource.

In the example, two types of the resources are required for building a house. First are Workers. These are laborers like electricians, plasterers, and so on, that are able to execute some specific tasks. Also, a company, which could be used as an outsourcing, should be given as a resource. A worker could use the second type of the resources which are Tools. These are machines which could be used by workers during execution of tasks. Hardly ever one resource is able to complete a task itself. Thus resources are grouped into work teams, which will be described in the next subsection.

Not all resources are necessary for the execution of some tasks. This could be determined by the Supervisor with the help of the third kind of UML diagrams, namely sequence diagrams.

C. Scenarios of the resource cooperation

Inspired by real world, where most tasks are executed by a group of resources, a concept of a team is introduced. The team is a set of resources that are able to execute a task. Any task may be executed by more than one team. We assume that the execution cost and time of the task are known. One resource may belong to different teams, but teams having the same resources cannot be scheduled at the same time period.

A use case is refined into one or more sequence diagrams to show how the case might be implemented with the help of detailed actions. Therefore, each sequence of the actions defines an actual work-flow and reflects a sequence of decisions the supervisor should take to perform a single task. Every task is executed by teams (single worker team is also possible, but rarely). Different teams are able to finish their work faster or cheaper, using more or less resources. The sequence diagrams have to define these scenarios.

Fig. 3 shows an example how the supervisor might interact with objects participating in the construction of the building[3]. Moreover, the sequence diagrams show options (with time and cost of their implementations) available to the supervisor of the enterprise.

It is characteristic for a Supervisor that while traversing a task graph it determines step by step what should be done at that point and that its selections are usually optional. This means that it actually decides what functionalities of the RT MOOS should be assigned to what objects and in which order these functionalities should be executed. Its decisions should be optimal, taking into account costs and the time of execution. Therefore, the quality of the supervisor should be as high as possible.

D. Library of resources

Sequence diagrams specify how cooperating objects are organized in teams and how do they work. For example, Fig. 3 shows 4 teams. Each of the teams could bring media to a house under construction, but with different workload and costs. From sequence diagrams a table is derived which determines

[2]Automatic generation of TG from UML diagrams is possible but will not be discussed in the paper.

[3]Remaining 16 sequences are very similar.

TABLE I
A LIBRARY OF RESOURCES.

Task #	Team # (for the task #)	Time	Cost	Time * Cost	Members
1:	0	125	6	750	WR1
1:	1	51	24	1224	WR1, WR2
1:	2	37	33	1221	WR1, WR2
...	...	...	...	...	...
1:	58	9	115	1035	WR1, WR2, WR4, WR5, WR7, WR8
...	...	...	...	...	...
1:	62	82	183	15006	WR1, WR2, WR3, WR4, WR5, WR7, WR8
2:	0	57	62	3534	WR3
...	...	...	...	...	...
2:	3	41	82	3362	WR1, WR2, WR5
...	...	...	...	...	...
2:	62	10	189	1890	WR3, WR4, WR5, WR7, WR8
...	...	...	...	...	...

a binding of tasks with teams. For the example, this is given in Table I.

Columns "Time", "Cost" and "Members" of the table are filled in with data from sequence diagrams. Units of "Time" or "Cost" are inessential. It could be a day or an hour, dollar or euro. It is only important that all costs and all times are defined using the same units.

Column "Time * Cost" does not give any new information, but it is helpful to accelerate the time of the computations.

V. CREATION OF SUPERVISORS

The second step of the method consists of initiating and evolving genotypes, corresponding to the supervisors, with the help of DGP. It is assumed that the supervisor selects options defining the strategy of the allocation of resources. The way in which it does, it is a specific feature of its mind, and it is contained in its genotype. A supervisor with the best genotype (allocating the resources optimally) will be generated with the help of DGP. DGP evolves genotypes, while genotype-to-phenotype mapping is used in the fitness computation, which is required for the genotype selection process. It is possible, that one phenotype may be created from two different genotypes, because genotype-to-phenotype mapping always generates systems that meet the system requirements.

A genotype corresponding to the supervisor has a form of a tree engineering the system. A root of the tree specifies a construction of an embryonic system, while all other nodes correspond to functions which progressively build up the whole system. If the system is defined by a task graph, then an embryo is a system executing the first task from the task graph. Thus, the number of possible embryos equals the number of teams, in the library of resources, which are capable of executing the first task. Embryonic systems are selected randomly for each attempt to create an initial population of supervisors.

TABLE II
SUPERVISOR'S OPTIONS

Step	Option	P
1	a. The fastest team	0.16(6)
	b. The cheapest team	0.16(6)
	c. The lowest time * cost	0.16(6)
	d. Determination by second gene	0.16(6)
	e. The fastest starting team	0.16(6)
	f. The fastest ending team	0.16(6)
2	List scheduling	1

A. Supervisor's options

The supervisor undertakes the following two actions:

- resource allocation and task assignment, that send an appropriate team to execute a particular task and hence, allocate members of the team,
- task scheduling (only when more than one task is assigned to the same resource), that schedules the tasks assigned to the resources. When the resource is unavailable, the execution of the task is delayed as long as the resource is not released.

Initial population of supervisors consists of randomly generated genotypes. It selects one of the options given in part 1 of its decision table. Table II contains the options which the supervisor may choose. The last column in Table II shows a probability of the selection.

The first option prefers a team, which requires the smallest period of time to execute a task. Second one prefers a team, which brings the lowest cost increase. Third option prefers a team with the best ratio of the costs to the time of the execution. Fourth option works in a different way. It allows us to use "a little pushed" teams, what cannot be obtained as a result of the remaining options. The next option prefers a team, which could start an execution of the task as soon as possible (other teams might be busy). The last option prefers a team whose members could be the first to finish a task (be freed). For the second action only one option is available, namely the list scheduling method.

B. Genotype

The genotypes have forms of binary trees corresponding to various procedures of synthesis of phenotypes (target solutions). Every node has the same structure presented on Fig. 4

The first field *isLeaf* determines a role of the node in a tree. When it is true (the node is a leaf), the strategy for tasks is described in the field named *"strategy"*, which stores an option from Table II. In this case information from the other fields is omitted. When the node is not a leaf, a content of the field *"strategy"* is not important. In this case, *cutPos* contains a number describing which group of tasks should be scheduled by the left node and which one by the right node. Thus, *nextLeft* and *nextRight* must not be null pointers.

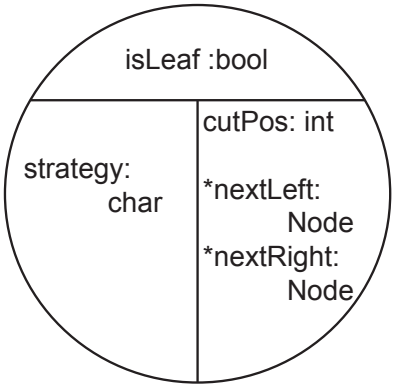


Fig. 4. A node of the genotype

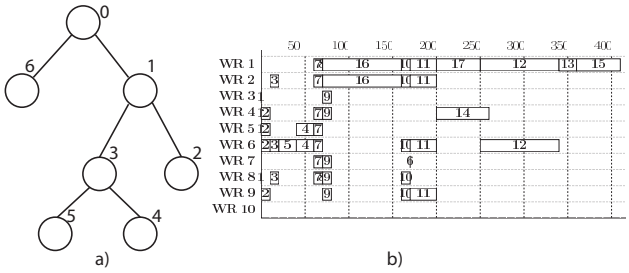


Fig. 5. A simple genotype (a) and the corresponding phenotype (b)

The simplest genotype consists of only one node, which is also a leaf and a root. A simple genotype and the corresponding phenotype are presented on Fig. 5.

During the evolution, a genotype grows but a size of the genotype tree is limited. If the tree exceeds the maximum size then too long branches are cut off. For example, if the maximum size is 6, every node on the sixth level which has a successor is changed into a leaf, and all its successors are destroyed.

An embryo of the tree could grow as an effect of genetic operators: mutation and crossover. An action associated with the mutation depends on the state of the node and is presented in the Table III.

The crossover is used to exchange information between two chromosomes. It is necessary to draw a point of cut a tree in both chromosomes. An example of the crossover is presented on Fig. 6

With every genotype an array is associated. Its size is equal to the number of tasks and contains indexes of teams. If for a task, strategy 'd' is chosen, the team with an index taken from the array is used. At the very beginning of the mutation, a place in the array is randomly chosen. Next a new index is randomly generated. During the crossover, parts of the arrays from both genotypes are swapped.

C. Genotype to phenotype mapping

The first step in a genotype-to-phenotype mapping is to assign strategies to tasks (that is teams from Table I to tasks

TABLE III
THE RULES OF MUTATIONS

Is a leaf?			
Yes		No	
Draw: switch leaf/node or not?			
Yes	No	Yes	No
Set *isLeaf* as FALSE. If *nextLeft* or *nextRight* is NULL - create a new leaf for it.	Draw new strategy	Set isLeaf as TRUE	Change value for a randomly chosen field: *cutPos*, *nextLeft* or *nextRight*

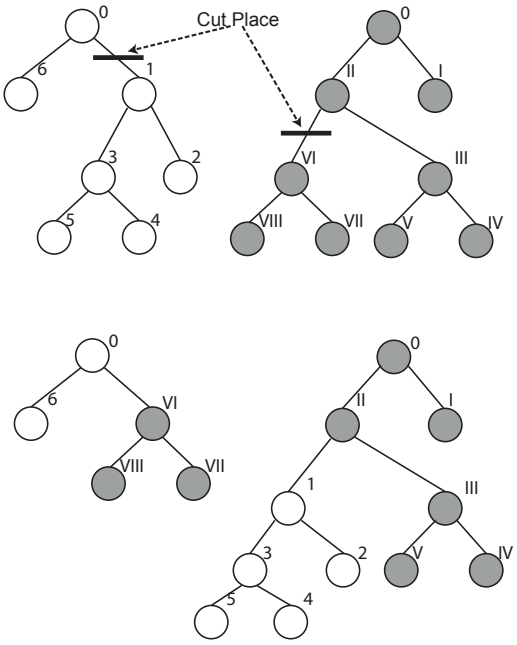


Fig. 6. An example of the crossover

from Fig. 2b, in the example). This step is illustrated on Fig. 7. Please note, that node 1 partitions tasks from 11 to 17 into two groups: from 11 to 18 and the rest. Because the first group is out of the range, in fact, there is only one group, which is taken by node 3.

In the second step all tasks without any predecessor, or with predecessors having already assigned teams, are being

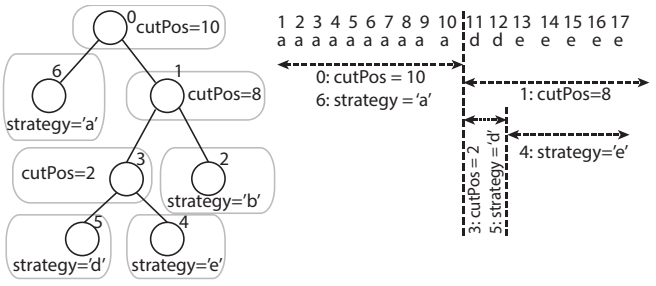


Fig. 7. The first step in genotype-to-phenotype mapping

searched for. For these tasks, teams are assigned according to their strategy. Then the step is repeated as long as there are tasks without assigned teams.

In the third step, the total cost of the solution could be calculated. For this purpose, the cost of each team from the resource library (Table I) is given.

D. Parameters of DGP

During the evolution, new populations of supervisors are created using genetic operations: reproduction, crossover (recombination) and mutation. After the genetic operations are performed on the current population, a new population replaces the current one. The number of individuals in each population is always equal:

$$\Pi = \alpha \prod_{i=1}^{n} s_i \tag{1}$$

where n is the number of tasks, s is the number of teams capable to solve specific problem and α is a constant between 0 and 1. If α is equal to 1, the population has as many individuals as many solutions of the problem exist. The evolution is controlled by parameters β, γ and δ, such that:

- $\Phi = \beta \cdot \Pi$ is the number of individuals created using the reproduction,
- $\Psi = \gamma \cdot \Pi$ is the number of individuals created using the crossover,
- $\Omega = \delta \cdot \Pi$ is the number of individuals created using the mutation and
- $\beta + \gamma + \delta = 1$

The last condition ensures that each of the created population will have the same number of individuals.

Finally, the selection of the best individuals by a tournament is chosen [21]. In this method, chromosomes (genotypes) are drawn with the same probability in quantity defined as a size of the tournament. From the drawn chromosomes the best one is taken. Hence, the tournament is repeated as many times as the number of chromosomes for a reproduction, crossover and mutation is required . A size of the tournament should not be too high, because the selection pressure is too strong and the evolution will be too greedy. It also could not be too low, because the time of finding any better result would be too long.

E. Fitness function

A fitness function determines the aim of DGP. In the presented approach, two options are possible. In the first one, the cheapest solution which has to be finished before a deadline is searched for. Such fitness function is applied when hard real time constraints have to be satisfied. In the second one, the DGP should find the fastest solution, which does not exceed a given budget. This case concerns systems with soft real-time requirements.

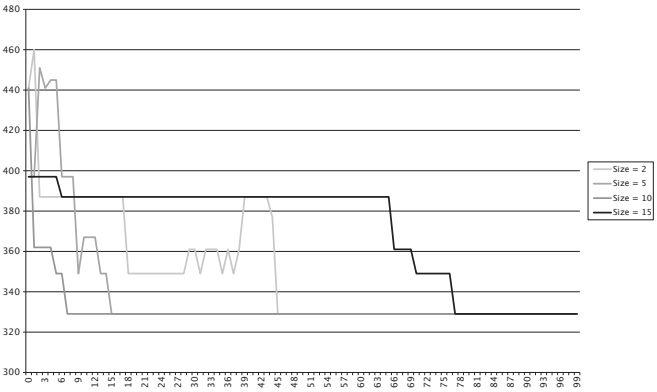


Fig. 8. Progress of the evolution for different tournament sizes

VI. Computational Experiments

High efficiency of the DGP may be achieved only when genetic parameters will be properly adjusted. The most important are the number of individuals created during reproduction, crossover, mutation and a size of the tournament. If their values are not chosen correctly, the DGP will find solutions far from the optimum or finding the best solution will take a lot of time.

The presented method was evaluated with the help of the example from Fig. 1. The deadline was equal 700 time units. During the experiments, the following values of genetic parameters were used:

- the evolution was stopped after 100 generations,
- each experiment was repeated 7 times, the mean of the best solutions received in each pass is given as the result,
- parameter α was equal $7 \cdot 10^{-30}$, thus the population size was equal 102.

First, the convergence of the DGP for different sizes of the tournament was explored. Tournament sizes equal 2, 5, 10 and 15 were examined (Fig. 8). We observed that the sizes 2 and 5 were too small. Very often the best solutions were skipped over, and not selected for further evolution. The best convergence was obtained for the size equal 10.

Next, we examined the influence of the crossover and the mutation on finding the best solution. For this purpose, an analysis of the best solutions achieved with different combination of parameters controlling the crossover and the mutation was performed. Fig. 9 presents the results obtained for different number of mutations. We may observe that for less than 10% of mutants in the population, the results are poor. The best result was obtained for 18% of the mutants. The number of mutants should not be too high. For more than 85% of the mutants the DGP usually produces also poor results. In this case, too much number of mutants probably disturbs the evolution. Fig. 10 presents results obtained for different numbers of individuals created using the crossover. The highest probability of obtaining the best results is when the crossover is applied for creation of at least 65% genotypes.

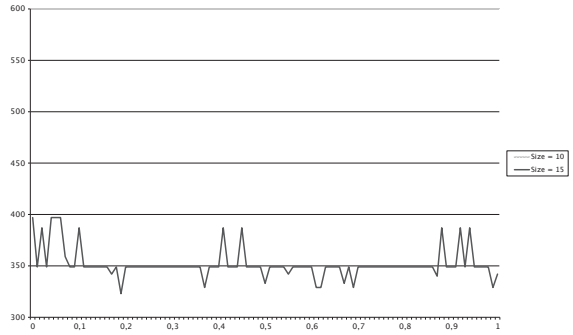


Fig. 9. The influence of δ

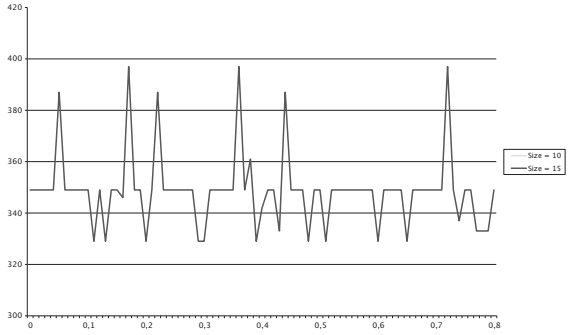


Fig. 10. The influence of γ

The results were also compared with greedy approach. The greedy algorithm assigns the cheapest teams to first tasks. However, if the deadline is to be exceeded, the algorithm assigns the fastest (usually the most expensive) teams to tasks at the end of the schedule as it is necessary to finish the work before the deadline. To the problem the greedy algorithm generates a solution which costs 1091 while the cost of the solution generated by the DGP is equal to 323. Thus, the result obtained with the help of our method is three times better.

To check whether our method led to global optimum or not, the entire space of the solutions was tested (Fig. 11). Out of $7.63 \cdot 10^{11}$ solutions, only 260 gave the best schedules, and a cost of the cheapest schedule was 323. So the answer is positive.

VII. Conclusions

A two-step procedure for automatic creation of supervisors of RT MOOS has been formulated. In the first step one has to specify functionalities of the system using early UML models (use case, activity, and sequence diagrams) and transform the models into a task graph and a library of resources of the system. In the second step one has to define decision options of a supervisor of the system and develop a genotype of the supervisor using DGP. An application of the procedure to RT MOOS, which was an enterprise of building houses, resulted in the creation of the best supervisor in acceptable time. It was evaluated from different points of view. Efficiency and precision were taken into account. Experiments showed

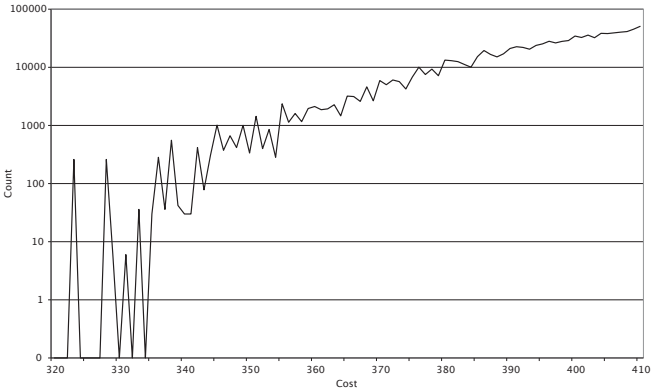


Fig. 11. The number of solutions to the problem for a given cost

that the supervisor can develop the best schedule, which corresponds to the global optimum.

Variant of the RCPSP presented in this paper is based on teams that are able to solve tasks. In general, a team is a group of workers and tools, but could also contain only one member. For different tasks, teams could contain different members, thus a team is able to start working when all of its members are idle. For this reason, it is possible, that choosing only the fastest teams do not yield the fastest solution. For the other hand, choosing only the cheapest teams could lead to the situation when the deadline of solution will be exceeded.

The influence of genetic parameters on the evolution of the genotype was also investigated. The most important parameter is δ, which defines the number of mutations. If it is too small or too large, DGP will have problems with escaping from local minima or it will behave too randomly. For the example presented in this paper, the best value for this parameter is about 20%.

The significant influence on the optimization has also a size of the tournament. If it is too small, DGP needs more time to find acceptable results. In the opposite case, if this value is too big, it leads DGP very quickly close to the global minimum, but has never achieved it. Actually, it stuck in local minima, because a variety of the population is decreasing.

The γ parameter corresponds to the number of crossovers. The crossover is responsible for a genotype tree development. Changing a genotype tree has also a bit similar effect of mutations, because after changing a position of the branch in the tree a new assignment of teams to tasks is achieved. Although this parameter has the smallest influence on results, it could be noticed that too high value of γ makes the solutions more random. If the value is too small, DGP need more time to achieve the optimal solution.

References

[1] C.Wei, P. Liu, Y. Tsai, “Resource-constrained project management using enhanced theory of constraint”, International Journal of Project Management, Vo. 20, No.7, 2002, pp.561-567. http://dx.doi.org/10.1016/S0263-7863(01)00063-1

[2] J. Blazewicz, J.K. Lenstra, A.H.G. Rinnooy Kan, Scheduling subject to resource constraints: Classification and complexity, Discrete Applied Mathematics, No.5,1983, pp.11-24. http://dx.doi.org/10.1016/0166-218X(83)90012-4

[3] Pawiński G. and Sapiecha K., "Cost-efficient Project Management Based on Distributed Processing Model.", Proceedings of The 2013 21st Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, Belfast 2013 http://dx.doi.org/10.1109/PDP.2013.30

[4] R. H. Möhring, A. S. Schulz, F. Stork, M. Uetz, "Solving Project Scheduling Problems by Minimum Cut Computations", Management Science, v.49 n.3, pp.330-350, March 2003. http://dx.doi.org/10.1287/mnsc.49.3.330.12737

[5] Hartmann S., Briskorn D., A survey of variants and extensions of the resource-constrained project scheduling problem, European journal of operational research : EJOR. - Amsterdam : Elsevier, Vol. 207., 1 (16.11.), pp. 1-15 (2010). http://dx.doi.org/10.1016/j.ejor.2009.11.005

[6] J.H.Holland, "Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology", Control, and Artificial Intelligence, University of Michigan Press, Ann Arbor, MI (reprinted 1992, MIT Press, Cambridge, MA).

[7] R.E.Keller, W.Banzhaf, "The evolution of genetic code in genetic programming", Proc. of the Genetic and Evolutionary Computation Conference, 1999, pp.1077-1082.

[8] G. Wilson, M. Heywood "Probabilistic Adaptive Mapping Developmental Genetic Programming (PAM DGP): A New Developmental Approach", Proceedings of the 9th International Conference on Parallel Problem Solving from Nature (PPSN IX), (Reykjavik 2007) http://dx.doi.org/10.1007/11844297_76

[9] L. C. Briand, Y. Labiche, A UML-Based Approach to System Testing, Software and Systems Modeling, vol. 1 (1), pp. 10-42, 2002. http://dx.doi.org/10.1007/s10270-002-0004-8

[10] J.L.M. Pasaje, M.G. Harbour, J.M. Drake, "MAST Real-Time View: a graphic UML tool for modeling object-oriented real-time systems", In proceeding of: IEEE 22nd Real-Time Systems Symposium, 2001. (RTSS 2001). http://dx.doi.org/10.1109/REAL.2001.990618

[11] H. Gomaa, Designing Concurrent, Distributed, and Real-Time Applications with UML. Addison-Wesley, 2000.

[12] R.Jigorea, S. Manolache, P.Eles, Zebo Peng, "Modelling of real-time embedded systems in an object-oriented design environment with UML," Proceedings. Third IEEE International Symposium on Object-Oriented Real-Time Distributed Computing, 2000, pp.210-213. http://dx.doi.org/10.1109/ISORC.2000.839532

[13] Alcaraz, J., & Maroto, C. (2001). A robust genetic algorithm for resource allocation in project scheduling. Annals of Operations Research, 102, 83-109. http://dx.doi.org/10.1023/A:1010949931021

[14] Hartmann, S. (1998). An competitive genetic algorithm for resource-constrained project scheduling. Naval Research Logistics, 45(7), 733-750. http://dx.doi.org/10.1002/(SICI)1520-6750(199810)45:7%3C733::AID-NAV5%3E3.3.CO;2-7

[15] Xiang Li, Lishan Kang, Wei Tan, "Optimized Research of Resource Constrained Project Scheduling Problem Based on Genetic Algorithms", Lecture Notes in Computer Science, Vol. 4683, 2007, pp 177-186. http://dx.doi.org/10.1007/978-3-540-74581-5_19

[16] Hossein Zoulfaghari, Javad Nematian, Nader Mahmoudi, and Mehdi Khodabandeh. 2013. A New Genetic Algorithm for the RCPSP in Large Scale. Int. J. Appl. Evol. Comput. 4, 2 (April 2013), 29-40. http://dx.doi.org/10.4018/jaec.2013040103

[17] Koza, J., Bennett III , F. H., Andre, D., Keane, M. A., 1998. Evolutionary Design of Analog Electrical Circuits Using Genetic Programming. In: I. C. Parmee (ed.), Adaptive Computing in Design and Manufacture. http://dx.doi.org/10.1007/978-3-540-85857-7_8

[18] S.Deniziak, A.Górski, "Hardware/Software Co-Synthesis of Distributed Embedded Systems Using Genetic Programming", Lecture Notes in Computer Science, Springer-Verlag, 2008, pp.83-93. http://dx.doi.org/10.1007/978-3-540-85857-7_8

[19] J.R.Koza, R.Poli, "Genetic Programming", In Edmund Burke and Graham Kendal, editors. "Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques", Chapter 5. Springer, 2005. http://dx.doi.org/10.1007/0-387-28356-0_5

[20] Binder R. V., Testing Object-Oriented Systems - Models, Patterns, and Tools, Addison-Wesley (1999)

[21] Z. Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs, Springer-Verlag Berlin Heidelberg, 1996.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 511–518
DOI: 10.15439/2014F33
ACSIS, Vol. 2

A Comparison between Different Chess Rating Systems for Ranking Evolutionary Algorithms

Niki Večeka*, Matej Črepinšeka*, Marjan Mernika* and Dejan Hrnčič†
*Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia
Email: {niki.vecek, marjan.mernik, matej.crepinsek}@uni-mb.si
†Adacta d.o.o., Maribor, Slovenia
Email: dejan.hrncic@adacta.si

***Abstract*—Chess Rating System for Evolutionary algorithms (CRS4EAs) is a novel method for comparing evolutionary algorithms which evaluates and ranks algorithms regarding the formula from the Glicko-2 chess rating system. It was empirically shown that CRS4EAs can be compared to the standard method for comparing algorithms - null hypothesis significance testing. The following paper examines the applications of chess rating systems beyond Glicko-2. The results of 15 evolutionary algorithms on 20 minimisation problems obtained using the Glicko-2 system were empirically compared to the Elo rating system, Chessmetrics rating system, and German Evaluation Number (DWZ). The results of the experiment showed that Glicko-2 is the most appropriate choice for evaluating and ranking evolutionary algorithms. Whilst other three systems' benefits were mainly the simple formulae, the ratings in Glicko-2 are proven to be more reliable, the detected significant differences are supported by confidence intervals, the inflation or deflation of ratings is easily detected, and the weight of individual results is set dynamically.**

***Index Terms*—chess rating system, ranking, evolutionary algorithms comparison, Glicko-2, Elo, Chessmetrics**

I. Introduction

A METHOD for comparing the algorithms is needed for determing whether one algorithm performs better than the other. As numerous effective evolutionary algorithms are appearing, a comparison with only one algorithm is now insufficient. This fact leads to the need for determing which of the multiple algorithms is better than the other. Which of them is the best and which the worst? A well-established method for comparing the experimental results of multiple evolutionary algorithms is Null Hypothesis Significance Testing (NHST) [22]. Whilst there are many variants of NHST, there are still some pitfalls regarding statistics and its application [2], [7], [13], [21] that imply that this field still needs attention. A novel method the Chess Rating System for Evolutionary Algorithms (CRS4EAs) [30] suggests using a chess rating system for evaluating the results and ranking the algorithms. CRS4EAs treats (i) evolutionary algorithms as chess players, (ii) one comparison between two algorithms as one game, and (iii) execution and evaluation of pairwise comparisons between all algorithms participating in the experiment as a tournament. Just like the standard comparison of two algorithms, one game in CRS4EA can have three outcomes: the first algorithm is better (and therefore wins), the second algorithm is better (and therefore wins), or they perform equally regarding predefined accuracy ϵ (they play a draw). It has been empirically shown that CRS4EAs is comparable to NHST, and can be used as a comparative method for evolutionary algorithms [30]. A CRS4EAs method is used within an open-source framework Evolutionary Algorithms Rating System (EARS) [8], [9]. CRS4EAs and EARS were developed to provide fairer and easier to understand comparisons between evolutionary algorithms. All the experiments in EARS are executed for the same number of optimisation problems, the algorithms are written in the same programming language (Java), have the same termination criteria, are initialised with the same random seed, and executed under the same hardware configuration. Hence, some factors that could affect the final results of the experiment were excluded [30]. The CRS4EAs uses the Glicko-2 chess rating system [18], since it is one of the newest and it consists of many preferences that look promising. In the proposed paper the Glicko-2 rating system is compared to three other better-known and well-established rating systems: Elo, Chessmetrics, and German Evaluation Number (DWZ). In order to compare these four rating systems the experiment was conducted for 15 evolutionary algorithms covering 20 minimisation problems. The analysis showed that comparing evolutionary algorithms the Glicko-2 was the most appropriate choice. One downside to the Glicko-2 is its complicated formulae, for the understanding of which mathematical and statistical knowledge is needed. The differences amongst players are more straightforward in the other three systems, however they are unsupported by any reliable measurements - they are arbitrary. Otherwise, Glicko-2 was shown to be more reliable: the detected significant differences are supported by a confidence interval, straightforward measurement for rating reliability, the control of conservativity/liberty is more correct, the weightings of individual results are set dynamically, improvement through time is considered in final results, the inflation or deflation of ratings is easily detected, and the selective pairing is not an issue. This paper presents the reasons why the first choice for rating system used in CRS4EAs was the Glicko-2.

The paper is structured as follows. Section II summarises four more popular chess rating systems. The formulae used in these systems are adapted for EARS and are used during the experiment. The CRS4EAs method and the experiment are introduced in Section III, followed by a detailed analysis of the obtained results. Section IV concludes the paper.

978-83-60810-58-3/$25.00

II. BACKGROUND

Chess is a strategic game of two players with three possible outcomes: the first player can win, the first player can lose, or the players can play a draw. Usually, the absolute power of a chess player is described using a number that is called a 'rating'. A player's rating is updated after each tournament the player participates in and each chess organisation has its own rating system with formulae that evaluate its players. In this section the more common chess rating systems are introduced. All the players are represented on the leaderboard, from best to worst and although there are different formulae behind updating the players' ratings, all of them have two things in common: a player's rating is always a positive integer and the player with the highest rating value is expected to be better. A player joins the tournament with k opponents in which the ith player has a rating R_i, and plays m games.

A. Elo

The best-known chess rating system is the Elo rating system [10] where the expected score of the ith player against the jth player is calculated using the formula in Eq. 1.

$$E(R_i, R_j) = \frac{1}{1 + 10^{(R_j - R_i)/400}} \tag{1}$$

The expected score of the ith against the jth player is the probability of i defeating j. Hence, the sum of the expected scores of the ith and jth players (against each other) equals 1. The score the ith player gained against the jth player is denoted by $S_{i,j}$ and equals 1 if the ith player won, 0 if ith player lost, or 0.5 for a draw. All the ratings are updated at the end of a tournament using the formula from Eq. 2. The new rating of the ith player is denoted by R'_i.

$$R'_i = R_i + K \sum_{j=1}^{m} \big(S_{i,j} - E(R_i, R_j)\big) \tag{2}$$

The K-factor is a constant that affect the emphasis of the difference between the actual score and the expected score. The USCF (United States Chess Federation) rating system implements the K-factor by dividing 800 by the sum of effective number of games a player's rating is based on (N_e) and the number of games the player completed during a tournament (m) [17]. Even though, the Elo system is famous for its simplicity and wide-usage, it has a few drawbacks such as properly setting the K-factor, an inaccurate distribution model, or unreliable rating.

B. Chessmetrics

The chess statistician Jeff Sonas proposed the usage of a more dynamic K-factor in his own chess rating system called Chessmetrics [27], described as 'a weighted and padded simultaneous performance rating'. Chessmetrics uses the following formula (Eq. 3) for updating the rating of the ith player.

$$R'_i = 43 + \frac{R_{per} * m + 4 * \sum_{j=1}^{k} R_j/k + 2300 * 3}{m + 7} \tag{3}$$

R_{per} is the performance rating calculated as $\sum_{j=1}^{k} R_j/k + (\sum_{j=1}^{m} S_{i,j}/m - 0.5) * 850$ and with the meaning that each 10% increase in percentage score corresponds to an 85 point advantage in the ratings [27].

C. German Evaluation Number (DWZ)

The simplest and one of the first rating systems was the Ingo rating system [20] by Anton Hoesslinger (1948), which has influenced many other rating system, including the Deutsche Wertungszahl (DWZ) [6]. DWZ is similar to the Elo rating system of the FIDE (World Chess Federation) but has improved in its own way since 1990 when it was first introduced. The expected score in DWZ is calculated using the same formula as the expected score in the Elo system (Eq. 1), whilst the rating is updated using the formula in Eq. 4.

$$R'_i = R_i + \sum_{j=1}^{m} \frac{800}{D + m} \big(S_{i,j} - E(R_i, R_j)\big) \tag{4}$$

D is the development coefficient (Eq. 5), dependent on the fundamental value D_0 (Eq. 6), the acceleration factor a (Eq. 7), and the breaking value b (Eq. 8).

$$\begin{aligned} D &= a * D_0 + b \\ 5 \leq D &\leq \begin{cases} \min(30, 5i) & \text{if } b = 0 \\ 150 & \text{if } b > 0 \end{cases} \end{aligned} \tag{5}$$

$$D_0 = \big(\frac{R_i}{1000}\big)^4 + J \tag{6}$$

The coefficient J differs according to the different ages of the players - the older the player, the bigger the J. The acceleration factor a (Eq. 7) cannot be higher than 1 or lower than 0.5, and is calculated only if a player younger than 20 years achieved more points than expected, otherwise a equals 1. The breaking value b (Eq. 8) is computed only if the player with a rating under 1300 achieved less points than expected, otherwise b equals 0.

$$a = \frac{R_i}{2000} \tag{7}$$

$$b = e^{\frac{1300 - R_i}{150}} - 1 \tag{8}$$

D. Glicko-2

One of main concerns about the Elo system is the possibility of a player winning the game and losing rating points, or losing the game and gaining rating points. Problems with unreliable ratings show in those games between players with the same rating, when one of them has not played for years and the other plays constantly - they would lose and gain the same amount of points. A less reliable rating is expected for the player who has not played in years, and a more reliable rating for the player who plays constantly. It is expected that if the first player wins his rating would go up more than the rating of the second player goes down. Because anything cannot be said about the player's gaming behaviour or the reliability of his power, Glickman [14] introduced a new chess rating system. The Glicko system [15] introduces a new value that represents

the reliability of a player's power - rating deviation RD - which is similar to standard deviation regarding statistics. RD_i is set to 350 at the beginning of the first tournament and updated (just as rating) at the end of each tournament. It decreases with each tournament the ith player participates in and increases with each tournament ith player skips. The maximum value of RD is 350, whilst the minimum is set by an organisation implementing the system (Glickman suggests 30). Rating deviation tells how reliable the player's rating is - the lower the RD the more reliable the rating. In 2012 Glickman updated its system and presented the Glicko-2 rating system [18], which is based on Glicko but has another variable that presents the reliability of the player's strength - rating volatility σ_i. The volatility indicates the degree of expected fluctuation in a player's rating. If σ_i is low the player performs at a consistent level, whilst high σ_i indicates erratic performances. Firstly, the rating R and rating deviation RD have to be converted from Glicko to the Glicko-2 rating system (Eq. 9).

$$\mu = \frac{R - 1500}{173.7178} \text{ and } \phi = \frac{RD}{173.7178} \tag{9}$$

The estimated variance v of the player's rating based only on game outcomes is calculated using the formula in Eq. 10.

$$v = \left(\sum_{j=1}^{m} g(\phi_j)^2 E(\mu_i, \mu_j, \phi_i)(1 - E(\mu_i, \mu_j, \phi_i)) \right)^{-1} \tag{10}$$

The gravity factor g (Eq. 11) and the expected score E (Eq. 12) are calculated using the following formulae.

$$g(\phi) = \frac{1}{\sqrt{1 + 3\phi^2/\Pi^2}} \tag{11}$$

$$E(\mu, \mu_i, \phi_i) = \frac{1}{1 + 10^{-g(\phi_i)(\mu - \mu_i)}} \tag{12}$$

Next, the estimated improvement in rating Δ (Eq. 13) has to be calculated where the pre-period rating μ_i is compared to the performance rating μ_j based only on the game outcomes $S_{i,j}$.

$$\Delta = v \sum_{j=1}^{m} g(\phi_j)(S_{i,j} - E(\mu_i, \mu_j, \phi_i)) \tag{13}$$

A new rating volatility σ' is found when using the Illinois algorithm [5] for a function $f(x) = \frac{e^x(\Delta^2 - \phi^2 - v - e^x)}{2(\phi^2 + v + e^x)^2} - \frac{x - \ln(\sigma^2)}{\tau^2}$ with accuracy of up to 6 decimal places. This method is used for finding zeros and once the zero x_0 of this function is found, σ' is set to $e^{x_0/2}$ and the pre-rating period value ϕ^* (Eq. 14) is calculated.

$$\phi^* = \sqrt{\phi^2 + \sigma'^2} \tag{14}$$

New values for rating deviation ϕ' (Eq. 15) and rating μ' (Eq. 16) are set.

$$\phi' = \frac{1}{\sqrt{\frac{1}{(\phi^*)^2} + \frac{1}{v}}} \tag{15}$$

$$\mu' = \mu + \phi' \sum_{i=1}^{m} g(\phi_i)(S_i - E(\mu, \mu_i, \phi_i)) \tag{16}$$

Finally, the new rating R' and new rating deviation RD' are converted from the Glicko-2 to the Glicko system using the formulae in Eq. 17.

$$R' = 173.7178\mu' + 1500 \text{ and } RD' = 173.7178\phi' \tag{17}$$

All of these systems have their own advantages (Table I), however, Glicko-2 contains most of them despite its more complicated formula (in comparison with other systems).

TABLE I: Preferences a chess rating contains.

Preference	Elo	Chessmetrics	DWZ	Glicko-2
Simple formula	✔	✔	✔	
Player's age influence			✔	
Dynamic weight factor		✔	✔	✔
Control over selective pairing				✔
Time varying impact				✔
Bayesian approach				✔
Straightforward measurement of rating inflation and deflation				✔
Straightforward measurement of rating reliability				✔
Straightforward measurement of differences between ratings				✔

Our implementations of these four algorithms were used in the following experiment.

III. Experiment

This experiment was conducted using the novel method for comparing and ranking the evolutionary algorithms CRS4EAs [30]. The experiment in CRS4EAs is executed as any other experiment, however each outcome of each algorithm regarding every optimisation problem must be saved for further comparison. In the CRS4EAs the run-by-run comparison the roles of the chess players adopt evolutionary algorithms. Each outcome in every run for every optimisation problem of one algorithm is compared to the corresponding outcome of the other algorithm. Such a comparison is called one 'game'. If the difference between compared outcomes is less than the predefined ϵ, the final score of this game is a draw, otherwise the algorithm with the outcome closer to the optimum of the optimisation problem wins. With k algorithms ($k - 1$ opponents), N optimisation problems, and n runs, one algorithm plays $n * N * (k - 1)$ games during one tournament. Hence, in our tournament $n * N * k * (k - 1)/2$ games are played. The whole process is presented in the flowchart in Fig. 1. The chess rating system used in CRS4EAs is Glicko-2, however due to this being an experiment, other chess rating systems were implemented as well.

In the presented experiment our implementations of $k = 15$ evolutionary algorithms were compared for $N = 20$ optimisation problems over $n = 100$ runs. The simplest algorithm used in the experiment was the basic random search (RWSi) [24]. Next being Teaching Learning Based Optimization (TLBO) [3], [25]. There were two variants of evolutionary strategies (ES(1+1) and CMA-ES) [19], [26], 10 variants of the Differential Evolution [4], [23], [29], [31], and the Self-adaptive

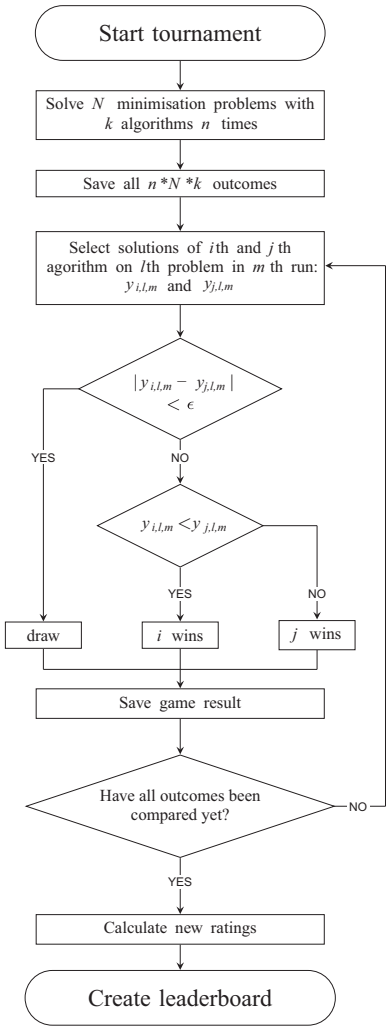


Fig. 1: Flowchart of experiment's execution in CRS4EAs.

Differential Evolution (jDE) [1]. The optimisation problems were from the Special Session and Competition on Large-Scale Global Optimization CEC 2010 [28]. The termination critera for each algorithm was maximum number of evaluations $Max\_FEs = 10^5$. The threshold for a draw was set at $\epsilon = 1.0e-06$, and the initial rating was set to 1500 for each rating system to provide a fairer comparison. Detailed descriptions of the algorithms and optimisation problems can be found in [30]. Other properties for the rating systems can be seen in Table II. R_{init} represents the initial rating for a new player, rating intervals and rating ranges present the values for detecting the differences in the powers of the algorithms, K is the K-factor in Elo, N_e is the number of games a player's rating is based on, m is the number of games the player completed during a tournament, J is the age coefficient from Chessmetrics, RD_{init} is the initial rating deviation for a new player, RD_{min} is the minimum rating deviation, and RD_{max} is the maximum rating deviation. Whilst Glicko-2 uses the straightforward values for detecting significant differences - R, RD, and rating interval - other systems do not consist of such preferences. Two algorithms are significantly different if their rating intervals do not overlap. In Glicko-2 the 99.7% rating (confidence) interval is defined by $[R-3RD, R+3RD]$. The rating range that distinguishes between the powers of two players in Elo equals 200 rating points. The minimum Elo rating can be 100 points, then the players are classified in categories: 100-199 points (J), 200-399 points (I), 400-599 points (H), 600-799 points (G), 800-999 points (F), 1000-1199 points (E), 1200-1399 points (D), 1400-1599 points (C), 1600-1799 points (B), 1800-1999 points (A), 2000-2199 points (expert), etc. The same is done for DWZ and Chessmetrics, but while DWZ uses the same categories as Elo, the Chessmetrics' categories differ by 100 points. However, it must be explicitly pointed out that this classification of categories is not a straighforward way of detecting significant differences as the confidence intervals in Glicko-2.

TABLE II: Properties for the chess ratings used during the experiment.

Chess rating system	Properties
Elo	$R_{init} = 1500$
	Rating range 200 points
	$K = \frac{800}{N_e+m}$
Chessmetrics	$R_{init} = 1500$
	Rating range 100 points
DWZ	$R_{init} = 1500$
	$J = 15$ for all players
	Rating range 200 points
Glicko-2	$R_{init} = 1500$
	$RD_{init} = 350$, $RD_{min} = 50$, $RD_{max} = 350$
	Rating interval $[R-3RD, R+3RD]$

The ratings evolutionary algorithms obtained for each rating system are presented on a group leaderboard in Table III. All the algorithms obtained minimum rating deviations of 50 points in the Glicko-2 system. Although, different formulae were used in different chess rating systems, the orders of the ratings were almost the same. The only rating system for which the order of algorithms was different was Elo where CMA-ES, DE/best/2/exp and DE/rand/1/exp go in reverse order. These three algorithms, however, are really close regarding rating points. In order to obtain a better picture the average ranking of the algorithms by data sets, i.e. Friedman ranking [11], [12] was added in the last column. A statistical analysis and comparison with NHST can be found in [30]. All the obtained ratings are displayed in Fig. 2 where distributions of ratings for each rating system are shown. Maximum and minimum overall rating values were obtained in Elo. These ratings were more scattered and there was a big gap (435 points) between the 7th algorithm DE/best/1/exp and the 8th algorithm DE/best/2/bin by dividing the algorithms into two groups: algorithms from 1 to 7 and algorithms from 7 to 15. The Chessmetrics and DWZ ratings seemed to be equally distributed, but the difference between the corresponding rating points varied between 20 to 59 points. The difference was bigger for better performing algorithms

TABLE III: Leaderboard with ratings the algorithms obtained using four different rating systems and the average ranking (AR) of the algorithms by data sets, i.e. Friedman ranking [11], [12].

i	Algorithm	Elo	Chessmetrics	DWZ	Glicko-2	AR
1	JDE/rand/1/bin	2014	1812	1753	1829	3.6
2	DE/rand/2/exp	1996	1772	1715	1779	3.425
3	CMA-ES	**1972**	1767	1711	1774	4.325
4	DE/best/2/exp	**1982**	1761	1705	1766	4.675
5	DE/rand/1/exp	**1985**	1758	1702	1762	4.325
6	DE/rand/2/bin	1940	1704	1651	1696	4.75
7	DE/best/1/exp	1890	1626	1578	1602	7.675
8	DE/best/2/bin	1455	1588	1542	1554	7.975
9	DE/rand-to-best/1/exp	1361	1575	1530	1540	7.05
10	DE/rand/1/bin	1221	1516	1475	1467	8.5
11	DE/rand-to-best/1/bin	1129	1375	1342	1294	10.8
12	TLBO	1078	1297	1268	1199	12.05
13	DE/best/1/bin	1057	1268	1241	1164	12.55
14	RWSi	1000	1178	1156	1054	13.7
15	ES(1+1)	983	1151	1131	1020	14.6

(59 for JDE/rand/1/bin) and smaller for worse performing algorithms (20 for ES(1+1)). The biggest gap in ratings for Glicko-2, DWZ, and Chessmetrics was between the 10th algorithm DE/rand/1/bin and the 11th algorithm DE/rand-to-best/1/bin. Algorithms DE/rand/2/exp ($i = 2$), CMA-ES ($i = 3$), DE/best/2/exp ($i = 4$), and DE/rand/1/exp ($i = 5$) were close in ratings for all four rating systems.

An interesting outlook regarding the results of a tournament is when examinating wins, losses, and draws (Table IV). This is not only useful in chess but also in comparison with evolutionary algorithms. The number of wins, losses, and draws can tell a lot about how one algorithm performed against another. For example, JDE/rand/1/bin was the overall best algorithm - it had the most wins and the least losses - but when comparing its performance with the performance of the worst algorithm ES(1+1) - with the least wins and the most losses - showed that ES(1+1) defeated JDE/rand/1/bin in 1 out of 2000 (=20*100) games. It could be concluded that the JDE/rand/1/bin performed with outliers as this is a phenomenon that is also detected with other worse algorithms: DE/rand-to-best/1/bin (2 outliers), TLBO (2 outliers), DE/best/1/bin (2 outliers), and RWSi (2 outliers). An interesting fact is that CMA-ES has more wins than DE/rand/2/exp but is ranked one place lower. This is due to the fact that CMA-ES also has more loses and less draws. However, as mentioned before the difference in ratings is small. Table IV also shows that the draws were less common in those games with low-ranked algorithms - even between the low-ranked algorithms themselves. The draws were fairly common in games between the first half of the algorithms, whilst in games with algorithms that were ranked lower than 8th place the draws hardly appeared. The most draws (1112) were played between DE/rand/2/exp and DE/rand/1/exp. DE/rand/2/exp, DE/rand/2/bin, and DE/rand-to-best/1/exp were the only three algorithms that won the absolute number of games (2000) against at least one opponent. DE/rand/2/exp won absolutely against TLBO, DE/best/1/bin, RWSi, and ES(1+1), DE/rand/2/bin against RWSi, and DE/rand-to-best/1/exp against ES(1+1).

The detected significant differences are shown in Fig. 3. As Chessmetrics has the lowest threshold for classifying players into groups (100 rating points), the highest distinctions (90) between players were detected within this system. Elo and DWZ had the same threshold (200 rating points), but more distinctions were detected in Elo, due to the fact that the obtained players' ratings in Elo had wider ranges. Chessmetrics detected 10 differences more than DWZ, 8 differences more than Elo, and there was no difference in those detected by DWZ or Elo and those Chessmtrics was not. DWZ detected 8 differences that Elo did not, and Elo detected 11 differences that DWZ did not. These differences are listed in Table V.

TABLE V: System marked with ✓ detected differences in the ratings of the listed algorithms, whilst the system marked with ✗ did not.

Chessmetrics ✓ DWZ ✗	Chessmetrics ✓ Elo ✗
JDE/rand/1/bin vs. DE/rand/2/exp	DE/rand/2/exp vs. DE/best/1/exp
JDE/rand/1/bin vs. CMA-ES	CMA-ES vs. DE/best/1/exp
JDE/rand/1/bin vs. DE/best/2/exp	DE/best/2/exp vs. DE/best/1/exp
JDE/rand/1/bin vs. DE/rand/1/exp	DE/rand/1/exp vs. DE/best/1/exp
JDE/rand/1/bin vs. DE/rand/2/bin	DE/rand/2/bin vs. DE/best/1/exp
DE/best/1/exp vs. DE/best/2/bin	DE/rand-to-best/1/bin vs. TLBO
DE/best/1/exp vs. DE/rand-to-best/1/exp	DE/rand-to-best/1/bin vs. DE/best/1/bin
DE/best/1/exp vs. DE/rand/1/bin	DE/rand-to-best/1/bin vs. RWSi
DE/rand-to-best/1/bin vs. TLBO	
DE/rand-to-best/1/bin vs. DE/best/1/bin	

DWZ ✓ Elo ✗	Elo ✓ DWZ ✗
DE/rand/2/exp vs. DE/best/1/exp	JDE/rand/1/bin vs. DE/rand/2/exp
CMA-ES vs. DE/best/1/exp	JDE/rand/1/bin vs. CMA-ES
DE/best/2/exp vs. DE/best/1/exp	JDE/rand/1/bin vs. DE/best/2/exp
DE/rand/1/exp vs. DE/best/1/exp	JDE/rand/1/bin vs. DE/rand/1/exp
DE/rand/2/bin vs. DE/best/1/exp	JDE/rand/1/bin vs. DE/rand/2/bin
DE/rand-to-best/1/bin vs. RWSi	DE/best/1/exp vs. DE/best/2/bin
TLBO vs. RWSi	DE/best/1/exp vs. DE/rand-to-best/1/exp
DE/best/1/bin vs. RWSi	DE/best/1/exp vs. DE/rand/1/bin
	DE/best/2/bin vs. DE/rand-to-best/1/exp
	DE/best/2/bin vs. DE/rand/1/bin
	RWSi vs. ES(1+1)

It appears that Elo, Chessmetrics, and DWZ are more liberal, and the conservativity could be increased with a wider rating range between categories. However controlling the conservativity in such way would not be as efficient as in Glicko-2 where conservativity is controlled by setting the minimal rating deviation and choosing an appropriate confidence interval. In Glicko-2 the algorithms' ratings were compared pairwisely, whilst with the other three systems algorithms were classified into groups and then compared regarding them. Also, the significances of the differences detected within Elo, Chessmetrics, and DWZ are unknown, as there was no statistical tool for measuring them and the choice of rating range is arbitrary. On the other hand, Glicko-2 detected 50 significant differences that were made with 99.7% confidence and were comparable to NHST [30]. The tests of significance used for NHST analysis were the Friedman

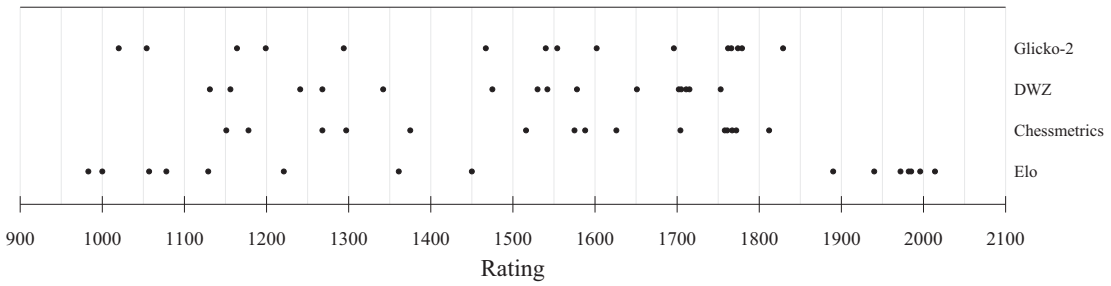


Fig. 2: Distributions of the ratings for all 4 rating systems.

TABLE IV: Wins (W), losses (L), and draws (D) algorithms obtained in tournaments against each other. When the number of losses is relatively small (e.g., < 75) the occurance of outliers is very likely.

i		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Σ
	W	/	642	769	750	775	747	1417	1389	1919	1976	1998	1998	1998	1998	1999	20375
1 JDE/rand/1/bin	L	/	423	659	407	365	397	64	245	67	24	2	2	2	2	1	2660
	D	/	935	572	843	860	856	519	366	14	0	0	0	0	0	0	4956
	W	423	/	711	519	474	624	1182	1299	1690	1843	1993	2000	2000	2000	2000	18758
2 DE/rand/2/exp	L	642	/	665	502	414	380	298	334	305	157	7	0	0	0	0	3704
	D	935	/	624	979	1112	996	520	367	5	0	0	0	0	0	0	5538
	W	659	665	/	765	688	853	1295	1371	1662	1843	1901	1910	1942	1898	1968	19420
3 CMA-ES	L	769	711	/	593	700	559	246	164	334	157	99	90	58	102	32	4614
	D	572	624	/	642	612	606	459	465	4	0	0	0	0	0	0	3984
	W	407	502	593	/	570	684	1175	1314	1633	1774	1992	1982	1999	1998	1999	18622
4 DE/best/2/exp	L	750	7519	765	/	547	462	309	306	356	226	8	18	1	2	1	11270
	D	843	979	642	/	883	854	516	380	11	0	0	0	0	0	0	5108
	W	365	414	700	547	/	555	1157	1244	1713	1846	1969	1976	1987	1984	1996	18453
5 DE/rand/1/exp	L	775	474	688	570	/	537	353	390	279	154	31	24	13	16	4	4308
	D	860	1112	612	883	/	908	490	366	8	0	0	0	0	0	0	5239
	W	397	380	559	462	537	/	1028	1080	1525	1584	1671	1905	1759	2000	1851	16738
6 DE/rand/2/bin	L	747	624	853	684	555	/	460	552	471	416	329	95	241	0	149	6176
	D	856	996	606	854	908	/	512	368	4	0	0	0	0	0	0	5104
	W	64	298	246	309	353	460	/	1114	1103	1333	1891	1924	1987	1977	1993	15052
7 DE/best/1/exp	L	1417	1182	1295	1175	1157	1028	/	530	893	667	109	76	13	23	7	9572
	D	519	520	459	516	490	512	/	356	4	0	0	0	0	0	0	3376
	W	245	334	164	306	390	552	530	/	979	1097	1691	1862	1985	1997	1996	14128
8 DE/best/2/bin	L	1389	1299	1371	1314	1244	1080	1114	/	1017	903	309	138	15	3	4	11200
	D	366	367	465	380	366	368	365	/	4	0	0	0	0	0	0	2681
	W	67	305	334	356	279	471	893	1017	/	1352	1982	1987	1999	1999	2000	15041
9 DE/rand-to-best/1/exp	L	1919	1690	1662	1633	1713	1525	1103	979	/	648	18	13	1	1	0	12905
	D	14	5	4	11	8	4	4	4	/	0	0	0	0	0	0	54
	W	24	157	157	223	154	416	667	903	648	/	1893	1930	1970	1974	1990	13106
10 DE/rand/1/bin	L	1976	1843	1843	1774	1846	1584	1333	1097	1352	/	107	70	30	26	10	14891
	D	0	0	0	0	0	0	0	0	0	/	0	0	0	0	0	0
	W	2	7	99	8	31	329	109	309	18	107	/	1618	1846	1976	1993	8452
11 DE/rand-to-best/1/bin	L	1998	1993	1901	1992	1969	1671	1891	1691	1982	1893	/	381	153	24	7	19546
	D	0	0	0	0	0	0	0	0	0	0	/	1	1	0	0	2
	W	2	0	90	18	24	95	76	138	13	70	381	/	1199	1967	1809	5882
12 TLBO	L	1988	2000	1910	1982	1976	1905	1924	1862	1987	1930	1618	/	801	33	191	22107
	D	0	0	0	0	0	0	0	0	0	0	1	/	0	0	0	1
	W	2	0	58	1	13	241	13	15	1	30	153	801	/	1770	1834	4932
13 DE/best/1/bin	L	1998	2000	1942	1999	1987	1759	1987	1985	1999	1970	1846	1199	/	230	166	23067
	D	0	0	0	0	0	0	0	0	0	0	1	0	/	0	0	1
	W	2	0	102	2	16	0	23	3	1	26	24	33	230	/	1501	1963
14 RWSi	L	1998	2000	1898	1998	1984	2000	1977	1997	1999	1974	1976	1967	1770	/	499	26037
	D	0	0	0	0	0	0	0	0	0	0	0	0	0	/	0	0
	W	1	0	32	1	4	149	7	4	0	10	7	191	166	499	/	1071
15 ES(1+1)	L	1999	2000	1968	1999	1996	1851	1993	1996	2000	1990	1993	1809	1834	1501	/	26929
	D	0	0	0	0	0	0	0	0	0	0	0	0	0	0	/	0

and Nemenyi tests with critical difference $CD = 4.79$. The first implied that there are significant differences between algorithms, and the other found 43 significant differences that were similar to those found with Glicko-2 (Fig. 3e). The executed experiment therefore showed that the Glicko-2 rating system is more appropriate for comparison and ranking of evolutionary algorithms. It provides more reliable ratings and more evident way of detecting significant differences. Hence, the preferences of the Glicko-2 (Table I) do not only contribute in tournaments between chess players but also in comparison between evolutionary algorithms.

IV. Conclusion

This paper conducted a comparison of four chess rating systems for ranking evolutionary algorithms. All the rating systems were implemented within EARS software, executed as an experiment, and analysed using the CRS4EAs method. The experiment showed that the Glicko-2 rating system is the most appropriate for ranking evolutionary algorithms. The main reason lies in the detection of significant differences amongst players and the formation of a confidence interval that allows direct comparison with null hypothesis significance testing. The other three systems - Elo, Chessmetrics, and DWZ - use simpler methods for detecting differences between ratings. Players are classified into categories and the differences in powers depend on the category the player belongs to. A new method for detecting the differences between players could increase the efficiencies of these systems, if the proposed method were dynamic (similar to Glicko-2). Otherwise, the results obtained from small tournaments (with a small number of algorithms or a small number of optimisation problems) would be unreliable. The conservativity/liberty of the method can be more efficiently controlled within Glicko-2. Elo, Chessmetrics, or DWZ can be improved by using some factors that are important for chess players (e.g., a player's age or the colour of pieces), but are irrelevant when comparing evolutionary algorithms. The results in CRS4EAs can be examined by observing the number of wins, losses, and draws amongst different players. Using this approach the outliers can be detected and the number of draws can indicate which algorithms are more likely to play a draw. In this paper we have empirically shown that various chess rating systems can be used for comparison amongst evolutionary algorithms and their rankings. The rationale as to why Glicko-2 may be a better choice than other chess systems for comparing the evolutionary algorithms has also been discussed in details. In the future, we will continue using Glicko-2 for CRS4EAs, with more focus on tuning the parameters. Glicko-2 was proven to be more reliable and dynamic than other older systems.

References

[1] J. Brest, S. Greiner, B. Bošković, M. Mernik, V. Žumer. Self-adapting control parameters in differential evolution: A comparative study on numerical benchmark problems. *IEEE Transactions on Evolutionary Computation*, 10(6):646–657, 2006.

[2] J. Cohen. The earth is round ($p < .05$). *American psychologist*, 49(12):997–1003, 1994.

[3] M. Črepinšek, S.H. Liu, L. Mernik. A Note on Teaching-Learning-Based Optimization Algorithm. *Information Sciences*, 212:79–93, 2012.

[4] S. Das, P.N. Suganthan. Differential evolution: A survey of the state-of-the-art. *IEEE Transactions on Evolutionary Computation*, 15(1):4–31, 2011.

[5] M. Dowell, P. Jarratt. A modified regula falsi method for computing the root of an equation. *BIT Numerical Mathematics*, 11(2):168–174, 1971.

[6] Deutscher Schachbund [Online]. Available: http://www.schachbund.de/wertungsordnung.html

[7] T. Dyba, V.B. Kampenes, D.I. Sjoberg. A systematic review of statistical power in software engineering experiments. *Information and Software Technology*, 48(8):745–755, 2006.

[8] Evolutionary Algorithms Rating System [Online]. Available: http://earatingsystem.appspot.com/

[9] Evolutionary Algorithms Rating System (Github) [Online]. Available: https://github.com/matejxxx/EARS

[10] A.E. Elo. The rating of chessplayers, past and present (Vol. 3). *Batsford*, 1978.

[11] M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32:675–701, 1937.

[12] M. Friedman. A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics*, 11:86–92, 1940.

[13] J. Gill. The insignificance of null hypothesis significance testing. *Political Research Quarterly*, 52(3):647–674, 1999.

[14] M.E. Glickman. A comprehensive guide to chess ratings. *American Chess Journal*, 3:59–102, 1995.

[15] M.E. Glickman. The glicko system. *Boston University*, 1995.

[16] M.E. Glickman. Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(3):377–394, 1999.

[17] M.E. Glickman. Dynamic paired comparison models with stochastic variances. *Journal of Applied Statistics*, 28(6):673–689, 2001.

[18] M.E. Glickman. Example of the Glicko-2 system. *Boston University*, 2012.

[19] N. Hansen. The CMA Evolution Strategy: A Comparing Review. *Towards a new evolutionary computation, Springer*, 75–102, 2006.

[20] D. Hooper, K. Whyld. The Oxford Companion to Chess. Oxford University Press, 1992.

[21] B.A. Kitchenham, S.L. Pfleeger, L.M. Pickard, P.W. Jones, D.C. Hoaglin, K. El Emam, J. Rosenberg. Preliminary guidelines for empirical research in software engineering. *IEEE Transactions on Software Engineering*, 28(8):721–734, 2002.

[22] J. Neyman, E. Pearson. On the problem of the most efficient test of statistical hypothesis. *Philosophical Transaction of the Royal Society of London - Series A*, 231:289–337, 1933.

[23] M.G. Epitropakis, V.P. Plagianakos, M.N. Vrahatis. Balancing the exploration and exploitation capabilities of the differential evolution algorithm. *IEEE World Congress on Computational Intelligence 2008*, 2686–2693, 2008.

[24] L.A. Rastrigin. The convergence of the random search method in the extremal control of a many-parameter system. *Automation and Remote Control*, 24(10):1337–1342, 1963.

[25] R.V. Rao, V.J. Savsani, D.P. Vakharia. Teaching-Learning-Based Optimization: An optimization method for continuous non-linear large scale problems. *Information Sciences*, 183(1):1–15, 2012.

[26] I. Rechenberg. Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution. *Frommann-Holzboog*, 1973.

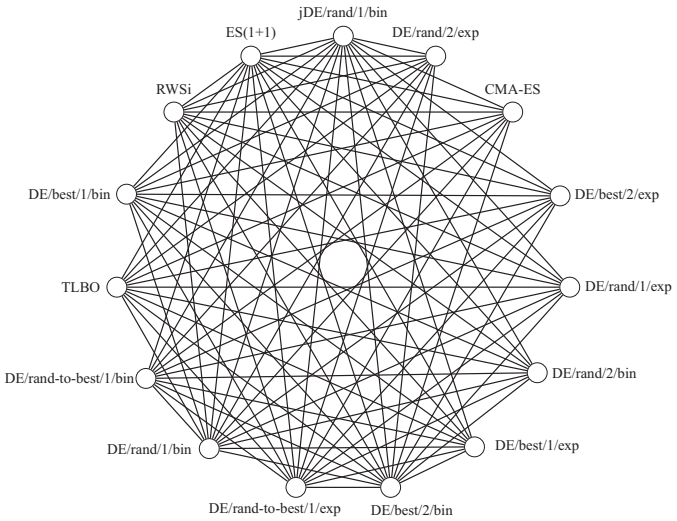
[27] J. Sonas. http://www.chessmetrics.com, Februar 2014.

[28] K. Tang, X. Li, P.N. Suganthan, Z. Yang, T. Weise. Benchmark Functions for the CEC2010 Special Session and Competition on Large-Scale Global Optimization. *Nature Inspired Computation and Applications Laboratory*, 2009.

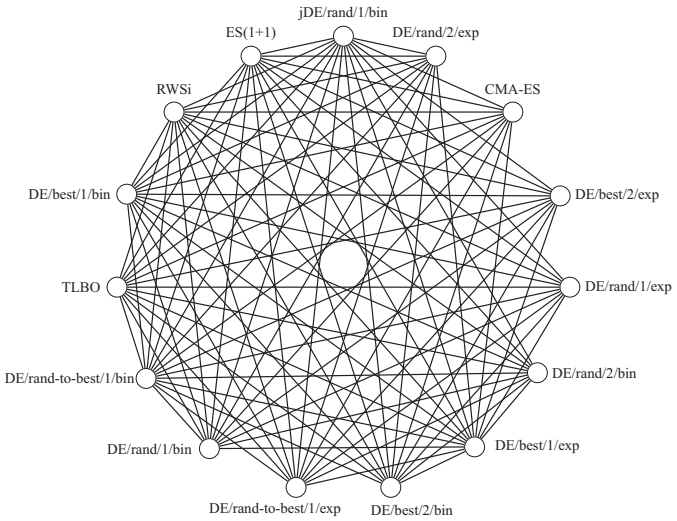
[29] J. Tvrdik. Adaptive differential evolution: application to nonlinear regression. *In Proceedings of the International Multiconference on Computer Science and Information Technology*, 193–202, 2007.

[30] N. Veček, M. Mernik, M. Črepinšek. A Chess Rating System for Evolutionary Algorithms - A New Method for the Comparison and Ranking of Evolutionary Algorithms. *Information Sciences*, 277:656-679, 2014.

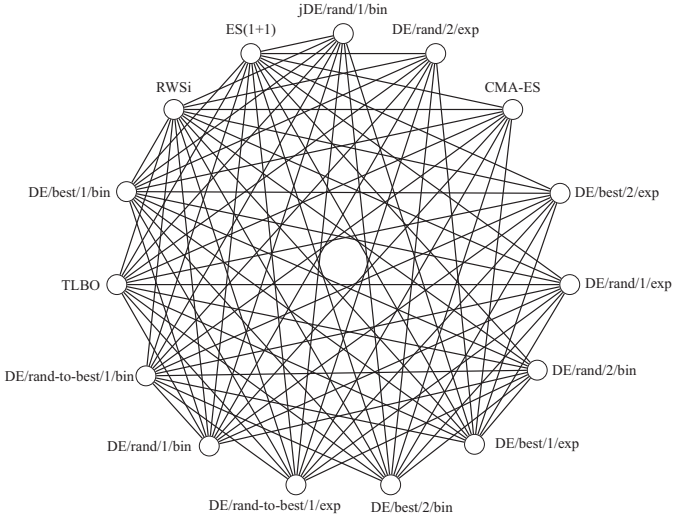
[31] D. Zaharie. A comparative analysis of crossover variants in differential evolution. *Proceedings of IMCSIT*, 171-181, 2007.



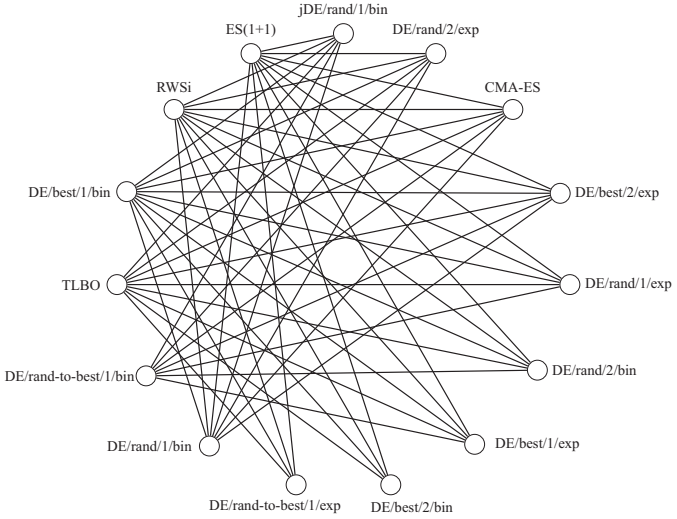
(a) Elo rating system (83 group distinctions)



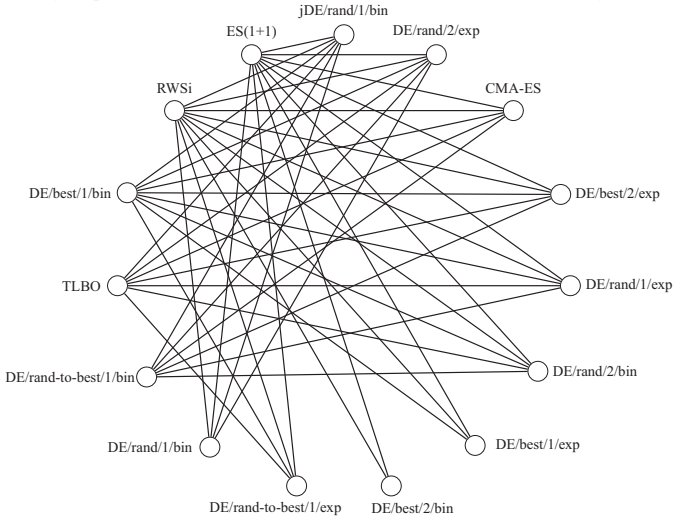
(b) Chessmetrics rating system (90 group distinctions)



(c) DWZ rating system (80 group distinctions)



(d) Glicko-2 rating system (50 significant differences)



(e) Null Hypothesis Significance Testing (43 significant differences)

Fig. 3: Detected differences amongst four rating systems. Two algorithms are connected when they are not within the same rating group (Fig. 3a, 3b, 3c) or are significantly different with probability 99.7% (Fig. 3d) or are significantly different with Null Hypothesis Significance Testing - Friedman test and $CD = 4.79$ (Fig. 3e).

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 519–527
DOI: 10.15439/2014F272
ACSIS, Vol. 2

Data-driven Genetic Algorithm in Bayesian estimation of the abrupt atmospheric contamination source

A. Wawrzynczak
1) National Centre for Nuclear Research, Świerk-Otwock, Poland
2) Institute of Computer Sciences, Siedlce Univesity, Poland
Email: a.wawrzynczak@ncbj.gov.pl

M. Jaroszynski
Institute of Computer Sciences, Siedlce Univesity, Poland
Email: marcinjaro89@gmail.com

M. Borysiewicz
National Centre for Nuclear Research, Świerk- Otwock, Poland
Email: manhaz@ncbj.gov.pl

***Abstract*—We have applied the methodology combining Bayesian inference with Genetic algorithm (GA) to the problem of the atmospheric contaminant source localization. The algorithms input data are the on-line arriving information about concentration of given substance registered by sensors' network. To achieve rapid-response event reconstructions the fast-running Gaussian plume dispersion model is adopted as the forward model. The proposed GA scan 5-dimensional parameters' space searching for the contaminant source coordinates (x,y), release strength (Q) and atmospheric transport dispersion coefficients. Based on the synthetic experiment data the GA parameters, best suitable for the contamination source localization algorithm performance were identified. We demonstrate that proposed GA configuration can successfully point out the parameters of abrupt contamination source. Results indicate the probability of a source to occur at a particular location with a particular release strength. We propose the termination criteria based on the probabilistic requirements regarding the parameters' value.**

I. Introduction

ACCIDENTAL atmospheric releases of hazardous material pose great risks to human health and the environment. In the event of an atmospheric release of chemical, but also radioactive biological materials, emergency responders need to quickly predict the current and future locations and concentrations of substance in the atmosphere. In this context it is valuable to develop the emergency system, which based on the concentration of dangerous substance by the sensors' network can inform about probable location of the release source. Moreover, the contamination source's location should be found as soon as possible. The most obvious way is to propose the simulation which gives the same substance point concentrations like registered by the sensors. However, to create the model realistically reproducing the real situation based only on the sparse point-concentration data is not trivial. This task requires specification of set of models' parameters, which depends on the applied model. The event reconstruction problem can be reformulated into a solution based on efficient sampling of an ensemble of simulations, guided by comparisons with data.

This work was supported by the Welcome Programme of the Foundation for Polish Science operated within the European Union Innovative Economy Operational Programme 2007-2013

A comprehensive literature review of past works on solutions of the inverse problem for atmospheric contaminant releases can be found in (e.g.[1]). The problem of the source term estimation was studied in literature grounded both on the deterministic and probabilistic approach. [2] implemented an algorithm based on integrating the adjoint of a linear dispersion model backward in time to solve a reconstruction problem. [3] introduced dynamic Bayesian modeling, and the Markov Chain Monte Carlo (MCMC) sampling approaches to reconstruct a contaminant source. The effectiveness of MCMC in the localization of the atmospheric contamination source based on the synthetic experiment data was presented in [4], [5]. The advantage of the Sequential Monte Carlo over the MCMC in the estimation of the probable values of the source coordinates was presented in [6].

The problem of finding the 'best fitted' model's parameters, for which a forward atmospheric dispersion model's output will reach agreement with real observations, can be considered as the optimization problem. Metaheuristics, such as genetic algorithms (GA), are broadly used to solve various optimization problems. GA was designed to imitate some of the processes that people can witness in natural environment [7]. By observing nature people noticed that many beings have evolved diametrically in the relatively short period of time. The concept of GA was to use the power of evolution to create a strong and universal tool reliable of solving optimization problems. The GAs are highly relevant for industrial applications, because they are capable of handling problems with non- linear constraints, multiple objectives, and dynamic components - properties that usually appear in the real-world problems (e.g. [8]). Since GA introduction and propagation the GA have been often used as an alternative to the conventional optimization methods and has been successfully applied in a variety of areas. For example it was used in control engineering [9], finding hardware bugs [11] and much more e.g. [10]. GA has been also used in

978-83-60810-58-3/$25.00

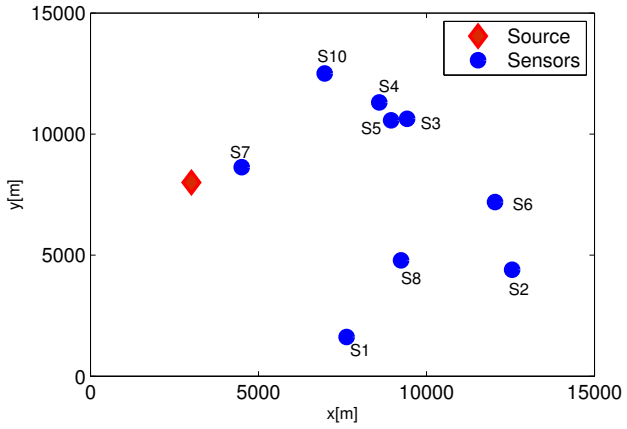


Fig. 1. Distribution of the sensors and the release source within the considered domain.

environmental sciences problem e.g. in the addressing air quality problem [12].

Application of the metaheuristic algorithms like GA requires defining the values of several algorithm components and parameters. These parameters have large impact on performance and efficiency of the algorithm (e.g., [13], [14], [15]). Therefore, it is important to estimate the algorithm's parameters best suitable for the considered optimization problem. The optimal values for the parameters depend mainly on: a) the problem; b) the domain of the problem to deal with; and c) computational time that can be spend for solving the problem. Usually, in the algorithm parameters tuning a compromise between solution quality and computational time should be achieved.

In this paper we apply the GA to the problem of localizing the abrupt atmospheric contamination source based on point-concentrations reported by the sensors network. Using the synthetic experiment data we demonstrate the efficient GA configuration best suitable for the algorithm performance.

A. Synthetic data

Our main goal is to conduct dynamic inference of an unknown atmospheric release. To test the proposed methods we require some concentration data. To satisfy this requirement we have performed the simulation with use of the atmospheric dispersion second-order Closure Integrated PUFF model (SCIPUFF) [16]. SCIPUFF is an ensemble mean dispersion model designed to compute the time-dependent field of expected concentrations resulting from one or more sources. The model solves the transport equations using a second-order closure scheme and treats releases as a collection of Gaussian puffs. In simulation we assumed that we have 10 sensors distributed over 15km x 15km area, the location of sensors was chosen randomly within the domain (Fig. 1). The atmospheric contamination source was located at $x = 3$ km, $y = 8$ km, $H = 25$ m within the domain. The simulated release was continuous with rate $Q = 8000 g/s$ and started one hour before first sensors measurements. The wind was directed along x axis with speed $5 m/s$. Further, in this paper we assume that the only algorithm input information we have, are reported every 15 minutes (in subsequent time steps) during 1.5 hour concentrations of dispersed substance registered by 10 sensors (Table I). We run algorithm searching for the source of contamination just after first information from sensors (t=1) and update the obtained probabilities with use of the developed algorithms by subsequent sensors registrations.

TABLE I
CONCENTRATION $[g/m^3]$ REPORTED BY SENSORS IN SUBSEQUENT TIME INTERVALS

Sensor	t=1	t=2	t=3	t=4	t=5	t=6
S1	0	0	0	0	0	0
S2	0	3.62E-09	4.93E-09	6.98E-09	4.15E-09	6.65E-09
S3	9.15E-09	2.88E-08	1.97E-08	1.88E-08	1.69E-08	1.62E-08
S4	3.83E-12	1.77E-11	4.89E-12	6.53E-12	2.31E-12	7.77E-12
S5	1.14E-08	1.83E-08	1.25E-08	1.20E-08	1.10E-08	1.03E-08
S6	2.91E-06	4.85E-04	4.77E-04	4.71E-04	4.43E-04	4.49E-04
S7	3.28E-05	3.27E-05	3.21E-05	3.13E-05	3.01E-05	2.87E-05
S8	2.29E-11	2.15E-10	1.05E-10	1.17E-10	7.56E-11	1.14E-10
S9	0	0	0	0	0	0
S10	0	0	0	0	0	0

II. RECONSTRUCTION PROCEDURE

A. Bayesian inference

The Bayes' theorem, as applied to an emergency release problem, can be stated as follows:

$$P(M|D) \propto P(D|M)P(M) \tag{1}$$

where M represents possible model configurations or parameters and D are observed data. For our application, Bayes' theorem describes the conditional probability $P(M|D)$ of certain source parameters (model configuration M) given observed measurements of concentration at sensor locations (D). This conditional probability $P(M|D)$ is also known as *a posteriori* distribution and is related to the probability of the data conforming to a given model configuration $P(D|M)$, and to the possible model configurations $P(M)$, before taking into account the measurements. The probability $P(D|M)$, for fixed D, is called the *likelihood* function, while $P(M)$- *a priori* distribution [17]. To estimate the unknown source parameters M using (1), the posteriori distribution $P(M|D)$ must be sampled. $P(D|M)$ quantifies the likelihood of a set of measurements D given the source parameters M.

Value of likelihood for a sample is computed by running a forward dispersion model with the given source parameters M and comparison of the model predicted concentrations in the points of sensors location (within a considered domain) with actual observations D. The closer the predicted values are to the measured ones, the higher is the likelihood of the sampled source parameters.

As the sampling procedure we use an GA to obtain the posterior distribution $P(M|D)$ of the source term parameters given the concentration measurements at sensor locations. This way we completely replace the Bayesian formulation with a

sampling procedure to explore the model parameters' space and to obtain a probability distribution for the source location.

B. The likelihood function

A measure indicating the quality of the current GA population is expressed in terms of a likelihood function. This function compares the predicted from model and observed data at the sensor locations as:

$$\lambda(M) = -\frac{\sum_{i=1}^{N}[log(C_i^M) - log(C_i^E)]^2}{2\sigma_{rel}^2}, \quad (2)$$

where λ is the likelihood function, C_i^M are the predicted by the forward model concentrations at the sensor locations i, C_i^E are the sensor measurements, N is the number of sensors, σ_{rel}^2 is an error parameter which can be updated accordingly to the expected errors in the observations at given observational time interval, here fixed to 0.2.

C. Posterior distribution

The posterior probability distribution (1) is computed directly from the resulting GA generations and is estimated as:

$$P(M|D) = \frac{1}{n}\sum_{i=1}^{n}\delta(M_i - M), \quad (3)$$

which represents the probability of a particular model configuration M giving results that match the observations at sensor locations. Equation (3) is a sum over the entire GA generation. Thus $\delta(M_i - M) = 1$ when $M_i = M$, and 0 otherwise. If in the generation many chromosomes have the same configuration $P(M|D)$ increases through the summation increasing the probability for those contamination source parameters.

D. Forward dispersion model

A forward model is needed to calculate the concentration C_i^M at the points i of sensor locations for the tested set of model parameters M at each GA step. As a testing forward model we selected the fast-running Gaussian plume dispersion model (e.g. [18]).

The Gaussian plume dispersion model for uniform steady wind conditions can be written as follows:

$$C(x,y,z) = \frac{Q}{2\pi\sigma_y\sigma_z U} exp\left[-\frac{1}{2}\left(\frac{y}{\sigma_y}\right)^2\right] \times \quad (4)$$
$$\{exp\left[-\frac{1}{2}\left(\frac{z-H}{\sigma_z}\right)^2\right] + exp\left[-\frac{1}{2}\left(\frac{z+H}{\sigma_z}\right)^2\right]\}$$

where $C(x,y,z)$ is the concentration at a particular location, U is the wind speed directed along x axis, Q is the emission rate or the source strength and H is the height of the release; y and z are the distance along horizontal and vertical direction, respectively. In the equation (4) σ_y and σ_z are the standard deviation of concentration distribution in the crosswind and vertical direction. These two parameters are defined empirically for different stability conditions [19], [20]. In this case we restrict the diffusion to the stability class C (Pasquill type stability for rural area). In scanning algorithm we assumed

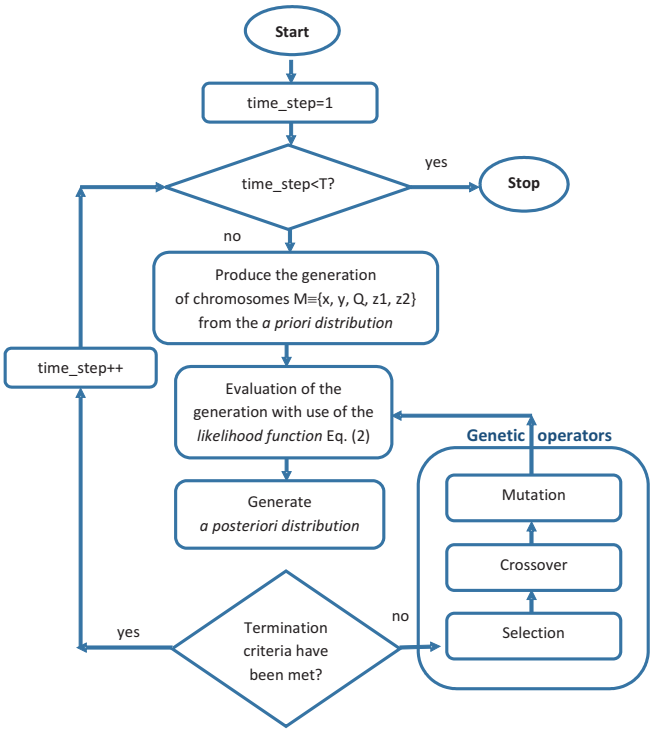


Fig. 2. Flow chart of the stochastic reconstruction procedure

that we do not know exact behavior of the plume and consider those coefficients as unknown. Thus, the parameters σ_y and σ_z are taken as: $\sigma_y = z_1 \cdot x \cdot (1 + x \cdot 4 \cdot 10^{-5})^{-0.5}, \sigma_z = z_2 \cdot x$ where values z_1 and z_2 are sampled by algorithm within interval $[0.001, 0.35]$.

To summarize, in this paper the searched model's parameter space is

$$M = (x, y, Q, z_1, z_2) \quad (5)$$

where x and y are coordinates of the release's source, Q release strength and z_1, z_2 are terms in the turbulent diffusion parametrization.

E. Genetic algorithm

The localization of the contamination source within the predefined domain requires the recognition of the atmospheric dispersion model parameters for which the model output at the sensors location meet the real data. In this context we can say that the problem can be seen as the optimization problem for which GA can be applied.

Fig. 2 presents the concept of GA's application in the Bayesian estimation of the unknown model parameters. The algorithm starts with the defining the initial population. The population is composed from the predefined number of chromosomes, $P(\tau) = x_1^\tau, \ldots, x_n^\tau$, for the generation τ, being initially randomly drawn from the admissible set of values. This set is explicitly defined by the space of explored parameters. GA chromosome is configured as binary value representing the real value of searched parameters. The quality

Algorithm 1 Rank Selection

```
ascSortMByLikelihoodFunction();
MProbabilityRange = 0;
FOR i=1 to N LOOP %N–population size
    M(i).rank = i-1; %M–chromosome
    probability = 2*(N-M(i).rank)/N*(N+1);
    MProbabilityRange += probability;
    M(i).probability = MProbabilityRange;
END LOOP
FOR i=1 to N LOOP
    randVal = drawNumberFrom0To1();
    FOR j=1 to N LOOP
        IF M(j).probability >= randVal
            newPopulation(i) = M(j);
            break;
        END IF
    END LOOP
END LOOP
```

Algorithm 2 Hard tournament selection

```
FOR i=1 to N LOOP
  FOR j=1 to TS LOOP
        tournamentGroup(j)=
                =drawSpecimenFromPopulation();
  END LOOP
  sortTournamentGroupByLikelihoodFunction();
  newPopulation(i) = getBestTournamentSpecimen();
END LOOP
```

of each chromosome in current population is evaluated based on the cost, or objective/likelihood function. Various objective functions can be applied; its form depends upon the problem being solved. We use the function presented by eq. (2). The 'improvement' of the current population can be done by the various genetic operators.

Information on the quality of population's chromosomes is used to perform selection. The portion of the population that is replaced in each generation is done based rank on the likelihood function (Eq.2) value obtained during the evaluation of the population (various in each algorithm iteration). Then, the crossover is performed. Crossover is process of replacing parents by their children in the current population. Children are created by blending of the parents at the randomly chosen crossover point. The number of crossovers that occurs within the population is determined by the crossover probability. Subsequently the current population is mutated. It changes the chromosome's features. By giving a chance of changing chromosome's individual bits mutation allows the algorithm to search for the entire solution's space and not to converge to local extremes. The number of mutations that occurs is determined by the mutation probability. After performing the selection crossover and mutation the new generation $(\tau + 1)$, being subject to the new evaluation, is established. After some number of generations the algorithm converges - it is expected that the best chromosome represents a near-optimum (reasonable) solution. The process stops when the termination criterion is fulfilled. The most common termination criterion is limited number of generations, but in this paper we present

Algorithm 3 Multi-point Crossover.

```
FOR i=1 to N LOOP %N–population size
    IF drawNumberFrom0To1() <= CP
      currentPopulation(i).isParrent(true);
    END IF
END LOOP

WHILE existsTwoNotUsedParents() LOOP
  firstParent  = popParent();
  secondParent = popParent();

  xCrossPoint = drawNumberFrom0ToParameterXLength();
  yCrossPoint = drawNumberFrom0ToParameterYLength();
  qCrossPoint = drawNumberFrom0ToParameterQLength();
  z1CrossPoint= drawNumberFrom0ToParameterZ1Length();
  z2CrossPoint= drawNumberFrom0ToParameterZ2Length();

  tmpXBin1 = firstParent.getXParameterBinaryForm();
  tmpYBin1 = firstParent.getYParameterBinaryForm();
  tmpQBin1 = firstParent.getQParameterBinaryForm();
  tmpZ1Bin1= firstParent.getZ1ParameterBinaryForm();
  tmpZ2Bin1= firstParent.getZ2ParameterBinaryForm();

  tmpXBin2 = secondParent.getXParameterBinaryForm();
  tmpYBin2 = secondParent.getYParameterBinaryForm();
  tmpQBin2 = secondParent.getQParameterBinaryForm();
  tmpZ1Bin2= secondParent.getZ1ParameterBinaryForm();
  tmpZ2Bin2= secondParent.getZ2ParameterBinaryForm();

  firstChildX  = tmpXBin1(0,CrossPoint)+
               + tmpXBin2(CrossPoint+1);
  firstChildY  = tmpYBin1(0,CrossPoint)
               + tmpYBin2(CrossPoint+1);
  firstChildQ  = tmpQBin1(0,CrossPoint)+
               + tmpQBin2(CrossPoint+1);
  firstChildZ1 = tmpZ1Bin1(0,CrossPoint)+
               + tmpZ1Bin2(CrossPoint+1);
  firstChildZ2 = tmpZ2Bin1(0,CrossPoint)+
               + tmpZ2Bin2(CrossPoint+1);

  secondChildX = tmpXBin2(0,CrossPoint)+
               + tmpXBin1(CrossPoint+1);
  secondChildY = tmpYBin2(0,CrossPoint)+
               + tmpYBin1(CrossPoint+1);
  secondChildQ = tmpQBin2(0,CrossPoint)+
               + tmpQBin1(CrossPoint+1);
  secondChildZ1= tmpZ1Bin2(0,CrossPoint)+
               + tmpZ1Bin1(CrossPoint+1);
  secondChildZ2= tmpZ2Bin2(0,CrossPoint)+
               + tmpZ2Bin1(CrossPoint+1);

  firstChild = firstChildX+firstChildY+firstChildQ
             + firstChildZ1+firstChildZ2;
  secondChild= secondChildX+secondChildY+secondChildQ
             + secondChildZ1+secondChildZ2;

  currentPopulation(firstParent.getId())=firstChild;
  currentPopulation(secondParent.getId())=secondChild;
END LOOP
```

other possibility.

In this paper the scanned parameters space M is five-dimensional i.e. $M \equiv \{x, y, Q, z1, z2\}$. Correspondingly each population's chromosome $M(i)$ stores the following information:

- x, y - coordinates of contamination's source in [m],
- Q - strength of release in [g/s],
- $z1$, $z2$ - terms in the turbulent diffusion parametrization.

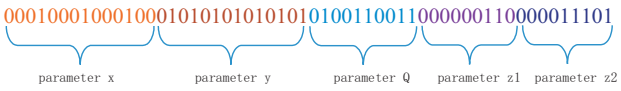


Fig. 3. Example of the chromosome representing the searched model's parameters

In the problem presented in this paper the parameters M are searched within the intervals $x \in \langle 0, 15000\rangle$, $y \in \langle 0, 15000\rangle$, $Q \in \langle 1, 8000\rangle$, $z1 \in \langle 0.001, 0.350\rangle$ and $z2 \in \langle 0.001, 0.350\rangle$. The parameters value precision P for parameters x,y equals $P_{x,y} = 1$ [m], for Q: $P_Q = 1$ [g/s], and $P_{z1} = P_{z2} = 0.001$. The example of the encoded chromosome presents Fig. 3.

1) Selection: There are many ways of dealing with GA selection e.g. roulette selection, rank selection, hard and soft tournament. For the problem presented in this paper the all mentioned methods were tested. The best results were achieved with selection based on rank and hard tournament selection. Results obtained applying these two selections are compared further in this paper. In the rank selection the better likelihood function results in the lower rank value leading to higher probability of being drawn to the next population. Pseudo code presents Algorithm 1. In the case of hard tournament selection of size 2, as the result of the tournament from each pair of the selected chromosomes one with the better objective function value passes to the next population. Pseudo code presents Algorithm 2.

2) Crossover: Similarly to the previous operator there are many ways of dealing with GA crossover e.g. single point crossover, multi point crossover, uniform crossover, arithmetic crossover. For a given problem the best results were achieved with by applying the multi-point crossover. Procedure begins with performing, for each chromosome, the test for being a parent according to the crossover probability CP. From the parents' population the unexploited pair is chosen, then one crossover point for each parameter encoded in the chromosome is drawn, i.e. five points for the problem presented. Parents are split at the crossover points for each encoded parameter, then (in term of each encoded parameter) bits are swap resulting in two children. Pseudo code presents Algorithm 3.

3) Mutation: The latter applied genetic operator is mutation. The most frequently used are uniform mutation and not-uniform mutation. For the given problem the best results were achieved with uniform mutation in which all chromosome's bits are mutated with the mutation probability MP. Pseudo code presents Algorithm 4.

In the reconstruction of the atmospheric contamination source the following GA configuration was applied:

- Size of population N=150;
- Selection:
  - rank selection,
  - hard tournament of size 2;
- Multi-point crossover with probability $CP = 0.75$, with 5 crossover points (5 is a number of searched parameters);
- Uniform mutation with probability $MP = 0.02$.

Algorithm 4 Uniform Mutation

```
FOR i=1 to N LOOP   %N–population size
  FOR j=1 to L LOOP %L–length of chromosome
                      %binary form
    IF drawNumberFrom0To1() <= MP
      currentPopulation(i).swapBitValue(j);
    END IF
  END LOOP
END LOOP
```

TABLE II
NUMBER OF GENERATIONS USED IN THE RECONSTRUCTION ALGORITHM WITH THE RANK SELECTION, $CP = 0.75$ AND $MP = 0.02$.

Time step	Generation's number	Forward dispersion model's runs
t=1	14	21 000
t=2	12	18 000
t=3	1	1 500
t=4	17	25 500
t=5	1	1 500
t=6	21	31 500
Summary	66	99 000

TABLE III
NUMBER OF GENERATIONS USED IN THE RECONSTRUCTION ALGORITHM WITH THE HARD TOURNAMENT OF SIZE 2 SELECTION, $CP = 0.75$ AND $MP = 0.02$.

Time step	Generation's number	Forward dispersion model's runs
t=1	140	210 000
t=2	124	186 000
t=3	62	93 000
t=4	97	145 500
t=5	113	169 500
t=6	216	324 000
Summary	752	1 128 000

The size of population, crossover probability and mutation probability were selected based on the numerical tests presented in [21].

III. RESULTS

We assume that the concentration from the sensors arrives subsequently in six time steps (Table I). We start to search for the source location (x, y), release rate (Q) and model parameters z_1 and z_2 after first sensors' measurements. Thus, reconstruction algorithm is run with obtaining the first measurements from the sensors ($t = 1$ at Table I). We assume that initially we have no *a priori* information about the parameters' values. So, the initial value of each parameter is draw randomly from the predefined interval with use of the uniform distribution.

Then generation is evaluated with use of the likelihood function (Eq. 2). The subsequent generations are iteratively updated by the applied genetic operators until the stop criterion is met. Of course there arises question how to specify the termination criteria? The usual criterion applied in GA is fixed number of generations. For the problem presented in this paper the time of giving the answer is crucial, so the constant number of generations is not optimal. In the task of the estimation of the source of the atmospheric contamination the most important is to

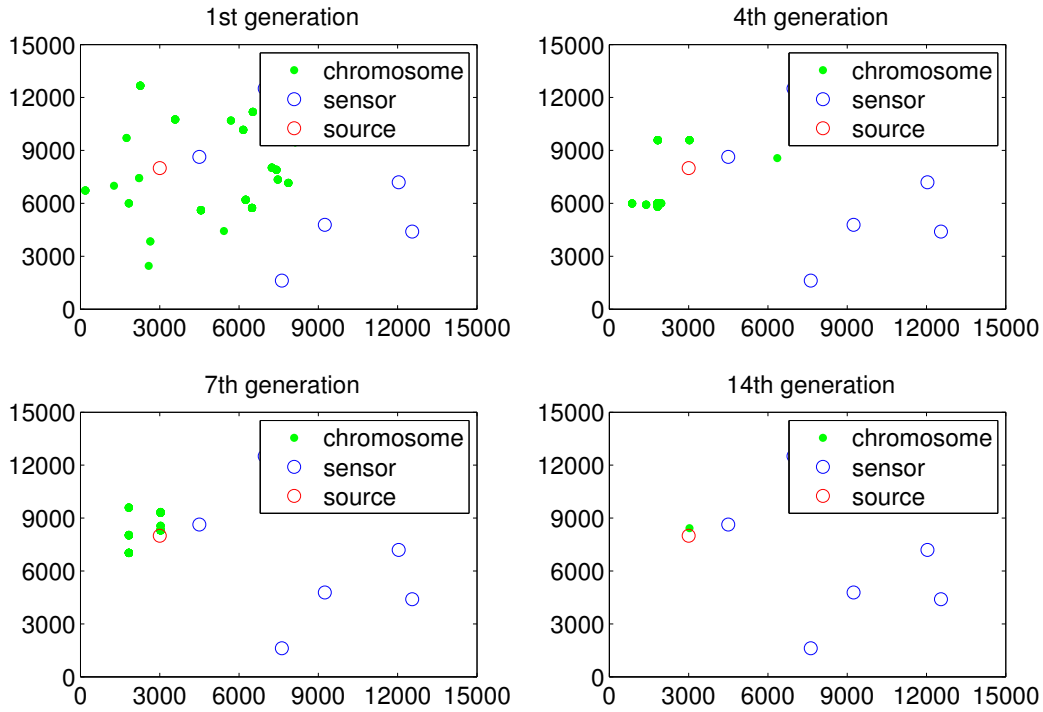


Fig. 4. Distribution of the x and y coordinates estimates during the GA runs for the given generation in 1st time step (rank selection).

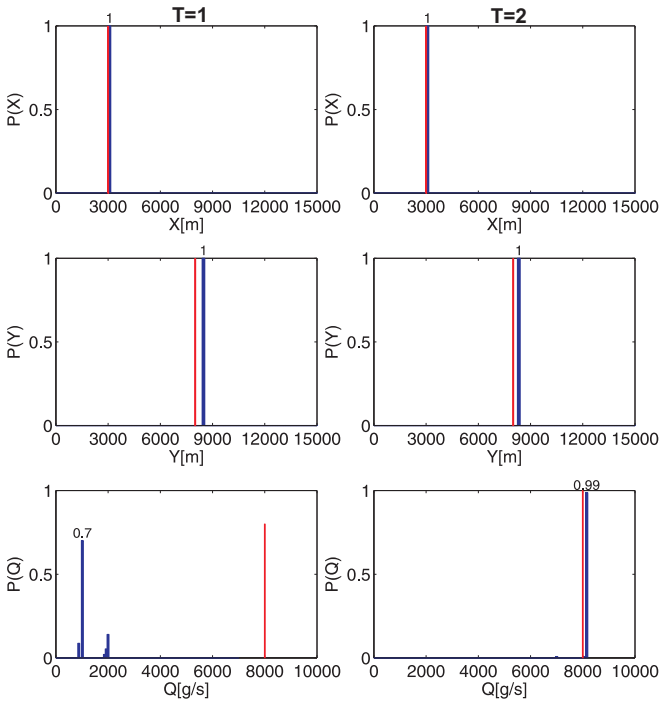


Fig. 5. Probability distributions of the models parameters x, y, and Q for the last generations in 1st and 2nd time step (rank selection). Vertical red lines represent the target value.

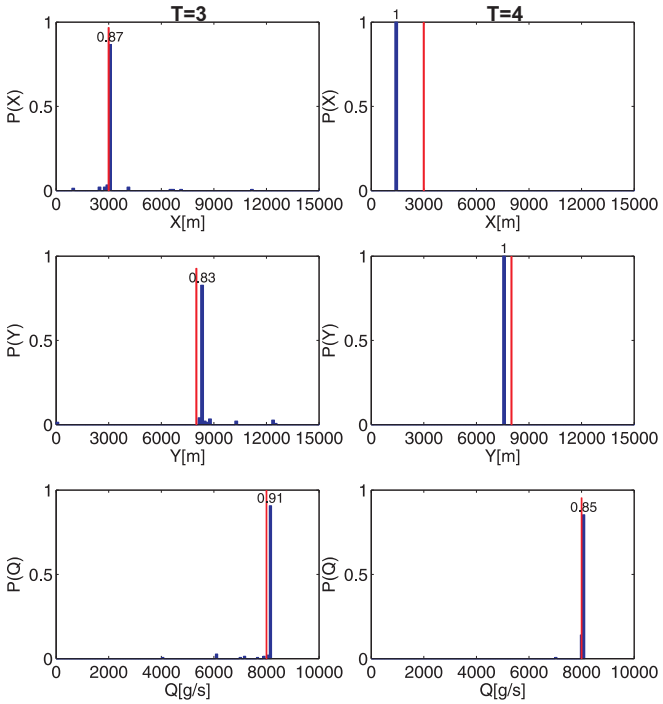


Fig. 6. Probability distributions of the models parameters x, y, and Q for the last generations in 3rd and 4th time step (rank selection). Vertical red lines represent the target value.

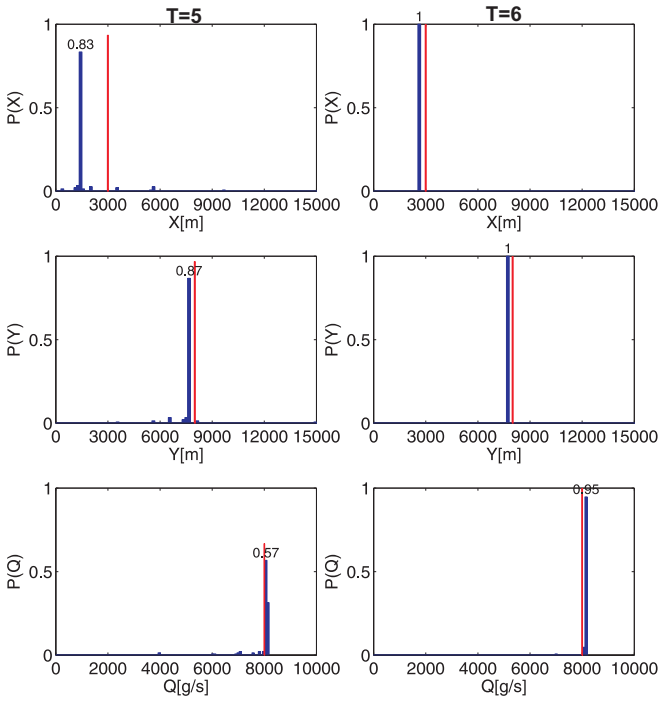


Fig. 7. Probability distributions of the models parameters x, y, and Q for the last generations in 5th and 6th time step (rank selection). Vertical red lines represent the target value.

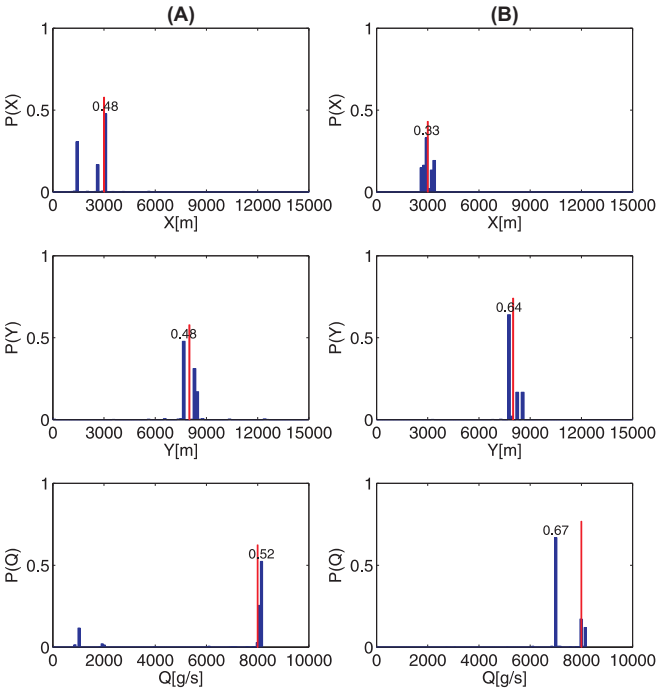


Fig. 8. Cumulative probability distributions of the models parameters x, y, and Q averaged over all time steps (A) rank selection, (B) hard tournament selection. Vertical red lines represent the target value.

estimate its location, to undertake the necessary action. Thus, crucial is assessment of the x and y coordinates of the source. Applying the Bayesian approach we can ask what probability of estimation of these parameters will be acceptable. So, after applying the last genetic operator, i.e. mutation, the histograms of x and y parameters encoded in the current chromosomes generation are evaluated. If many chromosomes have the same parameters configuration the probability of certain parameter's value increases. Consequently, the reconstruction algorithm is terminated when certain values of parameters x and y will be obtained with probability greater than 0.8. If this condition is fulfilled the *a posteriori* distributions of all parameters are calculated. Obtained *a posteriori* distributions are considered as the *a priori* distributions in the subsequent time step. Consequently, in the next time step, when new data from the sensors arrive the initial population is drawn uniformly from the *a priori* distribution i.e. *a posteriori* distribution from previous time step.

The number of generations required to fulfill the termination criterion in subsequent time steps for the rank selection is presented in Table II and for the hard tournament selection in Table III. Comparing the Tables it is obvious that the rank selection is much more effective. Fig. 4 illustrates the distribution of the estimated by the GA contamination source coordinates x and y in subsequent generations in the first time step. It is seen that at the beginning for the 1st generation the chromosomes are equally distributed within the scanned domain. However, the applied genetic operators improve population quality for further generations and the chromosomes gradually focus around the true source location. Finally, for 19th generation the estimated by the GA contamination source location approaches to the target location. Figs. 5, 6 and 7 present the a posteriori distributions for x, y and Q parameters obtained in the succeeding time steps. This distributions were obtained based on the chromosomes configurations in the last generation at given reconstruction algorithm iteration. Based on the searched parameters value, encoded in the final chromosomes population, the histogram for each parameter has been assessed. Obtained histograms shows which values of the parameters were the most frequent in the final generation, which directly is reflected in its probabilities.

Fig. 5 presents that the first sensors measurements allow to estimate the x and y parameters close to the target values, while the release strength Q is approached in the second time step. The probability distributions in subsequent time steps reflect how the sensor's data support or not the obtained distributions. The exact values of parameters differ in subsequent time steps. Below, as the estimated parameter value we provide the central value of the histogram bar with highest probability and as the error the half of the bar width. In the 6th time step the following parameters are estimated $P(x = 2625 \pm 75) = 1$, $P(y = 7725 \pm 75) = 1$ and $P(Q = 8120 \pm 40) = 0.95$. To effectively compare the results given by all proposed algorithms we have estimated the joint marginal distribution of x, y and Q parameters. Fig. 8ab present the *a posteriori* distribution averaged over all time steps for the GA algorithm

with rank selection and with the hard tournament selection, respectively. The algorithm applying rank selection as the most probable has pointed the parameters $P(x = 3075 \pm 75) = 0.48$, $P(y = 7725 \pm 75) = 0.48$ and $P(Q = 8120 \pm 40) = 0.52$, while the algorithm applying the hard tournament the parameters $P(x = 2925 \pm 75) = 0.33$, $P(y = 7725 \pm 75) = 0.64$ and $P(Q = 7000 \pm 40) = 0.67$. Fig. 9ab presents the probability distributions of the $z1$ and $z2$ parameters for the both selection methods. The algorithm applying rank selection returned the following values $P(z1 = 0.04375 \pm 0.00175) = 0.48$, $P(z2 = 0.00175 \pm 0.00175) = 0.79$ and algorithm applying the hard tournament $P(z1 = 0.05075 \pm 0.00175) = 0.48$, $P(z2 = 0.00175 \pm 0.00175) = 0.8$. We do not know the target values for these coefficient, as far the SCIPUFF model used to generate the synthetic concentration data do not allows to specify its directly. In the reconstruction procedure we could of course fix these coefficients according to the stability class pointed by the terrain and wind speed which in this case could be the stability class C for which $z1 = 0.22$ and $z2 = 0.2$. But our numerical tests showed that we obtain better results when we do not restrict the dispersion coefficients to the one given value. The 'freed' the dispersion coefficients in some acceptable interval assumption allows to better fit the Gaussian plum to the 'real' data.

Comparison of the obtained results leads to the conclusion that algorithms applying both selection methods return similar results for the x and y parameters, at the same time the algorithm using the hard tournament selection as the most probable denotes $Q = 7000$ which differs from the true release rate for $1000g/s$, while for the rank selection algorithm hits the target value. Consequently, we can pointed the algorithm applying the rank selection as more effective, as far it requires ~ 11 times less computational time than the hard tournament selection to return comparable results.

IV. Conclusion

We have presented a methodology to reconstruct a source causing an area of contamination, based on a set of measurements. The method combines Bayesian inference with the genetic algorithm and produces posterior probability distributions of the parameters describing the unknown source. Developed dynamic data-driven event reconstruction model couples data and pollutant dispersion simulations through Bayesian inference. This approach successfully provide the solution to the stated inverse problem i.e. having the downwind concentration measurement and knowledge of the wind field algorithm found the most probable location of the source and its strength.

We have proposed the termination criteria reflecting the probabilistic aspect of the obtained solution i.e. the GA is terminated when some of the searched parameters are pointed with satisfactorily probability. This approach allows to optimize the algorithm's computational time. We show that in the presented problem the rank selection is more efficient than the hard tournament selection.

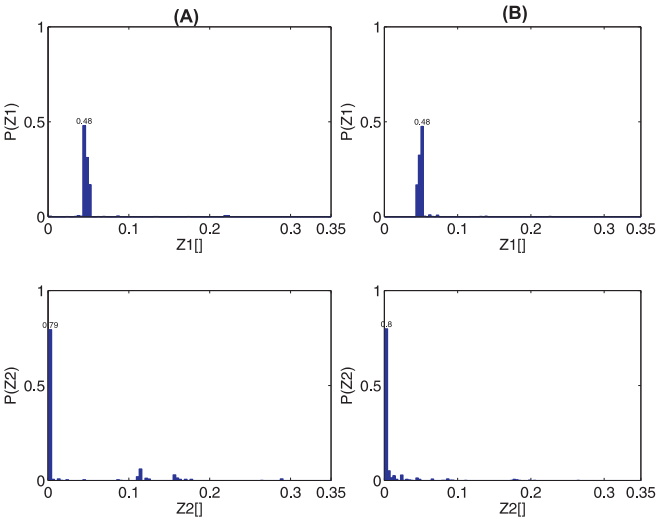


Fig. 9. Cumulative probability distributions of the models parameters $z1$ and $z2$ averaged over all time steps (A) rank selection, (B) hard tournament selection. Vertical red lines represent the value usually accepted in the Pasquili stability class C.

The probabilistic aspect of the solution optimally combines a probable answer with the uncertainties of the available data. Among several possible solutions, the Bayesian source reconstruction is solely able to find values of the model parameters that are more consistent with the currently available data.

Acknowlegements

Authors would like to thank the reviewers for helpful suggestions.

References

[1] Keats, A., E. Yee, and F.-S. Lien, (2007): Bayesian inference for source determination with applications to a complex urban environment. *Atmos. Environ.*, 41, 465-479, $doi : 10.1016/j.atmosenv.2006.08.044$.

[2] Pudykiewicz, J. A., (1998): Application of adjoint tracer transport equations for evaluating source parameters. *Atmos. Environ.*, 32, 303-3050, $doi : 10.1016/S1352 - 2310(97)00480 - 9$.

[3] Johannesson, G. et al., (2005): Sequential Monte-Carlo based framework for dynamic data-driven event reconstruction for atmospheric release.,*Proc. of the Joint Statistical Meeting*, Minneapolis, MN, American Statistical Association and Cosponsors, 73-80.

[4] Borysiewicz, M., Wawrzynczak A.,Kopka P,(2012):Stochastic algorithm for estimation of the model's unknown parameters via Bayesian inference, Proceedings of the Federated Conference on Computer Science and Information Systems pp. 501–508, IEEE Press, Wroclaw, ISBN 978-83-60810-51-4.

[5] Borysiewicz M., A.Wawrzynczak, P.Kopka,(2012): Bayesian-Based Methods for the Estimation of the Unknown Model's Parameters in the Case of the Localization of the Atmospheric Contamination Source, Foundations of Computing and Decision Sciences, 37, 4, 253–270, $doi : 10.2478/v10209 - 011 - 0014 - 9$.

[6] Wawrzynczak A., P. Kopka, M. Borysiewicz, (2014): Sequential Monte Carlo in Bayesian assessment of contaminant source localization based on the distributed sensors measurements, Lecture Notes in Computer Sciences 8385, PPAM 2013, Part II, ch.38, 407–417 , $doi : 10.1007/978 - 3 - 642 - 55195 - 6_38$.

[7] Holland J. H., (1992): Adaptation in Natural and Artificial Systems, 2nd Edn. Cambridge, MIT Press, 1992.

[8] Goldberg D. E., (2006): Genetic Algorithms in Search, Optimization and Machine Learning, Addison Wesley Longman, London, 2006.

[9] Fleming P. J., Fleming P. J., Purshouse R. C., and Purshouse R. C., (2001): Genetic Algorithms In Control Systems Engineering , In:Proceedings of the 12th IFAC World Congress, 383–390.

[10] Edited by Rustem Popa, (2012) Genetic Algorithms in Applications , ISBN 978-953-51-0400-1, InTech, Chapters published March 21, 2012 under CC BY 3.0 license doi : $10.5772/2675$

[11] Goodall R.M., Michail K., Whidborne J.F. Zolotas A.C, (2009): Optimised Configuration of Sensing Elements for Control and Fault Tolerance Applied to an Electro-Magnetic Suspension, PhD Thesis, Loughborough University, UK.

[12] Allen C.T, Haupt S. E., (2006): Source Characterization with a Genetic Algorithm-Coupled Dispersion-Backward Model Incorporating SCIPUFF, Department of Meteorology, The Pennsylvania State University, doi : $10.1175/JAM2459.1$.

[13] Eiben A . E., R. Hinterding and Z. Michalewicz, (1999): Parameter Control in Evolutionary Algorithms", IEEE Transactions on Evolutionary Computation, Vol. 3, No. 2, doi : $10.1109/4235.771166$

[14] Saremi A., T. Y. E. Mekkawy and G. G. Wang, (2007) Tuning the Parameters of a Memetic Algorithm to Solve Vehicle Routing Problem with Backhauls Using Design of Experiments", International Journal of Operations Research, Vol. 4, No. 4, 206–219.

[15] Roeva O., Stefka Fidanova, Marcin Paprzycki , Influence of the Population Size on the Genetic Algorithm Performance in Case of Cultivation Process Modelling , Proceedings of the 2013 Federated Conference on Computer Science and Information Systems, pages 371 - 376, 2013

[16] Sykes, R.I. et.al., (1998): PC-SCIPUFF Version 1.2PD Technical Documentation. ARAP Report No. 718. Titan Corporation,

[17] Gelman, A., J. Carlin, H. Stern, and D. Rubin, (2003): Bayesian Data Analysis. *Chapman & Hall/CRC*, 668 pp.

[18] Turner D. Bruce, (1994): Workbook of Atmospheric Dispersion Estimates, *Lewis Publishers*, USA

[19] Pasquill, F. (1961): The estimate of the dispersion of windborne material, *Meteorol Mag.*,90, 1063,: 33-49

[20] Gifford, F. A. Jr. (1960): Atmospheric dispersion calculation using generalized Gaussian Plum model, *Nuclear Safety*, 2(2):56-59,67-68

[21] Wawrzynczak A et al. (2014): Recognition of the atmospheric contamination source localization with the Genetic Algorithm, Studia Informatica, UPH, Siedlce (*submitted*)

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 529–538
DOI: 10.15439/2014F142
ACSIS, Vol. 2

Dispersive Flies Optimisation

Mohammad Majid al-Rifaie
Department of Computing
Goldsmiths University of London
London SE14 6NW, United Kingdom
Email: m.majid@gold.ac.uk

***Abstract*—One of the main sources of inspiration for techniques applicable to complex search space and optimisation problems is nature. This paper proposes a new metaheuristic – Dispersive Flies Optimisation or DFO – whose inspiration is beckoned from the swarming behaviour of flies over food sources in nature. The simplicity of the algorithm, which is the implementation of one such paradigm for continuous optimisation, facilitates the analysis of its behaviour. A series of experimental trials confirms the promising performance of the optimiser over a set of benchmarks, as well as its competitiveness when compared against three other well-known population based algorithms (Particle Swarm Optimisation, Differential Evolution algorithm and Genetic Algorithm). The convergence-independent diversity of DFO algorithm makes it a potentially suitable candidate for dynamically changing environment. In addition to diversity, the performance of the newly introduced algorithm is investigated using the three performance measures of accuracy, efficiency and reliability and its outperformance is demonstrated in the paper.**

I. Introduction

THROUGHOUT the history nature has been an inexplicable source of inspiration for scientists and researchers. Observations, many of which made unintentionally, have been triggering the inquisitive minds for hundreds of years. The task of resolving problems and its often present nature in the minds of scientists boosts the impact of these observations, which in cases led to discoveries. Among others, researchers in mathematics, physics and natural sciences have had their fair share of 'observations-leading-to-discoveries'.

Observing the magnificently choreographed movements of birds, behaviour of ants foraging, convergence of honey bees in search for food source and so forth has led several researchers to propose (inspired vs. identical) models used to solve various optimisation problems. Genetic Algorithm [1], Particle Swarm Optimisation [2] and Ant Colony Optimisation [3] are only few such techniques belonging to the broader category of swarm intelligence; it investigates collective intelligence and aims at modelling intelligence by looking at individuals in a social context and monitoring their interactions with one another as well as their interactions with the environment.

The work presented here aims at proposing a novel nature-inspired algorithm based on the behaviours of flies hovering over food sources. This model – Dispersive Flies Optimisation or DFO – is first formulated mathematically and then a set of experiments is conducted to examine its performance when presented with various problems.

II. Flies in Nature

Flies are insects of the order *Diptera*, which comprises a large order, containing an estimated $240,000$ species of mosquitoes, gnats, midges and others [4]. Flies exist in various types each exhibiting distinctive behaviour in different environments. What most flies have in common is their swarming behaviour which depends on several factors.

Swarming have been described in [5] where a difference of shape between low swarms over dung and high swarms over other markers have been logged. High swarms fluctuated in height; vertical movements of the swarms of *Anopheles franciscanus* (Culicidae) are said to be correlated with female presence at swarms [6]. Height change in mosquito swarm induced by a clarinet note [7] and the human voice [8] may have evolved as responses to the flight tone of female mosquitoes [9].

Swarms of flies are associated with visual markers ranging in size from cowpies and stones to church steeples [10]. The criteria used by insects to select markers may be quite subtle; it was noted in [11] that certain objects are used repeatedly by the mosquito *Aedes cataphylla* while similar objects nearby are neglected.

As explained in [12], various swarms of flies usually "flying in relation to a more or less conspicuous element of the landscape, a lakeshore, a road, a treetop, below the tip of a branch, in an opening in the forest canopy, above a cow, an outstanding leaf", and so on according to species (e.g. [13], [14]). Depending on the species, the size of the swarm may consist of a single individual or tens or thousands, related to a discrete swarm marker; or even countless millions in the zonal swarms of lake shores.

Several elements play a role in *disturbing* the swarms of flies; for instance, the presence of a threat causes the swarms to disperse, leaving their current marker; they return to the marker immediately after the threat is over. However, during this period if they discover another marker which matches their criteria closer, they adopt the new marker.

III. Dispersive Flies Optimisation

Dispersive Flies Optimisation (DFO) is an algorithm inspired by the swarming behaviour of flies hovering over food sources. As detailed in section II, the swarming behaviour of flies is determined by several factors and that the presence of threat could disturb their convergence on the marker (or the

978-83-60810-58-3/$25.00

optimum value). Therefore, having considered the formation of the swarms over the marker, the breaking or weakening of the swarms is also noted in the proposed algorithm.

In other words, the swarming behaviour of the flies, in Dispersive Flies Optimisation, consist of two tightly connected mechanisms, one is the formation of the swarms and the other is its breaking or weakening. The algorithm and the mathematical formulation of the update equations are introduced below.

The position vectors of the population are defined as:

$$\vec{x}_i^t = \left[x_{i1}^t, x_{i2}^t, ..., x_{iD}^t\right], \qquad i = 1, 2, ..., NP \tag{1}$$

where t is the current time step, D is the dimension of the problem space and NP is the number of flies (population size).

In the first generation, when $t = 0$, the i^{th} vector's j^{th} component is initialised as:

$$x_{id}^0 = x_{min,d} + r\left(x_{max,d} - x_{min,d}\right) \tag{2}$$

where r is a random number drawn from a uniform distribution on the unit interval $U(0,1)$; x_{min} and x_{max} are the lower and upper initialisation bounds of the d^{th} dimension, respectively. Therefore, a population of flies are randomly initialised with a position for each flies in the search space.

On each iteration, the components of the position vectors are independently updated, taking into account the component's value, the corresponding value of the best neighbouring fly (consider ring topology) with the best fitness, and the value of the best fly in the whole swarm:

$$x_{id}^t = x_{nb,d}^{t-1} + \mathrm{U}(0,1) \times \left(x_{sb,d}^{t-1} - x_{id}^{t-1}\right) \tag{3}$$

where $x_{nb,d}^{t-1}$ is the value of the *n*eighbour's *b*est fly in the d^{th} dimension at time step $t-1$; $x_{sb,d}^{t-1}$ is the value of the *s*warm's *b*est fly in the d^{th} dimension at time step $t-1$; and $\mathrm{U}(0,1)$ is the uniform distribution between 0 and 1.

The algorithm is characterised by two principle components: a dynamic rule for updating flies position (assisted by a social neighbouring network that informs this update), and communication of the results of the best found fly to other flies.

As stated earlier, the swarm is disturbed for various reasons; one of the positive impacts of such disturbances is the displacement of the disturbed flies which may lead to discovering a better position. To consider this eventuality, an element of stochasticity is introduced to the update process. Based on this, individual components of flies' position vectors are reset if the random number, r, generated from a uniform distribution on the unit interval $U(0,1)$ is less than the *disturbance threshold* or *dt*. This guarantees a proportionate disturbance to the otherwise permanent stagnation over a likely local minima.

Algorithm 1 summarises the DFO algorithm[1].

The next section briefly presents three population-based algorithms which will be used to compare the performance of DFO, and then the results of a series of experiments conducted on DFO over a set of benchmark functions are reported.

[1]The source code can be downloaded from the following page: http://doc.gold.ac.uk/~map01mm/DFO/

Algorithm 1 Dispersive Flies Optimisation

while FE $< 300,000$ **do**
  for $i = 1 \rightarrow NP$ **do**
    $\vec{x}_i$.fitness $\leftarrow f(\vec{x}_i)$
  end for
  $sb \leftarrow \{sb,\ \forall\ f(\vec{x}_{sb}) = \min\left(f(\vec{x}_1), f(\vec{x}_2), ..., f(\vec{x}_{NP})\right)\}$
  $nb \leftarrow \{nb,\ \forall\ f(\vec{x}_{nb}) = \min\left(f(\vec{x}_{\text{left}}), f(\vec{x}_{\text{right}})\right)\}$
  for $i = 1 \rightarrow NP$ **do**
    for $d = 1 \rightarrow \mathrm{D}$ **do**
      $\tau_d \leftarrow x_{nb,d}^{t-1} + \mathrm{U}(0,1) \times \left(x_{sb,d}^{t-1} - x_{id}^{t-1}\right)$
      if $(r < dt)$ **then**
        $\tau_d \leftarrow x_{min,d} + r\left(x_{max,d} - x_{min,d}\right)$
      end if
    end for
    $\vec{x}_i \leftarrow \vec{\tau}$
  end for
end while

IV. Population-Based Algorithms

The three algorithms introduced briefly in this section are variations of particle swarm optimisation (PSO), differential evolution algorithm (DE) and genetic algorithm (GA). One of the common features of these algorithms are the interactions between their population (i.e. information sharing), with the ultimate goal of finding the optima.

A. Particle Swarm Optimisation

Particle swarm optimisation (PSO) is population based optimization technique developed in 1995 by Kennedy and Eberhart [2]. It came about as a result of an attempt to graphically simulate the choreography of fish schooling or birds flying (e.g. pigeons, starlings, and shorebirds) in coordinated flocks that show strong synchronisation in turning, initiation of flights and landing, despite the fact that experimental researches to find leaders in such flocks failed [15].

A swarm in PSO algorithm comprises of a number of particles and each particle represents a point in a multi-dimensional problem space. The position of each particle, $\vec{x}$, is thus dependent on the particle's own experience and those of its neighbours. Each particle has a memory, containing the best position found so far during the course of the optimisation, which is called personal best or $\vec{p}$. Whereas the best position so far found throughout the population, or the local neighbourhood, is called neighbourhood best.

A standard particle swarm version, Clerc-Kennedy PSO (PSO-CK) or constriction PSO defines the position of each particle by adding a velocity to the current position. Here is the equation for updating the velocity and position of each particle:

$$v_{id}^t = \chi\left(v_{id}^{t-1} + c_1 r_1\left(p_{id} - x_{id}^{t-1}\right) + c_2 r_2\left(g_{id} - x_{id}^{t-1}\right)\right) \tag{4}$$

$$x_{id}^t = v_{id}^t + x_{id}^{t-1} \tag{5}$$

where χ which is the constriction factor is set to 0.72984 which is reported to be be working well in general [16]; v_{id}^{t-1} is the velocity of particle i in dimension d at time step $t-1$; $c_{1,2}$ are the learning factors (also referred to as acceleration constants) for personal best and neighbourhood best respectively (they are constant); $r_{1,2}$ are random numbers adding stochasticity to the algorithm and they are drawn from a uniform distribution on the unit interval $U(0,1)$; p_{id} is the personal best position of particle x_i in dimension d; and g_{id} is neighbourhood best. In the experiments reported in this work, local neighbourhood is used.

B. Differential Evolution Algorithm

Differential evolution (DE), an evolutionary algorithms (EAs), is a simple global numerical optimiser over continuous search spaces which was first introduced by Storn and Price [17].

DE is a population based stochastic algorithm, proposed to search for an optimum value in the feasible solution space. The parameter vectors of the population are defined as follows:

$$\vec{x}_i^g = \left[x_{i,1}^g, x_{i,2}^g, ..., x_{i,D}^g\right], i = 1, 2, ..., NP \tag{6}$$

where g is the current generation, D is the dimension of the problem space and NP is the population size. In the first generation, (when $g = 0$), the i^{th} vector's j^{th} component could be initialised as:

$$x_{i,j}^0 = x_{min,d} + r\left(x_{max,d} - x_{min,d}\right) \tag{7}$$

where r is a random number drawn from a uniform distribution on the unit interval $U(0,1)$, and x_{min}, x_{max} are the lower and upper bounds of the d^{th} dimension, respectively. The evolutionary process (mutation, crossover and selection) starts after the initialisation of the population.

1) Mutation: At each generation g, the mutation operation is applied to each member of the population x_i^g (target vector) resulting in the corresponding vector v_i^g (mutant vector). In this work, *DE/best/1* variation of mutation approaches is used:

$$v_i^g = x_{best}^g + F\left(x_{r_1}^g - x_{r_2}^g\right) \tag{8}$$

where r_1 and r_2 are different from i and are distinct random integers drawn from the range $[1, NP]$; In generation g, the vector with the best fitness value is x_{best}^g; and F is a positive control parameter for constricting the difference vectors and is set to 0.5.

2) Crossover: Crossover operation, improves population diversity through exchanging some components of v_i^g (mutant vector) with x_i^g (target vector) to generate u_i^g (trial vector). This process is led as follows:

$$u_{i,j}^g = \begin{cases} v_{i,j}^g, & \text{if } r \leq CR \text{ or } j = r_d \\ x_{i,j}^g, & \text{otherwise} \end{cases} \tag{9}$$

where r is a uniformly distributed random number drawn from the unit interval $U(0,1)$, r_d is randomly generated integer from the range $[1, D]$; this value guarantees that at least one component of the trial vector is different from the target vector. The value of CR, which is another control parameter and is set to 0.5, specifies the level of inheritance from v_i^g (mutant vector).

3) Selection: The selection operation decides whether x_i^g (target vector) or u_i^g (trial vector) would be able to to pass to the next generation $(g+1)$.In case of a minimisation problem, the vector with a smaller fitness value is admitted to the next generation:

$$x_i^{g+1} = \begin{cases} u_i^g, & \text{if } f\left(u_i^g\right) \leq f\left(x_i^g\right) \\ x_i^g, & \text{otherwise} \end{cases} \tag{10}$$

where $f(x)$ is the fitness function.

C. Genetic Algorithm

In this work, we use a real-valued Genetic Algorithm (GA) which has previously shown to work well on real-world problems [18], [19]. The GA works in the following way: the individuals are first randomly initialised and their fitness is evaluated through an objective function. Afterwards, in a iterative process, each individual has a probability of being exposed to recombination or mutation (or both). These probabilities are p_c and p_m respectively. The recombination operator used is arithmetic crossover and the mutation operator used is Cauchy mutation using an annealing scheme. At the end, in order to comb out the least fit individual, tournament selection [20] is utilised.

The reason behind using Cauchy mutation operator vs. the well-known Gaussian mutation operator is the thick tails of the Cauchy distribution that allows it to generate considerable changes, more frequently, compared to the Gaussian distribution. The Cauchy distribution is defined by:

$$C(x, \alpha, \beta) = \frac{1}{\beta\pi\left(1 + \left(\frac{x-\alpha}{\beta}\right)^2\right)} \tag{11}$$

where $\alpha \leq 0$, $\beta > 0$, $-\infty < x < \infty$ (α and β are parameters that affect the mean and spread of the distribution). As specified in [19], all of the solution parameters are subject to mutation and the variance is scaled with $0.1\times$ the range of the specific parameter in question.

In order to decrease the value of β as a function of the elapsed number of generations t, an annealing scheme was applied (α was set to 0):

$$\beta(t) = \frac{1}{1+t} \tag{12}$$

As for the arithmetic crossover, the offspring is generated as a weighted mean of each gene of the two parents:

$$\text{offspring}_i = r \times \text{parent1}_i + (1-r) \times \text{parent2}_i \tag{13}$$

where offspring_i is the i'th gene of the offspring, and parent1_i and parent2_i refer to the i'th gene of the two parents, respectively. The weight r is drawn from a uniform distribution on the unit interval $U(0,1)$.

In the experiments conducted in this paper, the probabilities of crossover and mutation of the individuals is set to $p_c = 0.7$ and $p_m = 0.9$ respectively. The tournament size of the tournament selection is set to two, and elitism with an elite size of one is deployed to maintain the best found solution in the population.

V. Experiments

This section presents a set of experiment investigating the performance of the newly introduced Dispersive Flies Optimisation (DFO) and discusses the results. Then, to understand whether disturbance plays an important role in the optimisation process, a *control* algorithm is presented DFO-c where no disturbance is inflicted upon the population of flies.

Recognising the lose of diversity as a common issue in all distribution based evolutionary optimisers (since dispersion reduces with convergence), the impact of disturbance on preserving the diversity of the population is also studied. Additionally, an optimal value for disturbance threshold, *dt*, is suggested. Afterwards the performance of DFO is compared against few other well-known population-based algorithms, namely Particle Swarm Optimisation (PSO), Differential Evolution (DE) and Genetic Algorithm (GA).

A. Experiment Setup

The benchmarks used in the experiments (see Table I) are divided in two sets, f_{1-14} and g_{1-14}; more details about these functions (e.g. global optima, mathematical formulas, etc.) are reported in [16] and [21]. The first set, f_{1-14}, have been used by several authors [22], [16], [23] and it contains the three classes of functions recommended by Yao *et al.* [24]: unimodal and high dimensional, multimodal and high dimensional, and low dimensional functions with few local minima. In order not to initialise the flies on or near a region in the search space known to have the global optimum, *region scaling* technique is used [25], which makes sure the flies are initialised at a corner of the search space where there are no optimal solutions.

The second test set, g_{1-14}, are the first fourteen functions of CEC 2005 test suite [21] and they present more challenging features of the common functions from the aforementioned test set (e.g. shifted by an arbitrary amount within the search space and/or rotated). This set has also been used for many researchers.

One hundred flies were used in the experiments and the termination criterion for the experiments is set to reaching $300,000$ function evaluations (FEs). There are 50 Monte Carlo simulations for each experiment and the results are averaged over these independent simulations. Apart from the disturbance threshold which is set to $dt = 0.001$, there are no adjustable parameters in DFO's update equation.

The aim of the experiments is to study and demonstrate the qualities of the newly introduced algorithm as a population based continuous optimiser. The behaviour of the DFO algorithm is compared against its control counterpart and some other population based algorithms (see Sections IV-A, IV-B and IV-C).

B. Performance measures and statistical analysis

In order to conduct the statistical analysis measuring the presence of any significant difference in the performance of the algorithms, Wilcoxon 1×1 non-parametric statistical test is deployed. The performance measures used in this paper are error, efficiency, reliability and diversity which are described below.

Error is defined by the quality of the best agent in terms of its closeness to the optimum position (if knowledge about the optimum position is known *a priori*, which is the case here). Another measure used is *efficiency* which is the number of function evaluations before reaching a specified error, and *reliability* is the percentage of trials where a specified error is reached. These performance measures are defined as below:

$$\text{Error} = |f(\vec{x}_g) - f(\vec{x}_o)| \quad (14)$$

$$\text{Efficiency} = \frac{1}{n}\sum_{i=1}^{n} \text{FEs} \quad (15)$$

$$\text{Reliability} = \frac{n^{'}}{n} \times 100 \quad (16)$$

where $\vec{x}_g$ is the best position found and $\vec{x}_o$ is the position of the known optimum solution; *n* is the number of trials in the experiment and $n^{'}$ is the number of successful trials, FEs is the number of function evaluations before reaching the specified error, which in these experiments, set to 10^{-8}.

In this work, *diversity*, which is the degree of convergence and divergence, is defined as a measure to study the population's behaviour with regard to exploration and exploitation. There are various approaches to measure diversity. The average distance around the population centre is shown [26] to be a robust measure in the presence of outliers and is defined as:

$$\text{Diversity} = \frac{1}{NP}\sum_{i=1}^{NP}\sqrt{\sum_{j=1}^{D}\left(x_i^j - \bar{x}^j\right)^2} \quad (17)$$

$$\bar{x}^j = \frac{1}{NP}\sum_{i=1}^{NP} x_i^j \quad (18)$$

where NP is the number of flies in the population, D is the dimensionality of the problem, x_i^j is the value of dimension j of agent i, and $\bar{x}^j$ is the average value of dimension j over all agents.

C. Performance of Dispersive Flies Optimisation

The error, efficiency and reliability results of DFO performance over the benchmarks are reported in Table II. The first five columns detail the error-related figures and the last column highlights the median efficiency along with the reliability (shown between brackets) of the algorithm in finding the optima. The algorithm exhibits a promising performance in optimising the presented problem set where half the benchmarks ($f_{1-2,5-11}$ and $g_{1-2,7,9}$) are optimised with the specified accuracy. The figures in the table are expanded in the following categories:

TABLE I
BENCHMARK FUNCTIONS

Fn	Name	Class	Dimension	Feasible Bounds
f_1	Sphere/Parabola	Unimodal	30	$(-100, 100)^D$
f_2	Schwefel 1.2	Unimodal	30	$(-100, 100)^D$
f_3	Generalized Rosenbrock	Multimodal	30	$(-30, 30)^D$
f_4	Generalized Schwefel 2.6	Multimodal	30	$(-500, 500)^D$
f_5	Generalized Rastrigin	Multimodal	30	$(-5.12, 5.12)^D$
f_6	Ackley	Multimodal	30	$(-32, 32)^D$
f_7	Generalized Griewank	Multimodal	30	$(-600, 600)^D$
f_8	Penalized Function P8	Multimodal	30	$(-50, 50)^D$
f_9	Penalized Function P16	Multimodal	30	$(-50, 50)^D$
f_{10}	Six-hump Camel-back	Low Dimensioal	2	$(-5, 5)^D$
f_{11}	Goldstein-Price	Low Dimensioal	2	$(-2, 2)^D$
f_{12}	Shekel 5	Low Dimensioal	4	$(0, 10)^D$
f_{13}	Shekel 7	Low Dimensioal	4	$(0, 10)^D$
f_{14}	Shekel 10	Low Dimensioal	4	$(0, 10)^D$
g_1	Shifted Sphere	Unimodal	30	$(-100, 100)^D$
g_2	Shifted Schwefel 1.2	Unimodal	30	$(-100, 100)^D$
g_3	Shifted Rotated High Conditioned Elliptic	Unimodal	30	$(-100, 100)^D$
g_4	Shifted Schwefel 1.2 with Noise in Fitness	Unimodal	30	$(-100, 100)^D$
g_5	Schwefel 2.6 with Global Optimum on Bounds	Unimodal	30	$(-100, 100)^D$
g_6	Shifted Rosenbrock	Multimodal	30	$(-100, 100)^D$
g_7	Shifted Rotated Griewank without Bounds	Multimodal	30	$(-600, 600)^D$
g_8	Shifted Rotated Ackley with Global Optimum on Bounds	Multimodal	30	$(-32, 32)^D$
g_9	Shifted Rastrigin	Multimodal	30	$(-5, 5)^D$
g_{10}	Shifted Rotated Rastrigin	Multimodal	30	$(-5, 5)^D$
g_{11}	Shifted Rotated Weierstrass	Multimodal	30	$(-0.5, 0.5)^D$
g_{12}	Schwefel Problem 2.13	Multimodal	30	$(-\pi, \pi)^D$
g_{13}	Expanded Extended Griewank plus Rosenbrock	Expanded	30	$(-5, 5)^D$
g_{14}	Shifted Rotated Expanded Scaffer	Expanded	30	$(-100, 100)^D$

1) Unimodal, high dimensional ($f_{1,2}, g_{1-5}$): The algorithm optimises 57% of the benchmarks in this category; while both functions in the first set are optimised ($f_{1,2}$), only two out of five benchmarks in the second and more challenging set are optimised to the specified accuracy. All optimised benchmarks achieve 100% success.

2) Low dimensional and few local minima (f_{10-14}): In this category, 40% of the benchmarks are optimised, with 100% reliability for f_{10} and 32% for f_{11}. However, none of the Shekel functions (f_{12-14}) are optimised; Shekel is known to be a challenging function to optimise due to the presence of several broad sub-optimal minima; also the proximity of a small number of optima to the Shekel parameter $\vec{a}_i$ is another reason for the difficulty of optimising these set of functions.

3) Multimodal, high dimensional (f_{3-9}, g_{6-14}): The optimiser is able to optimise 50% of the benchmarks in this category (f_{5-9} and $g_{7,9}$), 71% of which achieve 100% success rate (all except f_7, g_7 with 28% and 10% success rates respectively). The optimiser exhibit a promising performance when dealing with the difficult Rosenbrock functions (f_3, g_6), reaching the error of 10^{-4} and 10^{-3} respectively. The algorithm performs exceptionally well in optimising the infamous Rastrigin functions, both common and shifted mode (i.e. f_5 and g_9), achieving 100% success rate; however it does show weakness in the more challenging g_{10} rotated version.

The success of the optimiser in optimising the notorious Rastrigin function in its common and shifted modes will be discussed in the context of DFO's dimension-to-dimension disturbance mechanism induced by the algorithm.

In order to provide a better understanding of the behaviour of the algorithm, in the next section, the disturbance is discarded and the diversity of the algorithm is studied.

D. Diversity in DFO

Most swarm intelligence and evolutionary techniques commence with exploration and, over time (i.e. function evaluations or iterations), lean towards exploitation. Maintaining the right balance between exploration and exploitation phases has proved to be difficult. The absence of the aforementioned balance leads to a weaker diversity when encountering a local minimum and thus the common problem of pre-mature convergence to a local minimum surfaces.

Similar to other swarm intelligence and evolutionary algorithms, DFO commences with exploration and over time, through its mechanism (i.e. gradual decrease in the distance between the members of the population and as such,

TABLE II
DFO – DISPERSIVE FLIES OPTIMISATION

	Min.	Max.	Median	Mean	StdDev	Eff. (Rel.)
f_1	6.46E-47	1.97E-40	1.75E-43	1.07E-41	3.49E-41	46850 (100%)
f_2	2.24E-12	6.01E-10	6.46E-11	1.08E-10	1.26E-10	239850 (100%)
f_3	1.74E-04	1.45E+01	3.65E-01	2.17E+00	3.62E+00	∞ (0%)
f_4	3.89E-07	5.05E-03	2.87E-05	2.49E-04	7.81E-04	∞ (0%)
f_5	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	84850 (100%)
f_6	2.84E-14	6.39E-14	3.91E-14	3.88E-14	6.49E-15	121200 (100%)
f_7	0.00E+00	1.54E-01	1.85E-02	3.25E-02	3.74E-02	47450 (28%)
f_8	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	50950 (100%)
f_9	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	55550 (100%)
f_{10}	0.00E+00	2.22E-16	0.00E+00	4.00E-17	8.62E-17	1700 (100%)
f_{11}	0.00E+00	8.10E+01	8.10E+01	5.51E+01	3.82E+01	2100 (32%)
f_{12}	5.05E+00	5.05E+00	5.05E+00	5.05E+00	0.00E+00	∞ (0%)
f_{13}	5.27E+00	5.27E+00	5.27E+00	5.27E+00	0.00E+00	∞ (0%)
f_{14}	5.36E+00	5.36E+00	5.36E+00	5.36E+00	0.00E+00	∞ (0%)
g_1	5.68E-14	2.27E-13	1.71E-13	1.49E-13	4.28E-14	45300 (100%)
g_2	4.55E-12	9.78E-10	3.88E-11	1.03E-10	1.57E-10	234100 (100%)
g_3	3.58E+05	3.22E+06	1.40E+06	1.38E+06	6.23E+05	∞ (0%)
g_4	1.40E+00	2.38E+02	2.18E+01	3.71E+01	4.74E+01	∞ (0%)
g_5	3.47E+03	1.82E+04	8.95E+03	9.26E+03	3.17E+03	∞ (0%)
g_6	1.66E-03	1.51E+02	3.06E+00	1.41E+01	3.05E+01	∞ (0%)
g_7	3.31E-11	2.64E-01	1.97E-02	2.93E-02	4.05E-02	236800 (10%)
g_8	2.00E+01	2.02E+01	2.01E+01	2.01E+01	3.11E-02	∞ (0%)
g_9	1.14E-13	2.27E-13	1.71E-13	1.52E-13	3.71E-14	89450 (100%)
g_{10}	1.29E+02	3.42E+02	2.34E+02	2.38E+02	5.62E+01	∞ (0%)
g_{11}	2.46E+01	4.02E+01	3.11E+01	3.12E+01	3.23E+00	∞ (0%)
g_{12}	9.73E+01	1.58E+04	2.34E+03	3.62E+03	3.51E+03	∞ (0%)
g_{13}	9.34E-01	2.01E+00	1.48E+00	1.48E+00	3.07E-01	∞ (0%)
g_{14}	1.23E+01	1.40E+01	1.35E+01	1.35E+01	3.69E-01	∞ (0%)

each agent's local and global best positions), moves towards exploitation. However, having implemented the disturbance threshold, a dose of diversity (i.e. dt) is introduced in the population throughout the optimisation process, aiming to enhance the diversity of the algorithm.

Figure 1 illustrates the convergence of the population towards the optima and their diversities in three random trials over three benchmarks (i.e. $g_{1,7,9}$ chosen from the second set) as examples from unimodal and multimodal functions. The difference between the error and the diversity values demonstrates the algorithm's ability in exploration while converging to the optima whose fitness reach as low as 10^{-13} in g_1 and g_9.

Exploring the role of disturbance in increasing diversity, a control algorithm is proposed (DFO-c) where there is no disturbance ($dt = 0$) during the position update process.

The graphs in Fig. 2 illustrate the diversity of DFO-c populations in randomly chosen trials over three sample benchmarks (again $g_{1,7,9}$). The graphs illustrate that the diversity of the population in DFO-c is less than DFO, thus emphasising the impact of disturbance in injecting diversity which in turn facilitates the escape from local minima (e.g. as demonstrated in case of the highly multimodal Rastrigin functions f_5, g_9).

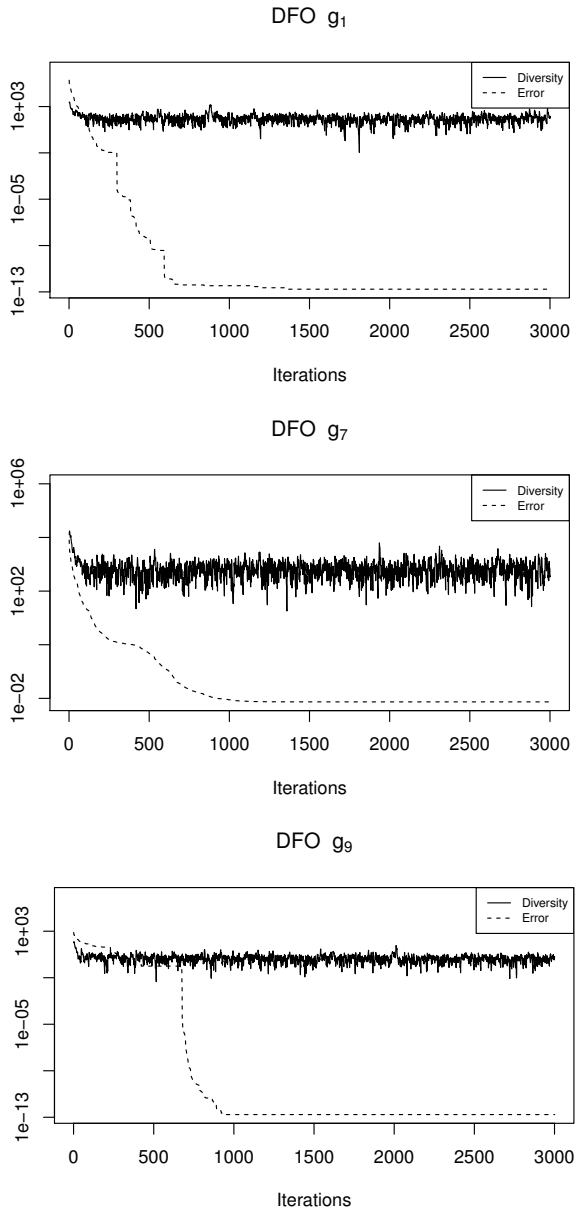


Fig. 1. DFO: diversity and error in $g_{1,7,9}$.

Note the gradual shrinkage of diversity in g_9 ($\approx 10^{-13}$) which is a clear indication of a premature convergence to a local minima with very poor chance of escape.

In order to compare the performance of DFO and its control counterpart, Table III presents the result of optimising the benchmarks using DFO-c. Additionally, a statistical analysis is conducted and the output is reported in Table IV where the performance is compared using the three aforementioned measures of error, efficiency and reliability (see Section V-B for the definitions of the measures). The results show that in 89% of cases (where there is a significant difference between the two algorithms), DFO is performing significantly better than its control counterpart (DFO-c) which is stripped from the

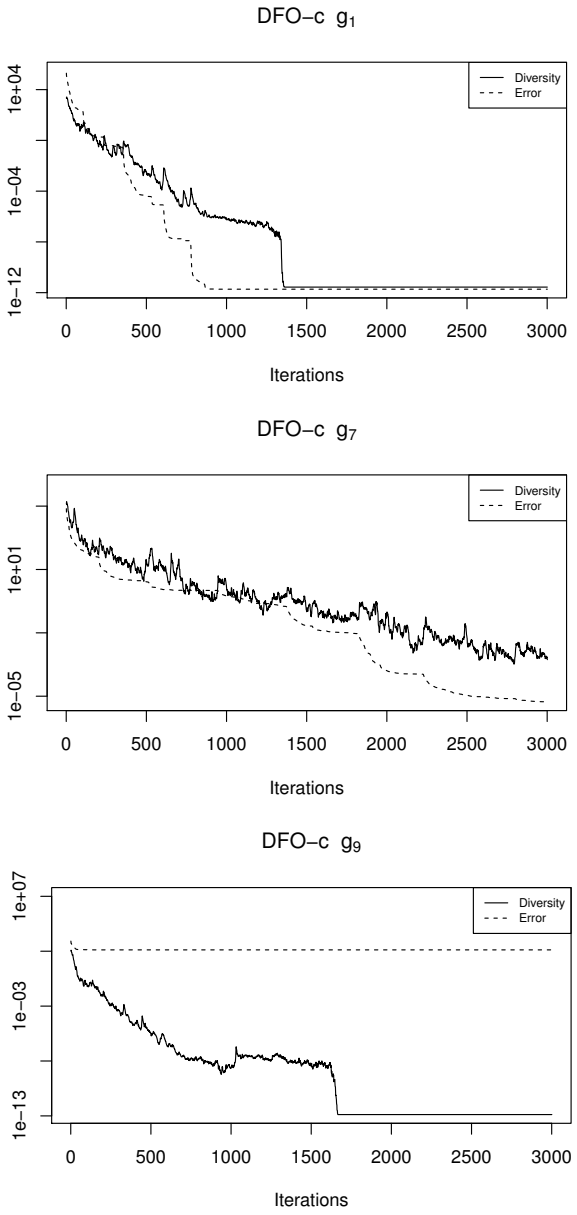


Fig. 2. DFO-c: diversity and error in $g_{1,7,9}$.

TABLE III
DFO-C – CONTROL DFO ALGORITHM

	Min.	Max.	Median	Mean	StdDev	Eff. (Rel.)
f_1	1.44E-56	3.09E-36	1.27E-45	9.65E-38	4.55E-37	65400 (100%)
f_2	7.29E-09	3.23E+01	1.28E-04	7.60E-01	4.60E+00	298200 (2%)
f_3	5.27E-05	1.61E+02	5.08E+00	1.67E+01	3.08E+01	∞ (0%)
f_4	4.48E-09	3.20E+03	1.55E+03	1.40E+03	8.66E+02	141500 (2%)
f_5	1.87E+02	4.17E+02	2.96E+02	2.94E+02	5.76E+01	∞ (0%)
f_6	1.97E+01	2.00E+01	1.98E+01	1.98E+01	5.24E-02	∞ (0%)
f_7	2.22E-16	6.00E+00	9.30E-02	3.51E-01	8.72E-01	64050 (8%)
f_8	1.03E-32	3.30E+02	2.14E+00	2.35E+01	5.84E+01	132950 (24%)
f_9	0.00E+00	1.57E+02	1.54E-01	5.35E+00	2.27E+01	176500 (30%)
f_{10}	0.00E+00	2.22E-16	0.00E+00	7.99E-17	1.08E-16	1700 (100%)
f_{11}	0.00E+00	8.10E+01	8.10E+01	5.99E+01	3.59E+01	2100 (26%)
f_{12}	5.05E+00	5.05E+00	5.05E+00	5.05E+00	0.00E+00	∞ (0%)
f_{13}	5.27E+00	5.27E+00	5.27E+00	5.27E+00	0.00E+00	∞ (0%)
f_{14}	5.36E+00	5.36E+00	5.36E+00	5.36E+00	0.00E+00	∞ (0%)
g_1	5.68E-14	9.37E-05	1.14E-13	1.91E-06	1.33E-05	70600 (94%)
g_2	1.68E-09	2.23E+01	1.23E-04	4.63E-01	3.14E+00	257700 (2%)
g_3	2.18E+05	5.38E+06	1.67E+06	1.73E+06	9.39E+05	∞ (0%)
g_4	2.23E+02	1.74E+04	1.80E+03	2.91E+03	3.36E+03	∞ (0%)
g_5	5.79E+03	1.38E+04	8.50E+03	8.69E+03	2.00E+03	∞ (0%)
g_6	2.25E-04	9.53E+01	8.61E+00	1.68E+01	2.52E+01	∞ (0%)
g_7	3.01E-10	2.13E-01	3.02E-02	4.17E-02	4.41E-02	263900 (2%)
g_8	2.00E+01	2.02E+01	2.00E+01	2.01E+01	3.89E-02	∞ (0%)
g_9	8.36E+01	2.64E+02	1.62E+02	1.64E+02	4.61E+01	∞ (0%)
g_{10}	1.22E+02	4.93E+02	2.69E+02	2.71E+02	7.69E+01	∞ (0%)
g_{11}	1.98E+01	4.11E+01	3.10E+01	3.13E+01	3.97E+00	∞ (0%)
g_{12}	2.32E+02	1.38E+04	3.04E+03	4.78E+03	3.88E+03	∞ (0%)
g_{13}	4.79E+00	3.56E+01	1.47E+01	1.58E+01	6.47E+00	∞ (0%)
g_{14}	1.28E+01	1.45E+01	1.36E+01	1.37E+01	3.38E-01	∞ (0%)

diversity inducing disturbance. Furthermore, in all multimodal functions (f_{3-9} and g_{6-12}), whenever there is a statistically significant difference between DFO and DFO-c, the former demonstrates significant outperformance over the later.

Following on the results from measuring error, Table IV also shows that in terms of efficiency and reliability measures, DFO is 79% more efficient than its control counterpart, and 92% more reliable.

E. Fine Tuning Disturbance Threshold

The role of disturbance in increasing the diversity of DFO population is discussed earlier (Section V-D). Also, the importance of disturbance is investigated on the optimisation capability of DFO by introducing a control algorithm which lacks the disturbance mechanism and the results demonstrate the positive impact of this mechanism.

The aim of this section is to recommend a value for the disturbance threshold, dt. The range of disturbance probabilities used in this experiment is between 1 to 10^{-9} and the values were chosen according to:

$$dt_n = 10^{-n}, \quad 0 \leq n \leq 9$$

Fig. 3 illustrates the performance of DFO using these dt probabilities. Both set of benchmarks (i.e. f_{1-14} and g_{1-14}) have been used to find a suitable value for the disturbance threshold. As the heat map highlights, the optimal range is $10^{-2} < dt < 10^{-4}$ and the overall recommended value of $dt = 10^{-3}$ is suggested as a good compromise.

F. Comparing DFO with other Population-Based Optimisers

Having presented the performance of the DFO algorithm (taking into account the three performance measures of error, efficiency and reliability, as well as the diversity of its population and the impact of disturbance on its behaviour), this section focuses on contrasting the introduced algorithm with few well-known optimisation algorithms. The three population

TABLE IV
COMPARING DFO AND DFO-C PERFORMANCE

Based on Wilcoxon 1×1 Non-Parametric Statistical Test, if the *error* difference between each pair of algorithms is significant at the 5% level, the pairs are marked. X–o shows DFO is significantly outperforming its counterpart algorithm; and o–X shows that the algorithm compared to DFO is significantly better than DFO. In terms of the *efficiency* and *reliability* measures, 1 – 0 (or 0 – 1) indicates that the left (or right) algorithm is more efficient/reliable. The figures, n – m, in the last row present a count of the number of X's or 1's in the respective columns.

	DFO – DFO-c		
	Error	**Efficiency**	**Reliability**
f_1	o – X	1 – 0	–
f_2	X – o	1 – 0	1 – 0
f_3	X – o	–	–
f_4	X – o	0 – 1	0 – 1
f_5	X – o	1 – 0	1 – 0
f_6	X – o	1 – 0	1 – 0
f_7	X – o	1 – 0	1 – 0
f_8	X – o	1 – 0	1 – 0
f_9	X – o	1 – 0	1 – 0
f_{10}	o – X	0 – 1	–
f_{11}	–	0 – 1	1 – 0
f_{12}	–	–	–
f_{13}	–	–	–
f_{14}	–	–	–
g_1	–	1 – 0	1 – 0
g_2	X – o	1 – 0	1 – 0
g_3	X – o	–	–
g_4	X – o	–	–
g_5	–	–	–
g_6	–	–	–
g_7	X – o	1 – 0	1 – 0
g_8	–	–	–
g_9	X – o	1 – 0	1 – 0
g_{10}	X – o	–	–
g_{11}	–	–	–
g_{12}	–	–	–
g_{13}	X – o	–	–
g_{14}	X – o	–	–
	16 – 2	11 – 3	11 – 1

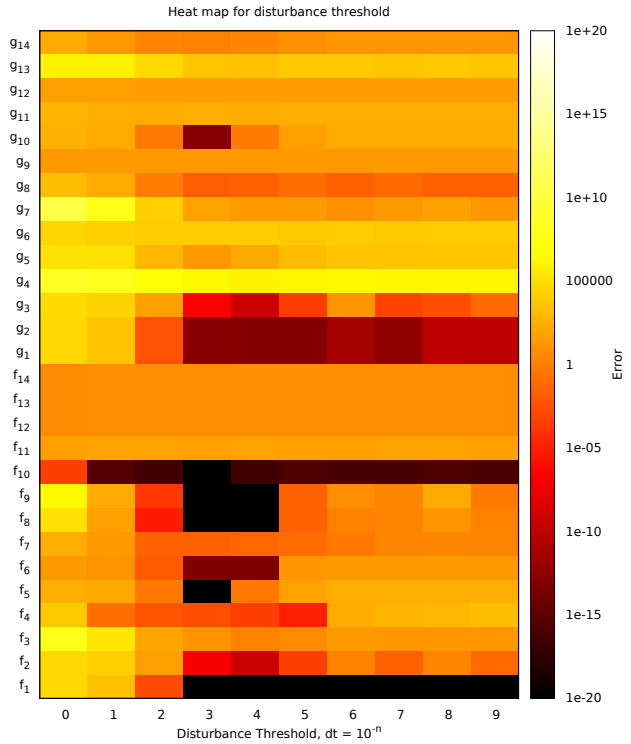


Fig. 3. Fine tuning disturbance threshold

TABLE V
DE (DIFFERENTIAL EVOLUTION), PSO (PARTICLE SWARM OPTIMISATION) AND GA (GENETIC ALGORITHM)

	DE		**PSO**		**GA**	
	Error	**Eff. (Rel.)**	**Error**	**Eff. (Rel.)**	**Error**	**Eff. (Rel.)**
g_1	1.38E-13	21500 (100%)	5.23E-14	656236 (100%)	5.04E-05	∞ (0%)
g_2	1.72E-07	∞ (0%)	1.33E-01	∞ (0%)	1.21E+04	∞ (0%)
g_3	9.65E+06	∞ (0%)	1.52E+06	∞ (0%)	1.47E+07	∞ (0%)
g_4	4.92E-01	∞ (0%)	7.89E+03	∞ (0%)	5.13E+04	∞ (0%)
g_5	2.34E+03	∞ (0%)	5.04E+03	∞ (0%)	2.09E+04	∞ (0%)
g_6	2.30E+00	265800 (12%)	2.16E+01	∞ (0%)	7.23E+02	∞ (0%)
g_7	5.39E-01	∞ (0%)	1.04E-02	279653 (10%)	5.48E+03	∞ (0%)
g_8	2.09E+01	∞ (0%)	2.09E+01	∞ (0%)	2.04E+01	∞ (0%)
g_9	3.47E+01	∞ (0%)	9.59E+01	∞ (0%)	2.20E+01	∞ (0%)
g_{10}	1.47E+02	∞ (0%)	1.14E+02	∞ (0%)	1.39E+02	∞ (0%)
g_{11}	3.65E+01	∞ (0%)	3.00E+01	∞ (0%)	1.17E+01	∞ (0%)
g_{12}	5.85E+05	∞ (0%)	9.51E+03	∞ (0%)	8.14E+03	∞ (0%)
g_{13}	5.70E+00	∞ (0%)	5.35E+00	∞ (0%)	2.70E+00	∞ (0%)
g_{14}	1.34E+01	∞ (0%)	1.25E+01	∞ (0%)	1.39E+01	∞ (0%)

algorithms deployed for this comparison are Differential Evolution (DE), Particle Swarm Optimisation (PSO) and Genetic Algorithm (GA). These algorithms are briefly described earlier in Sections IV-A, IV-B and IV-C. Generic versions of each algorithm are used against the generic version of Dispersive Flies Optimisation. In this comparison, only the second and the more challenging set of benchmarks, g_{1-14} are used. Table V presents the optimising results of the aforementioned algorithms, and as shown, the algorithms have optimised some of the benchmark to the specified accuracy, 10^{-8}. Table VI shows the result of the statistical analysis comparing DFO with the other three optimisers. Based on this comparison, whenever there is a significant difference between the performance of DFO and the other algorithms, DFO significantly outperforms DE, PSO and GA in 66.67%, 58.33% and 85.71% of the cases, respectively. Table VII summaries the efficiency results of the three optimisers with that of DFO; note that only the efficiency of functions reaching the specified error is given. As shown in the table, DFO, in the majority of cases, outperforms the other algorithms. In other words, although, when compared with DE, DFO only outperforms marginally (60%), it outperforms both PSO and GA in all cases (100%). The reliability comparison of DFO with the other optimisers is given in Table VIII. DFO is shown to be the most reliable algorithm in this comparison.

TABLE VI
COMPARING ERROR IN DFO WITH DE, PSO AND GA

Based on Wilcoxon 1×1 Non-Parametric Statistical Test, if the difference between each pair of algorithms is significant at the 5% level, the pairs are marked. X–o shows that the left algorithm is significantly better than the right one; and o–X shows that the right one is significantly better than the left. n – m in the row labeled Σ is a count of the number of X's in the columns above.

	DFO - DE	**DFO - PSO**	**DFO - GA**
g_1	–	o – X	X – o
g_2	X – o	X – o	X – o
g_3	X – o	–	X – o
g_4	o – X	X – o	X – o
g_5	o – X	o – X	X – o
g_6	o – X	X – o	X – o
g_7	X – o	o – X	X – o
g_8	X – o	X – o	X – o
g_9	X – o	X – o	X – o
g_{10}	o – X	o – X	o – X
g_{11}	X – o	–	o – X
g_{12}	X – o	X – o	X – o
g_{13}	X – o	X – o	X – o
g_{14}	–	o – X	X – o
$\sum$	8 – 4	7 – 5	12 – 2

TABLE VII
COMPARING EFFICIENCY IN DFO WITH DE, PSO AND GA

In this table, 1 – 0 (0 – 1) indicates that the left (right) algorithm is more efficient. The figures, n – m, in the last row present a count of the number of 1's in the respective columns. Note that non-applicable functions have been removed from the table.

	DFO - DE	**DFO - PSO**	**DFO - GA**
g_1	0 – 1	1 – 0	1 – 0
g_2	1 – 0	1 – 0	1 – 0
g_6	0 – 1	–	–
g_7	1 – 0	1 – 0	1 – 0
g_9	1 – 0	1 – 0	1 – 0
$\sum$	3 – 2	4 – 0	4 – 0

TABLE VIII
COMPARING RELIABILITY IN DFO WITH DE, PSO AND GA

In this table, 1 – 0 (0 – 1) indicates that the left (right) algorithm is more reliable. The figures, n – m, in the last row present a count of the number of 1's in the respective columns. Note that non-applicable functions have been removed from the table.

	DFO - DE	**DFO - PSO**	**DFO - GA**
g_2	1 – 0	1 – 0	1 – 0
g_6	0 – 1	–	–
g_7	1 – 0	–	1 – 0
g_9	1 – 0	1 – 0	1 – 0
$\sum$	3 – 1	2 – 0	4 – 0

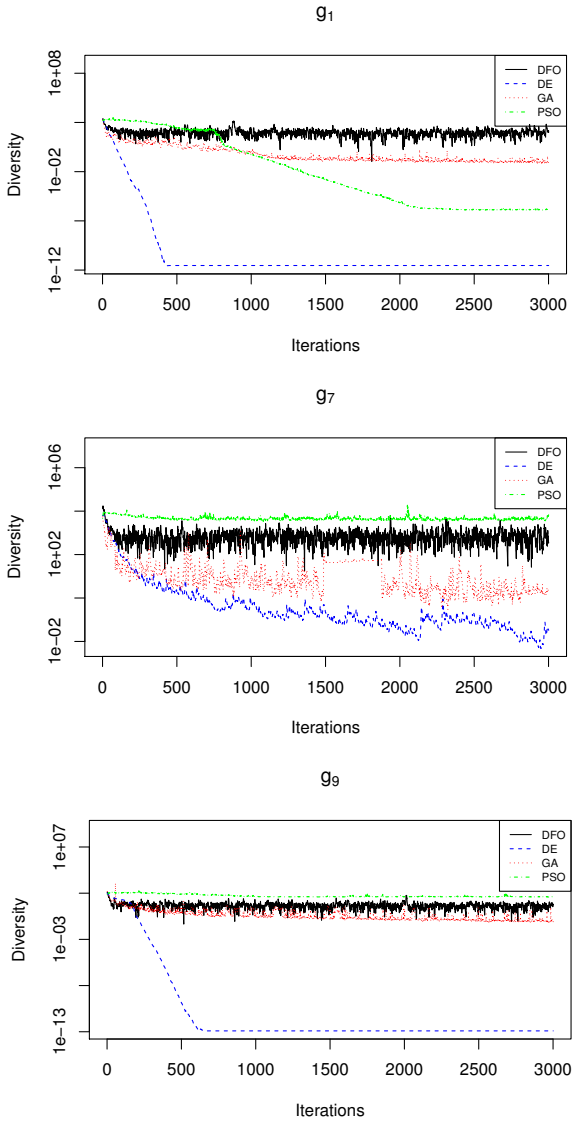


Fig. 4. Diversity of the population in DFO, DE, PSO and GA over three random trials in $g_{1,7 \text{ and } 9}$.

While DFO outperforms DE in 75% of cases, it show 100% outperformance when compared with PSO and GA. In order to compare the diversity of the DFO algorithm with the other three optimisers, three benchmarks were chosen from unimodal and multimodal categories ($g_{1,7,9}$). The result of this comparison is illustrated in Fig. 4. It is shown that DE has the least diversity in both uni- and multimodal functions. On the other hand, the diversity of the population in PSO decreases as the population converges towards an optimum (see g_1); however, when convergence does not occur (e.g. in $g_{7,9}$), PSO maintain its high diversity throughout the optimisation process. GA shows a similar pattern to that of PSO in multimodal functions, which is the gradual diversity decrease over time; however it maintains a higher diversity for the unimodal function than PSO (perhaps attributable to the difference in the fitness of the best positions found in both algorithms). In

terms of DFO, diversity is less convergence-dependent and more stable across all modalities.

VI. Conclusion

Dispersive Flies Optimisation (DFO), a simple numerical optimiser over continuous search spaces, is a population based stochastic algorithm, proposed to search for an optimum value in the feasible solution space; despite its simplicity, the algorithm's competitiveness over an exemplar set of benchmark functions is demonstrated.

As part of the study and in an experiment, a control algorithm is proposed to investigate the behaviour of the optimiser. In this experiment, the algorithm's induced disturbance mechanism shows the ability to maintain a stable and convergence-independent diversity throughout the optimisation process. Additionally, a suitable value is recommended for the *disturbance threshold* which is the only parameter in the update equations to be optimised. This parameter controls the level of diversity by injecting a component-wise disturbance (or restart) in the flies, aiming to preserve a balance between exploration and exploitation.

In addition to diversity, DFO's performance has been investigated using three other performance measures (i.e. error, efficiency and reliability). Using these measures, it is established that the newly introduced algorithm, outperforms few generic population based algorithms (i.e. differential evolution, particle swarm optimisation and genetic algorithm) in all of the aforementioned measures over the presented benchmarks. In other words, DFO is more efficient and reliable in 84.62% and 90% of the cases, respectively; furthermore, when there exists a statistically significant difference, DFO converges to better solutions in 71.05% of problem set.

A. Future Research

Much further research remains to be conducted on this simple new concept and paradigm. Among the possible future research are investigating the algorithm for an adaptive disturbance threshold, dt. Additionally, optimising multi-objective real world problems is yet to be researched; this would be a continuation of an earlier set of works on the deployment of population-based algorithms for detecting metastasis in bone scans and calcifications in mammographs [27]. At last, but not least, given the demonstrated stable and convergence-independent diversity of Dispersive Flies Optimisation (in the context of the presented benchmarks), another exciting future research is to investigate the performance of DFO in the context of dynamic optimisation problems.

References

[1] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1989.

[2] J. Kennedy and R. C. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, vol. IV. Piscataway, NJ: IEEE Service Center, 1995, pp. 1942–1948.

[3] M. Dorigo, M. Birattari, and T. Stutzle, "Ant colony optimization," *Computational Intelligence Magazine, IEEE*, vol. 1, no. 4, pp. 28–39, 2006.

[4] B. M. Wiegmann and D. K. Yeates, *Tree of Life: Diptera*. The Tree of Life Web Project, 1996.

[5] J. Downes, "Observations on the swarming flight and mating of culicoides (diptera: Ceratopogonidae) 1," *Transactions of the Royal Entomological Society of London*, vol. 106, no. 5, pp. 213–236, 1955.

[6] J. N. Belkin, N. Ehmann, and G. Heid, "Preliminary field observations on the behavior of the adults of anopheles franciscanus mccracken in southern california," *Mosq News*, vol. 11, pp. 23–31, 1951.

[7] H. T. Nielsen, "Swarming and some other habits of mansonia perturbans and psorophora ferox (diptera: Culicidae)," *Behaviour*, pp. 67–89, 1964.

[8] F. Knab, "The swarming of culex pipiens," *Psyche: A Journal of Entomology*, vol. 13, no. 5, pp. 123–133, 1906.

[9] L. M. Roth, "A study of mosquito behavior. an experimental laboratory study of the sexual behavior of aedes aegypti (linnaeus)," *American Midland Naturalist*, vol. 40, no. 2, pp. 265–352, 1948.

[10] R. T. Sullivan, "Insect swarming and mating," *The Florida Entomologist*, vol. 64, no. 1, pp. 44–65, 1981.

[11] W. Klassen and B. Hocking, "The influence of a deep river valley system on the dispersal of aedes mosquitos," *Bulletin of Entomological Research*, vol. 55, no. 02, pp. 289–304, 1964.

[12] J. Downes, "The swarming and mating flight of diptera," *Annual review of entomology*, vol. 14, no. 1, pp. 271–298, 1969.

[13] J. Downes, "Assembly and mating in the biting nematocera," *Intern. Congr. Entomol. Proc. 10th, Montreal*, pp. 425–34, 1958.

[14] R. L. Blickle, "Observations on the hovering and mating of tabanus bishopp," *Stone. Ann. Entomol. Soc. 52*, pp. 183–90, 1958.

[15] F. Heppner and U. Grenander, "A stochastic nonlinear model for coordinated bird flocks." *American Association for the Advancement of Science, Washington, DC(USA).*, 1990.

[16] D. Bratton and J. Kennedy, "Defining a standard for particle swarm optimization," in *Proc of the Swarm Intelligence Symposium*. Honolulu, Hawaii, USA: IEEE, 2007, pp. 120–127.

[17] R. Storn and K. Price, "Differential evolution - a simple and efficient adaptive scheme for global optimization over continuous spaces," 1995, tR-95-012, [online]. Available: http://www.icsi.berkeley.edu/ storn/litera.html.

[18] R. Thomsen, "Flexible ligand docking using evolutionary algorithms: investigating the effects of variation operators and local search hybrids," *Biosystems*, vol. 72, no. 1-2, pp. 57–73, 2003.

[19] J. Vesterstrom and R. Thomsen, "A comparative study of differential evolution, particle swarm optimization, and evolutionary algorithms on numerical benchmark problems," in *Evolutionary Computation, 2004. CEC2004. Congress on*, vol. 2, 2004, pp. 1980–1987.

[20] T. Back, D. B. Fogel, and Z. Michalewicz, *Handbook of evolutionary computation*. IOP Publishing Ltd., 1997.

[21] P. N. Suganthan, N. Hansen, J. J. Liang, K. Deb, Y. P. Chen, A. Auger, and S. Tiwari, "Problem definitions and evaluation criteria for the CEC 2005 special session on real-parameter optimization," Nanyang Technological University, Singapore and Kanpur Genetic Algorithms Laboratory, IIT Kanpur, Tech. Rep., 2005.

[22] J. Peña, "Theoretical and empirical study of particle swarms with additive stochasticity and different recombination operators," in *Proceedings of the 10th annual conference on Genetic and evolutionary computation*, ser. GECCO '08. New York, NY, USA: ACM, 2008, pp. 95–102. [Online]. Available: http://doi.acm.org/10.1145/1389095.1389109

[23] C.-Y. Lee and X. Yao, "Evolutionary programming using mutations based on the lévy probability distribution," *Evolutionary Computation, IEEE Transactions on*, vol. 8, no. 1, pp. 1–13, 2004.

[24] X. Yao, Y. Liu, and G. Lin, "Evolutionary programming made faster," *Evolutionary Computation, IEEE Transactions on*, vol. 3, no. 2, pp. 82–102, 1999.

[25] D. Gehlhaar and D. Fogel, "Tuning evolutionary programming for conformationally flexible molecular docking," in *Evolutionary Programming V: Proc. of the Fifth Annual Conference on Evolutionary Programming*, 1996, pp. 419–429.

[26] O. Olorunda and A. P. Engelbrecht, "Measuring exploration/exploitation in particle swarms using swarm diversity," in *Evolutionary Computation, 2008. CEC 2008.(IEEE World Congress on Computational Intelligence). IEEE Congress on*. IEEE, 2008, pp. 1128–1134.

[27] M. M. al-Rifaie and A. Aber, "Identifying metastasis in bone scans with stochastic diffusion search," in *Information Technology in Medicine and Education (ITME)*. IEEE, 2012. [Online]. Available: http://dx.doi.org/10.1109/ITiME.2012.6291355

Computer Science & Systems

CSS is a FedCSIS conference area aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to more technical aspects of computer science and related disciplines. The CSNS area spans themes ranging from hardware issues close to the discipline of computer engineering via software issues tackled by the theory and applications of computer science and to communications issues of interest to distributed and network systems. Events that constitute CSNS are:

- CANA'14 - 7th Workshop on Computer Aspects of Numerical Algorithms
- MMAP'14 - 7th International Symposium on Multimedia Applications and Processing
- SCoDiS-LaSCoG'14 - 3nd Workshop on Scalable Computing in Distributed Systems and 8th Workshop on Large Scale Computations on Grids

7th Computer Aspects of Numerical Algorithms

NUMERICAL algorithms are widely used by scientists engaged in various areas. There is a special need of highly efficient and easy-to-use scalable tools for solving large scale problems. The workshop is devoted to numerical algorithms with the particular attention to the latest scientific trends in this area and to problems related to implementation of libraries of efficient numerical algorithms. The goal of the workshop is meeting of researchers from various institutes and exchanging of their experience, and integrations of scientific centers.

Topics

- Parallel numerical algorithms
- Novel data formats for dense and sparse matrices
- Libraries for numerical computations
- Numerical algorithms testing and benchmarking
- Analysis of rounding errors of numerical algorithms
- Languages, tools and environments for programming numerical algorithms
- Numerical algorithms on GPUs
- Paradigms of programming numerical algorithms
- Contemporary computer architectures
- Heterogeneous numerical algorithms
- Applications of numerical algorithms in science and technology

Event Chairs

Bylina, Beata, Maria Curie-Sklodowska University, Poland

Bylina, Jarosław, Maria Curie-Sklodowska University, Poland

Stpiczyński, Przemysław, Maria Curie-Sklodowska University, Poland

Program Committee

Amodio, Pierluigi, Università di Bari, Italy

Anastassi, Zacharias, Qatar University, Qatar

Banaś, Krzysztof, AGH University of Science and Technology, Poland

Brugnano, Luigi, Universita' di Firenze, Italy

Czachorski, Tadeusz, IITiS

Filippone, Salvatore, University Rome Tor Vergata, Italy

Fourneau, Jean-Michel

Gansterer, Wilfried, University of Vienna, Austria

Georgiev, Krassimir, IICT - BAS, Bulgaria

Gimenez, Domingo, University of Murcia, Spain

Gravvanis, George, Democritus University of Thrace, Greece

Kozielski, Stanislaw

Kucaba-Pietal, Anna, Politechnika Rzeszowska, Poland

Lirkov, Ivan, Institute of Information and Communication Technologies , Bulgarian Academy of Sciences, Bulgaria

Maksimov, Vyacheslav, Institute of Mathematics and Mechanics, Russia

Marowka, Ami, Bar-Ilan University, Israel

Meini, Beatrice, Universita di Pisa, Italy

Minev, Peter, University of Alberta, Canada

Mycka, Jerzy, UMCS

Pekergin, Nihal

Petcu, Dana, West University of Timisoara, Romania

Pultarova, Ivana, Czech Technical University in Prague, Czech Republic

Satco, Bianca-Renata, Stefan cel Mare University of Suceava, Romania

Sedukhin, Stanislav, The University of Aizu, Japan

Sergeichuk, Vladimir, Institute of Mathematics of NAS of Ukraine, Ukraine

Srinivasan, Natesan, Indian Institute of Technology, India

Szajowski, Krzysztof, Institute of Mathematics and Computer Science, Poland

Telek, Miklos

Trivedi, Kishor S., Duke University, United States

Tudruj, Marek, Inst. of Comp. Science Polish Academy of Sciences/Polish-Japanese Institute of Information Technology, Poland

Tůma, Miroslav, Academy of Sciences of the Czech Republic, Czech Republic

Ustimenko, Vasyl, Marie Curie-Sklodowska University, Poland

Vazhenin, Alexander, University of Aizu, Japan

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 543–552
DOI: 10.15439/2014F183
ACSIS, Vol. 2

Solving Systems of Polynomial Equations: a Novel End Condition and Root Computation Method

Maciej Bartoszuk
Interdisciplinary PhD Studies Program,
Systems Research Institute, Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
Email: m.bartoszuk@phd.ipipan.waw.pl

***Abstract*—In this paper we present an improvement of the algorithm based on recursive de Casteljau subdivision over an n-dimensional bounded domain (simplex or box). The modification consists of a novel end condition and a way of calculation the root in subdomain. Both innovations are based on linear approximation of polynomials in a system. This improvement results in that our approach takes almost half of the time of the standard approach: it can be stopped much earlier than using standard diameter condition and getting midpoint of a subdomain as a root.**

I. Introduction

SOLVERS of systems of algebraic equations are a very important part of today's CAD/CAM systems. Issues such as the numerical representation of curves and surfaces [1], physical contacts between objects, collision detection, representations of Voronoi diagrams of set-theoretic models [2] or formulation of configuration space of a robot in motion planning applications [3] are reduced to finding solutions of the systems of polynomial equations.

Finding roots of polynomials (one equation) is well researched. Unfortunately, the problem of solving systems of polynomial equations goes back to the ancient Greeks and Chinese and still there is no one good solving method. There are a number of algorithms invented over the years based on different approaches. For details, see section II. It is worth noticing that in many of applications, especially CAD/CAM, like collision detection or curves intersection, there is no need to find all solutions in $\mathbb{R}^n$ and problem is narrowed to a bounded domain, which can be approximated (bounded) by n-dimensional box or simplex. That is why algorithms which can be applied only to a specific domain are so desirable.

The method described in this paper is a multidimensional bisection algorithm that uses multivariate Bernstein representation of polynomials, de Casteljau subdivision and convex hull property. Brief review of the algorithm is presented in next sections, but for further details we refer readers to [4], [5].

First of all, the method can be applied to an n-dimensional box or simplex domain. Such domains are used in curves and surfaces representation, so the algorithm can be used naturally in CAD/CAM applications. The further advantages of the method are a numerical stability, finding all zeros in the domain and no need for entering a starting point. A major drawback is the exponential complexity $O(2^n)$ where n is a number of equations. However, a number of publications in recent years have proved that this algorithm is effective for surfaces or curves intersections, where a number of equations in the system is equal to one or two [6].

In addition, research on the use of graphics cards (CUDA technology) were conducted to improve the execution time of the algorithm [7].

Our contributions are twofold:

1) We show novel end condition in multidimensional bisection algorithm which is based on linear approximation of equations
2) We show a novel way of computing the root in multidimensional bisection algorithm which is based on solving a linear system of the linear approximations

In this paper we consider a unit n-simplex as a domain of a system of polynomial equations.

The paper is organized as follows: in Section II we discuss related work, in Section III we briefly introduce the theoretical background for a multidimensional bisection algorithm. After that we describe two geometrical interpretations in Section IV. In the Section V we discuss the details of our method (especially end condition and calculation of the root in subdomain). In the Section VI reader can find numerical results obtained for various sets of equations. Section VII provides conclusions drawn from the presented analysis.

II. Related Work

Current methods of solving systems of polynomial equations can be divided into three groups: symbolical, geometrical and numerical (in particular subdivision) solvers.

A. Symbolical/Algebraic solvers

Symbolical methods are based on resultants and Gröbner bases. Those methods eliminate variables and reduce problem to finding roots for univariate polynomial using rational univariate representation. Those methods, however, are efficient only for systems of three up to four polynomials of low degree, such as 2 or 3. After reducing those polynomials, we obtain one univariate polynomial of degree close to 15. As it was shown by Wilkinson [8], computing roots of polynomial of degree greater than 15 can be ill-conditioned. What is more, these methods are difficult to implement for computers that

978-83-60810-58-3/$25.00

use finite precision arithmetic and that also slows down the resulting algorithm. Symbolical methods should be considered a success in theoretical area, but practical impact is unclear. An example of such an approach is [9], [10].

B. Geometrical solvers

For some specific applications, particular methods have been developed. Those methods use geometrical formulation of a problem. For example, for curves and surface intersection or ray tracing the algorithms are based on subdivision. However, these methods have limited applications in the general case [11].

C. Numerical/Analytic solvers

Numerical methods are probably the most known. They can be classified into iterative methods and homotopy methods. The most popular iterative method is Newton's method and it works well locally and only if initial point is a good guess, which is difficult in solving systems of polynomial equations. Other methods are Newton like methods, minimization methods or Weierstrass method [12], [13].

Homotopic methods have a well-explored theoretical background. They are based on proceeding path in a complex space. Theoretically, every path should converge to an isolated solution. In practice, however, there are many issues as the paths are not geometrically isolated, which causes problems with robustness of the approach. Moreover, those methods are rather computationally demanding. Example of this approach is [14], [15], [16].

Subdivision methods use an exclusion criterion to remove a domain if it does not contain a root. These solvers are often used to isolate the real roots. Exclusion criteria are based on Taylor exclusion function [17], interval arithmetic [18], Turan test [19], Sturm method [20], Descartes rule [21]. In this paper we propose an improvement to subdivision algorithm, where excluding is based on properties of Bernstein representation of polynomials. More information about that method can be found in [4], [5], [7].

Interestingly, computing on GPU becomes more and more popular, recently there are attempts to calculate many numerical problems on GPU, in particular solving polynomial equations systems [7], [15], [22], [23].

III. Problem formulation and algorithm idea

Consider a set of n polynomial equations in n independent variables

$$\mathbf{p}(\mathbf{x}) = \mathbf{0} \tag{1}$$

where $\mathbf{p} = (p_1, p_2, \ldots, p_n)\colon S_1^n \to \mathbb{R}^n$ (S_1^n is a unit n-simplex). The problem is to calculate numerically, with a given accuracy ϵ, all real roots $\{\mathbf{x}_0\}$ of the system (1). The method can also be used when the number of equations is not equal to the number of variables, but it is not in the scope of this paper.

This algorithm uses a simplex (also known as barycentric) Bernstein representation of multivariate polynomials. Vectors of this basis are as follows (N - degree of Bernstein polynomial):

$$B^N_{(j_1,\ldots,j_n)}(x_1,\ldots,x_n) = \frac{N!}{j_0!\cdot j_1!\cdot\ldots\cdot j_n!}\cdot x_0^{j_0}\cdot x_1^{j_1}\cdot\ldots\cdot x_n^{j_n} \tag{2}$$

where

$$j_k \in \mathbb{N}\cup\{0\}, x_k \in [0,1]\wedge x_0+\cdots+x_n \le 1\wedge k\in\{1,\ldots,n\}$$

and

$$j_1+\cdots+j_n \le N, j_0 = N-(j_1+\cdots+j_n), x_0 = 1-(x_1+\cdots+x_n)$$

so each polynomial p_k, after conversion to Bernstein representation b_k is of the form

$$b_k(\mathbf{x}) = \sum_{(j_1,\ldots,j_n)\in\{(j_1,\ldots,j_n):j_1+\cdots+j_n\le N\}} b^{(k)}_{(j_1,\ldots,j_n)} B^N_{(j_1,\ldots,j_n)}(\mathbf{x}) \tag{3}$$

and $b^{(k)}_{(j_1,\ldots,j_n)}$ are called Bernstein coefficients. We will also use coefficients of the system:

$$\mathbf{b}_{(j_1,\ldots,j_n)} = \begin{bmatrix} b^{(1)}_{(j_1,\ldots,j_n)} \\ b^{(2)}_{(j_1,\ldots,j_n)} \\ \vdots \\ b^{(n)}_{(j_1,\ldots,j_n)} \end{bmatrix}_{n\times 1}$$

From that point we will refer to a system of polynomial equations in Bernstein basis as $\mathbf{b}(\mathbf{x})$, so we can rewrite the main problem as

$$\mathbf{b}(\mathbf{x}) = \mathbf{0} \tag{4}$$

Definition III.1 (Extreme coefficients of polynomial in Bernstein form)**.** Extreme coefficients of polynomial in Bernstein form are

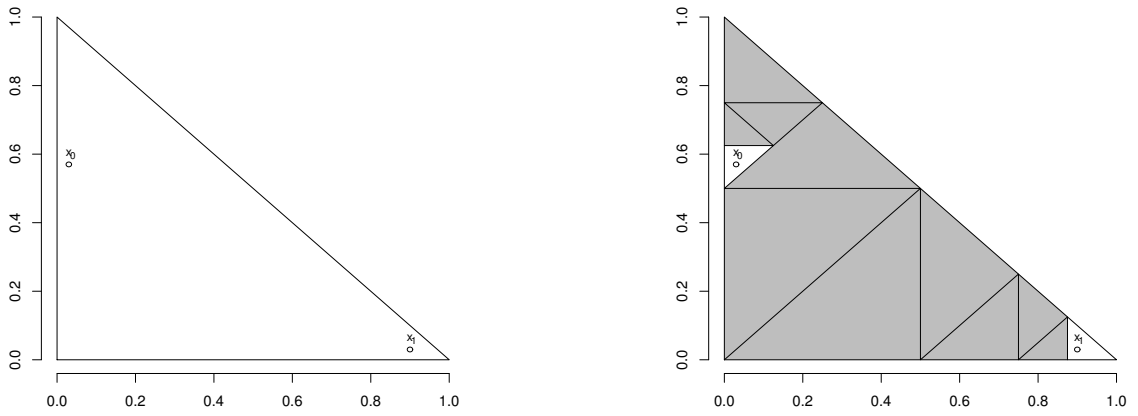
$$I_e = \{(j_1,\ldots,j_n)\colon j_1+\cdots+j_n \le N$$
$$\wedge(\forall_{k\in\{1,\ldots,n\}} j_k = 0 \vee \exists_{k\in\{1,\ldots,n\}} j_k = N)\wedge j_k\in\mathbb{N}\cup\{0\}\}$$

In other words, extreme coefficients of polynomial b_k are $b^{(k)}_{0,\ldots,0}$, $b^{(k)}_{N,0,\ldots,0}$, …, $b^{(k)}_{0,0,\ldots,N}$.

Example III.1. *For polynomial b_k of three variables and second degree extreme coefficients are $b^{(k)}_{000}$,$b^{(k)}_{200}$,$b^{(k)}_{020}$,$b^{(k)}_{002}$.*

A pseudocode is listed in Algorithm 1. Firstly, we have to convert input polynomials from $p_1,\ldots,p_n$ to $b_1,\ldots,b_n$. An efficient, robust algorithm is presented e.g. in [24]. After that, we create two queues S and Q, the first with root of the system and second with areas (simplexes) to be processed.

Routine is that: get one subsimplex (in the very beginning it is the initial simplex, where our system is defined), which means get coefficients of polynomials in the system defined on that subsimplex. After that, perform root exclusion tests. The tests can exclude subdomain, where there is no root with 100% certainty. Therefore, all of those tests have 100% true negative ratio. However, tests differ in false positive ratio. The fastest



(a) Domain before applying the algorithm

(b) Divided domain after applying the algorithm

Fig. 1: Algorithm idea, division of domain

Algorithm 1 Bisection algorithm to solve a system of polynomial equations

Require: Polynomials $b_1 \dots b_n$ in a system
Ensure: List of roots of the system
Initialize empty queue Q of n-tuples of polynomials;
Add $b_1 \dots b_n$ to Q
Initialize empty queue S of solutions;
while Q is not empty **do**
  Get n-tuple of $b_1 \dots b_n$ polynomials from Q
  Perform end condition test
  if End condition is true **then**
    Calculate root and add it to S queue
    CONTINUE;
  end if
  Divide $b_1 \dots b_n$ polynomials using de Casteljau method into $b_1^a \dots b_n^a$ and $b_1^b \dots b_n^b$ polynomials
  Perform tests (from fastest to slowest) and determine if every $b_1^a \dots b_k^a$ is suspected to have a root.
  if Every $b_1^a \dots b_n^a$ is suspected to have a root **then**
    Add $b_1^a \dots b_n^a$ into Q;
  end if
  Perform tests (from fastest to slowest) and determine if every $b_1^b \dots b_n^b$ is suspected to have a root.
  if Every $b_1^b \dots b_n^b$ is suspected to have a root **then**
    Add $b_1^b \dots b_n^b$ into Q;
  end if
end while
return S

test has the highest false positive ratio (the most subdomains are not excluded from consideration even though they should be). The slowest test has the lowest false positive ratio and is used only for those subdomains, which are not excluded by earlier tests. Details about the tests can be found in section V.

It should be stressed here that in general, de Casteljau division produces $n+1$ smaller simplexes from one input simplex. However, for simplicity and universality of the algorithm regardless of the number of equations, we make subdivision along one consecutive variable x_i, producing 2 smaller simplexes. Subdivision along the consecutive variable x_i has one more important advantage: we are assured that subdomains (diameter of subdomain) will be decreasing. More information about the diameter can be found in the section V. In other words, input simplex is a domain, and two output simplexes are subdomains, which added (in a set theory sense) are equal to input domain. Intersection of interiors of the subdomains is empty. Obtained polynomials, $b_1^a, \dots, b_n^a$ and $b_1^b, \dots, b_n^b$, are the same as input polynomials, but specified to new, smaller (scaled) unit simplex subdomain.

Subdomains not excluded from consideration are enqueued to Q.

Every subdomain dequeued from Q is tested for the end condition. If it is true, the root is calculated and enqueued in S. Thus the subdomain is excluded from consideration. The end condition and the root calculation are the centerpiece of this paper and will be discussed in section V.

When queue Q is empty, algorithm returns queue of solutions S.

In the Fig. 1 we can see an exemplary use of the algorithm. We have a two-dimensional domain (it means $n = 2$, so there are two polynomial equations in a system) and two roots x_0 and x_1 in the domain. The grey area represents a subdomain excluded by tests. We can see that the de Casteljau division is

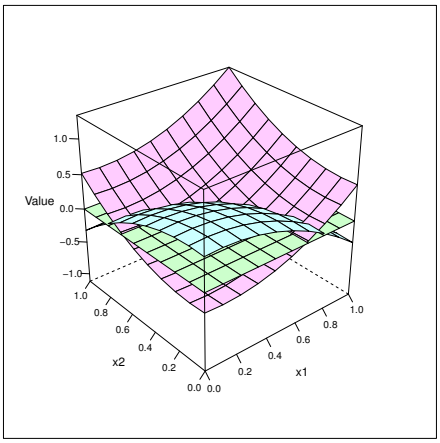


Fig. 2: Extended space representation for 2-dimensional box

performed along the longest edge of subsimplex, so diameters of the consecutive subsimplexes are getting smaller.

IV. Geometrical interpretations

Note that the problem can be geometrically interpreted in two ways.

A. Extended space representation

The extended space representation is used in the end condition and the root calculation. Consider a coordinate system with $n+1$ axes. n of them represent a domain. If we consider polynomial $p_k(\mathbf{x})$, the vector $\mathbf{x}$ has coordinates in those n axes. The $n+1$ axis is the value of the polynomial, so it is the value of $p_k(\mathbf{x})$. It implies that the polynomial $p_k(\mathbf{x})$ creates a surface in the coordinate system, where every point of that surface consists of n coordinates from domain (simplex) and last one which is value of the polynomial in that point.

Solving a system of polynomial equations (and equations in general) is finding points, where all of the surfaces have a zero value in the last, $n+1$, coordinate. The root is a point that consist of first n coordinates of such a point.

In the Fig. 2 we can see an example for n-dimensional box, where $n = 2$. One polynomial in the domain is pink, the second is blue, and value of zero is represented by a green plane.

B. n-dimensional mapping representation

The n-dimensional mapping representation can be applied to any system of functions. We can see the system as mapping $\mathbf{B}\colon S_1^n \to \mathbb{R}^n$. This function maps simplex S_1^n to $\mathbf{B}(S_1^n)$, so coordinate system has n axes.

This geometrical interpretation is used to understand tests that exclude subdomains from consideration. Those tests base on a Bernstein polynomial convex hull property, where coefficients $\mathbf{b}_{(j_1,\ldots,j_n)}$ of the system are the corners. More details can be found in [4].

Please note that:

$$(\mathbf{0} \in \mathbf{B}(S_1^n)) \Leftrightarrow ((\exists \mathbf{x}_0 \in S_1^n)\colon \mathbf{B}(\mathbf{x}_0) = \mathbf{0}) \tag{5}$$

Of course, even though this logical relationship is true, it is not very practical. It would be computationally demanding to check $\mathbf{0} \in \mathbf{B}(S_1^n)$. However, we can use the convex hull property of Bernstein polynomial:

$$(\exists \mathbf{x}_0 \in S_1^n\colon \mathbf{w}(\mathbf{x}_0) = \mathbf{0}) \Rightarrow (\mathbf{0} \in \operatorname{conv}(\{\mathbf{b}_{(j_1,\ldots,j_n)}\colon j_1 + \cdots + j_n \leq N\}) \tag{6}$$

Contraposition, however, of the above may be more convenient in practice:

$$(\mathbf{0} \notin \operatorname{conv}(\{\mathbf{b}_{(j_1,\ldots,j_n)}\colon j_1 + \cdots + j_n \leq N\}) \Rightarrow (\forall \mathbf{x}_0 \in S_1^n\colon \mathbf{w}(\mathbf{x}_0) \neq \mathbf{0}) \tag{7}$$

After obtaining certainty that the convex hull of the system of polynomial equations does not contain $\mathbf{0}$ point, we can see that we are able to exclude that system as it has no root. And if it contains $\mathbf{0}$, we can only suspect that a root is present in the system (because it is only in a convex hull).

We can see an example in the Fig. 3. It is a system of two polynomial equations of second degree. In Fig. 3a we can see a domain of the system and in the 3b there is the same system after mapping $\mathbf{B}(S_1^n)$. Coefficients of the system are marked as circles. Convex hull of the system is a green polygon. Vertices of red polygon are extreme coefficients of the system and they are marked as red circles. It is worth noticing that this system has two roots in the domain and both of them map to $\mathbf{0}$ in $\mathbf{B}(S_1^n)$.

Definition IV.1 (Hyperplane)**.** A hyperplane passing through the point h_0 and defined by a vector $\mathbf{p}$ is a set of points:

$$H^m(\mathbf{p}, \mathbf{h}_0) = \{\mathbf{h} \in \mathbb{R}^m\colon \mathbf{p}^T \cdot (\mathbf{h} - \mathbf{h}_0) = 0, \mathbf{h_0}, \mathbf{p} \in \mathbb{R}^m, \mathbf{p} \neq \mathbf{0}\} \tag{8}$$

where $\cdot$ is a dot product.

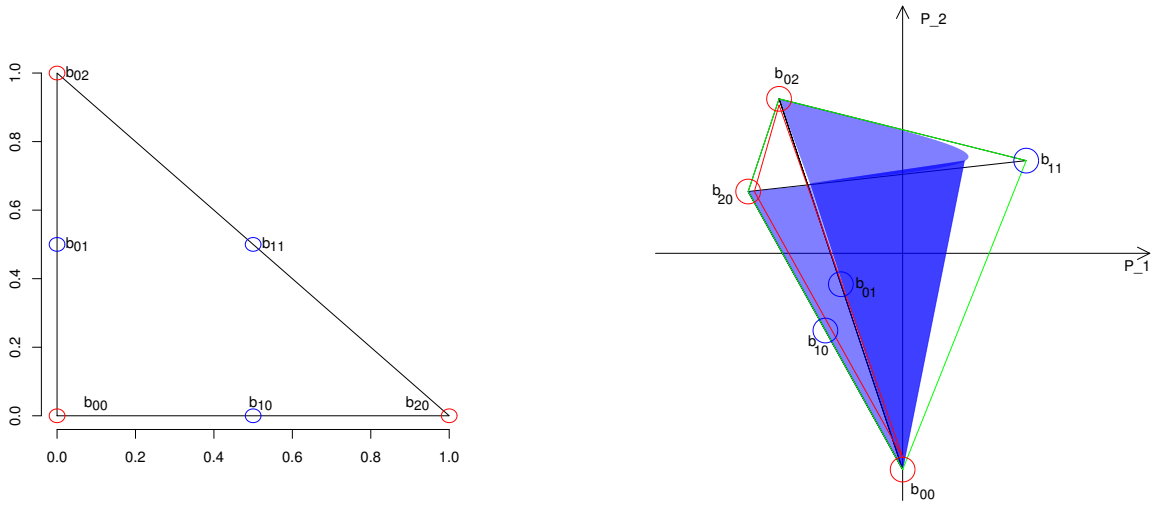
V. Algorithm details

All three described tests are based on hyperplanes. It can be proved that if we can find a hyperplane which separates the convex hull and $\mathbf{0}$, then convex hull does not include $\mathbf{0}$ point. More complicated tests use more interesting (and computationally damanding) hyperplanes.

As we wrote above, all three tests are based on finding a hyperplane that separates the convex hull and $\mathbf{0}$ point. Testing if all vertices of the convex hull are on one side of a hyperplane can be computed by checking a sign of dot product. More details about excluding tests can be found in [4].

A. Test of signs

This test checks hyperplanes $H^n(\mathbf{e}_i, \mathbf{0})$, where $\mathbf{e}_i$ is a base unit vector, for $i = 1, \ldots, n$. It can be performed as checking, if every polynomial in a system has coefficients of both signs, positive and negative. If at least one polynomial has all coefficients positive only or negative only, this test rejects that system as it does not have a root.



(a) S_1^n (b) $\mathbf{B}(S_1^n)$

Fig. 3: n-dimensional mapping representation

B. Test of midpoint

Let $M = \binom{n+N}{N}$ be a number of polynomial's coefficients. This equality can be proved by *Stars and bars* theorem, which can be found in [25].

Midpoint test (which should be performed as a second one) calculates a vector which is an average of Bernstein coefficients of set:

$$\mathbf{p_{mid}} = \frac{1}{M} \sum_{(j_1,\ldots,j_n)\in\{(j_1,\ldots,j_n):j_1+\cdots+j_n\leq N\}} \mathbf{b}_{(j_1,\ldots,j_n)}$$

Test verifies if all coefficients (points) are on one side of hyperplane $H^n(\mathbf{p_{mid}}, 0)$.

C. Test of hyperplanes

The last, most complex test is a side test. Hyperplanes in this test are determined by Bernstein coefficients of the system. We get every n linearly independent points from Bernstein coefficients $\mathbf{b}_{(j_1,\ldots,j_n)}$ and create a hyperplane H^n. Again, coefficients of the system are tested if they are on one side of the hyperplane (exclude) or not (further subdivision is needed).

There are two variants of this test. The first is that we get n linear independent points from extreme coefficients only. The second takes all coefficients into account. We have decided to implement the second variant in our work, as it tests more hyperplanes and can exclude more subdomains.

D. End condition and a root computation

The end condition used in literature is checking a diameter of the subdomain. A diameter, by definition, is the longest side of a simplex. If it is less than tolerance ϵ, algorithm assumes this area has a root and returns the midpoint of the simplex (adds it to the list of solutions). Such an approach can be found in [4], [5], [7].

We have found better approach, where a lower recursion level is needed. Using the de Casteljau division gives us smaller domains and we have observed that polynomials in these smaller domains are getting closer to linear functions.

We decided to check if polynomials on the subdomain are close enough to linear functions. If they are, then we change the problem to a system of linear equations (all polynomials on the subdomain are substituted to linear approximations) and solve it (find intersections of those approximations). Details are discussed below. In this part of the paper we should think of it according to the extended space representation, where every polynomial makes its own surface over the domain.

Definition V.1 (Control points of polynomial in Bernstein form)**.** Control point $\mathbf{b}^{(\mathbf{k})}_{(\mathbf{j_1},\ldots,\mathbf{j_n})}$ corresponding to a coefficient $b^{(k)}_{(j_1,\ldots,j_n)}$ of a polynomial b_k is a point (according to extended space representation) which first n coordinates are coordinates of the corresponding point of the domain (see Fig. 3a) and last $n+1$ coordinate is the value of coefficient $b^{(k)}_{(j_1,\ldots,j_n)}$.

Definition V.2 (Extreme control points of polynomial in Bernstein form)**.** Extreme control point $\mathbf{b}^{(\mathbf{k})}_{\mathbf{i_e}}$ $(i_e \in I_e)$ is a control point corresponding to an extreme coefficient $b^{(k)}_{i_e}$ of a polynomial b_k.

We can see an example in the Fig. 4. It is a surface created by polynomial according to extended space representation. Extreme control points are marked as red and the rest of control points are marked as blue.

After defining extreme control points, we can say that, after concluding that a polynomial is sufficiently linear on a subdomain, we can create a hyperplane passing through the extreme control points of the polynomial and the hyperplane will be an aformentioned linear approximiation of the equation.

The last thing we have to specify is a method to determine

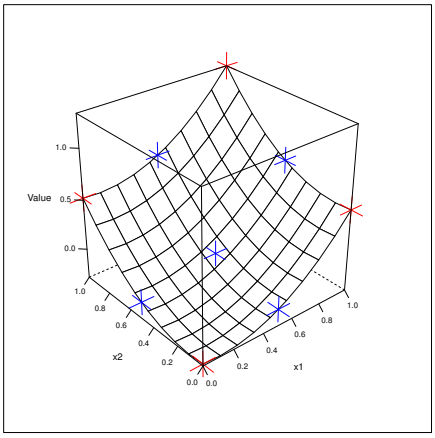


Fig. 4: Control points. Extreme control points are red.

when a polynomial is so similar to a linear function, that we can approximate the polynomial with the linear function with given error ϵ.

It should be stressed that there are $n+1$ extreme control points and we would denote them as ${}^{\mathbf{e}}\mathbf{b}_{\mathbf{0}}^{(\mathbf{k})}$, ${}^{\mathbf{e}}\mathbf{b}_{\mathbf{1}}^{(\mathbf{k})}$, ..., ${}^{\mathbf{e}}\mathbf{b}_{\mathbf{n}}^{(\mathbf{k})}$ for polynomial b_k.

Definition V.3 (Thickness of a polynomial in Bernstein form)**.** Consider extreme control points ${}^{\mathbf{e}}\mathbf{b}_{\mathbf{0}}^{(\mathbf{k})}$, ${}^{\mathbf{e}}\mathbf{b}_{\mathbf{1}}^{(\mathbf{k})}$, ..., ${}^{\mathbf{e}}\mathbf{b}_{\mathbf{n}}^{(\mathbf{k})}$ of polynomial b_k. Vectors ${}^{\mathbf{e}}\mathbf{b}_{\mathbf{1}}^{(\mathbf{k})} - {}^{\mathbf{e}}\mathbf{b}_{\mathbf{0}}^{(\mathbf{k})}$, ..., ${}^{\mathbf{e}}\mathbf{b}_{\mathbf{n}}^{(\mathbf{k})} - {}^{\mathbf{e}}\mathbf{b}_{\mathbf{0}}^{(\mathbf{k})}$ span an n-dimensional subspace. For every control point $\mathbf{b}_{(\mathbf{j_1},\ldots,\mathbf{j_n})}^{(\mathbf{k})}$ (not only *extreme* control points) we can calculate the distance from the subspace by projecting $\mathbf{b}_{(\mathbf{j_1},\ldots,\mathbf{j_n})}^{(\mathbf{k})}$ on this subspace and computing the distance $d_{(j_1,\ldots,j_n)}^{(k)}$ from this point $\mathbf{b}_{(\mathbf{j_1},\ldots,\mathbf{j_n})}^{(\mathbf{k})}$ to its projection. *Thickness* of a polynomial b_k is $\max\{d_{(j_1,\ldots,j_n)}^{(k)} : j_1 + \cdots + j_n \leq N\}$.

E. Calculating thickness

Assume that we have vector $\mathbf{p} \in \mathbb{R}^{n+1}$ and subspace of n-dimensions U. We want to project $\mathbf{p}$ on U. It means $\mathbb{R}^{n+1} = U \oplus V$ where $\dim V = 1$. We say that V is the complementary (orthogonal) subspace to U. That means that $\mathbf{p}$ breaks up into:

$$\mathbf{p} = \mathrm{proj}(\mathbf{p}, U) + \mathrm{proj}(\mathbf{p}, V) \tag{9}$$

So to find $\mathrm{proj}(\mathbf{p}, U)$, we can simply find $\mathrm{proj}(\mathbf{p}, V)$, which is a projection onto a 1-dimensional subspace.

Assume that vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ span U. We want to find vector $\mathbf{v}$ which is orthogonal to those vectors:

$$\begin{aligned} \mathbf{v} \cdot \mathbf{v}_1 &= 0 \\ \mathbf{v} \cdot \mathbf{v}_2 &= 0 \\ &\vdots \\ \mathbf{v} \cdot \mathbf{v}_n &= 0 \end{aligned}$$

We obtain linear system with n equations and $n+1$ unknowns. We have to add one constraint, e.g. $\mathbf{v}$ is unit vector or one of its coordinates is equal to 1. When we have $\mathbf{v}$, we get:

$$\mathrm{proj}(\mathbf{p}, V) = \frac{\mathbf{p} \cdot \mathbf{v}}{\mathbf{v} \cdot \mathbf{v}} \mathbf{v} \tag{10}$$

So

$$\mathrm{proj}(\mathbf{p}, U) = \mathbf{p} - \mathrm{proj}(\mathbf{p}, V) = \mathbf{p} - \frac{\mathbf{p} \cdot \mathbf{v}}{\mathbf{v} \cdot \mathbf{v}} \mathbf{v} \tag{11}$$

End condition is checking if thicknesses of all polynomials in system are less than ϵ. If so, linear approximations l_k of polynomials b_k are created and system of linear equations is solved. Solution of the linear system is taken as a root of the system of polynomial equations. If not, further subdivision is needed.

VI. Experimental results

The numerical experiments were performed on a PC with Intel Core i5-2500K CPU 3.30GHz and 8 GB of RAM. The goal of our research was to compare time and recursion level for different end conditions and root computation methods.

The input systems of equations differed in the number n of variables and equations (2 or 3) and the degree N (ranging from 1 to 20). Intersection of two or three curves or surfaces is a common case in CAD/CAM applications and that is why we limited our research to 2 or 3 equations.

Four cases were tested:

1) end condition was thickness of polynomials less than ϵ and root computation was solving a linear system of equations (our approach),
2) end condition was diameter of subdomain less than ϵ and root computation was returning midpoint of subdomain (standard approach),
3) end condition was diameter of subdomain less than ϵ and root computation was solving a linear system of equations,
4) end condition was thickness of polynomials less than ϵ and root computation was returning midpoint of subdomain.

Last two scenarios, which are mixture of our and standard approach, did not yield better results. In third case too many divisions were performed for given error ϵ, so time results were unsatisfactory. In fourth case too few divisions were performed, so accuracy of solutions was too small. Therefore results of genuine approaches are presented only.

We present results on Fig. 5–7. The tolerance ϵ for all tests was equal to 10^{-6}. It should be noted that all polynomials in a system were normalized so all coefficients were in set $[-1; 1]$. All polynomials in one tested system are of the same degree.

In Fig. 5 we can see maximal recursion level (number of de Casteljau divisions) needed to compute roots depending on the polynomials' degree. For standard approach it is constant number, because there is always the same number of divisions needed to obtain given diameter equal to ϵ of subdomain. Unlike standard approach, recursion level in our approach varies. Number of divisions depends on degree of polynomials. When polynomial is of higher degree, more recursions are required to get desirable linear approximation. When we use standard end condition, linear system is treated as always and many unnecessary subdivisions are performed.

It can be seen that obtaining the same thickness of hyperplane as diameter of subdomain (equal to ϵ) needs about half the number of recursions. What is more, a number of necessary recursions grows very slowly in degree of polynomials.

Nice feature of our approach is that, when a system of polynomial equations is in fact linear (all polynomials are polynomials of first degree), this case is detected in the first recursion level and system is solved by method dedicated for those systems.

In Fig. 6 we can see how much time it takes to compute roots depending on the polynomials' degree. As we can see, smaller number of recursion results in shorter time of computation. On average, our approach takes 60% time of the standard approach.

It can be seen that the method is good for polynomials of low degree. Detailed time of computation for those polynomials can be seen in Fig. 7. Unfortunately, for higher polynomials' degree time of computation becomes unacceptable: over 0.015 seconds for two equations (up to 0.05 for 20 degree) and over 0.5 seconds for three equations (up to 2.5 seconds for 20 degree).

We compared our approach with well-known application for computations, Mathematica. In Mathematica we used the function NSolve which is designed for numerical computation (Mathematica can perform symbol computations as well). We chose arguments "Reals" and "6", which means we are interested in real roots only and results should have 6 significant digits. What is more, in addition to the polynomial equations, we added inequalities such as $x \geq 0$ or $x + y \leq 1$.

Results of this comparison can be seen in Fig. 8. As we can see, the computation time of Mathematica is incomparable to our method so much we decided to plot it on a logarithmic scale. For example, for polynomials of fifth degree time of our method's computation is 0.0011 seconds, while Mathematica needs 0.0590 seconds. The time difference is even greater as a degree gets bigger and for degree of 16 execution times are 0.0190 and 4505 seconds, respectively.

Mathematica's function NSolve finds all roots in whole domain. We did not find method other than adding inequalities to bound domain. Adding inequalities means that NSolve finds all roots and after that it excludes those, which are not in domain. This is reflected at computation time. It is much worse than our approach. There is no sense using NSolve for polynomials of degree higher than five if we can bound domain in simplex.

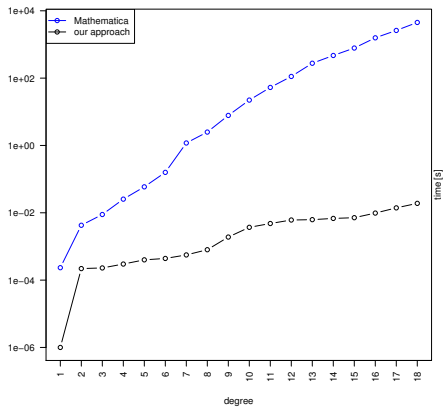


Fig. 8: Comparison of Mathematica and our approach computation time

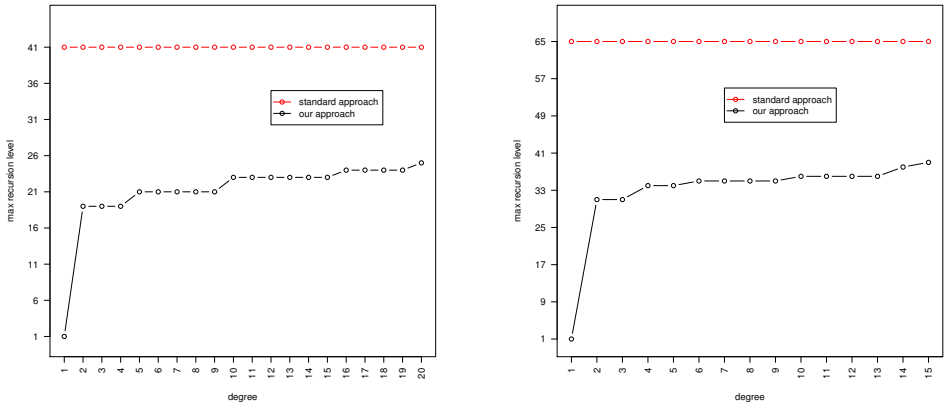
We tried to use NSolve for a system of three polynomial equations, but a difference of computation time was even bigger. For example, for three polynomials of fifth degree time of execution NSolve was 976 seconds.

VII. Conclusion

In this paper a new algorithm isolating the roots of systems of polynomial equations has been introduced. An algorithm based on the general multivariate Bernstein representation and the systematic search of a given domain has been implemented.

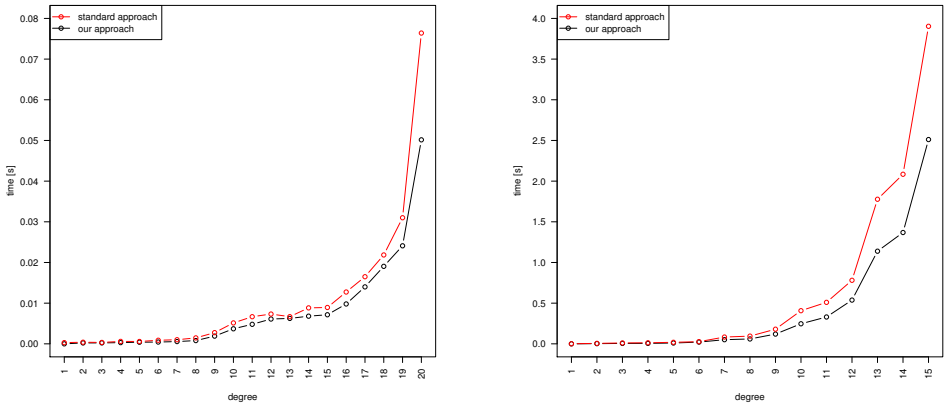
The main objective of this research was verification of applicability of novel end condition and computation of root by linear approximations. It was shown that this novel approach is nearly two times faster than standard methods.

Main advantage of this method is finding all roots in the given domain. Many known methods base on Newton's method which can converge to the solution in another simplex [12], [13]. This implies that some solutions may be missed and some may be found several times. Other methods, e.g. symbolic or homotopic, find all roots in $\mathbb{R}^n$, from which we can choose those in a domain (n-dimensional box or simplex). Unfortunately, this approach unnecessarily increase computation time, especially when most roots are outside of the domain, because many solutions are calculated and after that they are excluded from final set of roots. What is more, our algorithm is one of the fastest methods when there are no roots, because it is detected in the very first step. The next case, for which this method is doing exceptionally well is system of linear equations. In first step *thickness* is equal to zero and system is solved (approximations l_k and polynomials p_k from the system are the same).



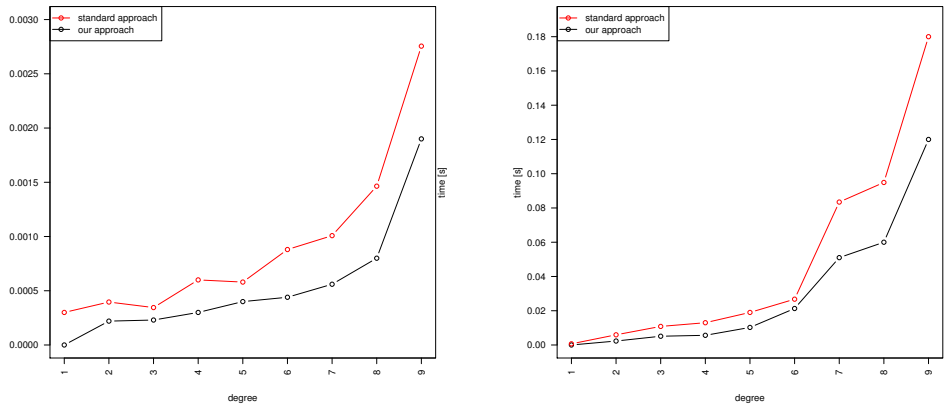
(a) Recursion level for two equations (b) Recursion level for three equations

Fig. 5: Recursion level depending on the polynomials' degree



(a) Computation time for two equations (b) Computation time for three equations

Fig. 6: Computation time depending on the polynomials' degree



(a) Computation time for two equations (b) Computation time for three equations

Fig. 7: Computation time depending on the polynomials' degree for 1–9 degrees

The algorithm has one interesting feature, when system has infinitely many solutions. If given ϵ is not too big, it can calculate some points which are roots of a system, and after connecting those points, we obtain a polyline, which approximates an infinite family of roots. A method for detecting this case in the beginning is needed.

The algorithm is at a disadvantage in the case where a root is in a corner or on a side of a subdomain. That causes many subdivisions and many simplexes to be considered because in many areas tests are passed. Detecting this case and evaluating only one simplex should be another area of exploration. However, it should be noted that a little perturbation of coefficients solves the problem.

A way for further development can be indicated. Some of the equations in the system can be close enough to linear function in earlier iterations of the algorithm than others. So we can substitute those equations to linear approximations. After that, one variable in the system (for one linear approximation) can be ousted. It would be very effective, because it would decrease the dimension of the problem. The algorithm has exponential computational complexity according to number of variables. That is why decreasing the number of variables is so important. One of the way to eliminate variables in a system of polynomial equations is using Gröbner basis.

Summarizing, this is an excellent method for surface intersections or for the visualization of curves and surfaces, because of finding all roots in a given domain. This is a well-conditioned numerical algorithm which can be useful for three to four equations [4]. It has some disadvantages like exponential computational complexity according to number of unknowns or vulnerability to roots in the corners or sides of a domain, but when we exclude those cases, it can be a powerful tool in computer graphics and computer-aided design.

ACKNOWLEDGMENT

Maciej Bartoszuk would like to acknowledge the support by the European Union from resources of the European Social Fund, Project PO KL "Information technologies: Research and their interdisciplinary applications", agreement UDA-POKL.04.01.01-00-051/10-00 via the Interdisciplinary PhD Studies Program.

REFERENCES

[1] T. Sederberg, D. Anderson, and R. Goldman, "Implicit representation of parametric curves and surfaces," *Computer Vision, Graphics, and Image Processing*, vol. 28, no. 1, 1984. doi: 10.1016/0734-189X(84)90140-3. [Online]. Available: http://dx.doi.org/10.1016/0734-189X(84)90140-3

[2] D. Lavender, A. Bowyer, J. Davenport, A. Wallis, and J. Woodwark, "Voronoi diagrams of set-theoretic solid models," *IEEE Computer Graphics and Applications*, pp. 69–77, September 1992. doi: 10.1109/38.156016. [Online]. Available: http://dx.doi.org/10.1109/38.156016

[3] J. Canny, "The Complexity of Robot Motion Planning," *MIT Press*, 1988. doi: 10.1017/S0263574700000151 ACM Doctoral Dissertation Award. [Online]. Available: http://dx.doi.org/10.1017/S0263574700000151

[4] J. Porter-Sobieraj, "Application of simplexes to the solving systems of algebraic equations," *Proc. of the IInd International Conference on Advances in Production Engineering 2001, Oficyna Wydawnicza PW, part II*, pp. 23–32, 2001.

[5] K. Marciniak, E. Pawelec, and J. Porter-Sobieraj, "Method for finding all solutions of systems of polynomial equations," *Proc. 7th IEEE Int. Conf. Methods and Models in Automation and Robotics*, vol. 1, pp. 155–158, 2001.

[6] J. Seland and T. Dokken, "Real-time algebraic surface visualization," *Geometrical Modeling, Numerical Simulation, and Optimization, Springer, Heidelberg*, pp. 163–183, 2007. doi: 10.1007/978-3-540-68783-2_6. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-68783-2_6

[7] J. Porter-Sobieraj and R. Klopotek, "Solving systems of polynomial equations on a GPU," *Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on, Series IEEE Computer Society Press(2012)*, pp. 539–544, 2012.

[8] J. Wilkinson, "The evaluation of the zeros of ill-conditioned polynomials. parts i and ii," *Numer. Math., 1:150-166 and 167-180*, 1959. doi: 10.1007/BF01386381. [Online]. Available: http://dx.doi.org/10.1007/BF01386381

[9] J. Chen, S. Lazard, L. Peñaranda, M. Pouget, F. Rouillier, and E. P. Tsigaridas, "On the topology of planar algebraic curves," *Mathematics for Computer Science. Special issue on Computational Geometry and Computer Aided Geometric Design*, vol. 4, no. 1, pp. 113–137, 2010. doi: 10.1145/1542362.1542424. [Online]. Available: http://dx.doi.org/10.1145/1542362.1542424

[10] S. Gao, F. V. IV, and M. Wang, "A new algorithm for computing grobner bases," 2010.

[11] L. Buse, M. Elkadi, and B. Mourrain, "Generalized resultants over unirational algebraic varieties," *J. of Symbolic Computation*, vol. 29, pp. 515–526, 2000. doi: 10.1006/jsco.1999.0304. [Online]. Available: http://dx.doi.org/10.1006/jsco.1999.0304

[12] D. Bini, "Numerical computation of polynomial zeros by means of aberth's method," *Numerical Algorithms*, vol. 13, no. 2, pp. 179–200, 1996. doi: 10.1007/BF02207694. [Online]. Available: http://dx.doi.org/10.1007/BF02207694

[13] B. Mourrain and O. Ruatta, "Relation between roots and coefficients, interpolation and application to system solving," *JSC*, vol. 33, pp. 679–699, 2002. doi: 10.1006/jsco.2002.0530. [Online]. Available: http://dx.doi.org/10.1006/jsco.2002.0530

[14] H.-J. Su, J. McCarthy, M. Sosonkina, and L. Watson, "Algorithm 857: POLSYS GLP: A parallel general linear product homotopy code for solving polynomial systems of equations," *ACM Trans. Math. Softw*, vol. 32, no. 4, pp. 561–579, 2006. doi: 10.1145/1186785.1186789. [Online]. Available: http://dx.doi.org/10.1145/1186785.1186789

[15] J. Verschelde and G. Yoffe, "Polynomial homotopies on multicore workstations," *Proceedings of the 4th International Workshop on Parallel Symbolic Computation (PASCO 2010), ACM*, pp. 131–140, 2010. doi: 10.1145/1837210.1837230. [Online]. Available: http://dx.doi.org/10.1145/1837210.1837230

[16] J. Verschelde, "Algorithm 795: PHCpack: A general-purpose solver for polynomial systems by homotopy continuation," *ACM Transactions on Mathematical Software (TOMS)*, vol. 25, no. 2, pp. 251–276, 1999. doi: 10.1145/317275.317286. [Online]. Available: http://dx.doi.org/10.1145/317275.317286

[17] J.-P. Dedieu and J.-C. Yakoubsohn, "Computing the real roots of a polynomial by the exclusion algorithm," *Numerical Algorithms*, vol. 4, no. 1, pp. 1–24, 1993. doi: 10.1007/BF02142738. [Online]. Available: http://dx.doi.org/10.1007/BF02142738

[18] R. B. Kearfott, "Interval arithmetic techniques in the computational solution of nonlinear systems of equations: Introduction, examples and comparisons," *Lectures in Applied Mathemetics. AMS Press*, pp. 337–357, 1990.

[19] V. Pan, "Optimal and nearly optimal algorithms for approximating polynomial zeros," *Computers & Mathematics with Applications*, vol. 31, no. 12, pp. 97–138, 1996. doi: 10.1016/0898-1221(96)00080-6. [Online]. Available: http://dx.doi.org/10.1016/0898-1221(96)00080-6

[20] M. Roy, "Basic algorithms in real algebraic geometry: from Sturm theorem to the existential theory of reals," *Exposition in Mathematics*, vol. 23, pp. 1–67, 1996.

[21] F. Rouillier and P. Zimmermann, "Efficient isolation of a polynomial real roots," *J. Comput. Appl. Math.*, 2003. doi: 10.1016/j.cam.2003.08.015. [Online]. Available: http://dx.doi.org/10.1016/j.cam.2003.08.015

[22] J. Verschelde and G. Yoffe, "Evaluating polynomials in several variables and their derivatives on a GPU computing processor," *IEEE Computer Society*, pp. 1391–1399, 2012. doi: 10.1109/IPDPSW.2012.177. [Online]. Available: http://dx.doi.org/10.1109/IPDPSW.2012.177

[23] S. Tomov, R. Nath, H. Ltaief, and J. Dongarra, "Dense linear algebra solvers for multicore with GPU accelerators," *Proceedings of the IEEE International Symposium on Parallel and Distributed Processing Workshops (IPDSW 2010) IEEE Computer Society*, pp. 1–8, 2010. doi: 10.1109/IPDPSW.2010.5470941. [Online]. Available: http://dx.doi.org/10.1109/IPDPSW.2010.5470941

[24] W. N. Waggenspack and D. C. Anderson, "Converting standard bivariate polynomials to Bernstein form over arbitrary triangular regions," *Comput. Aided Des.*, vol. 18, no. 10, pp. 529–532, Dec. 1986. doi: 10.1016/0010-4485(86)90040-0. [Online]. Available: http://dx.doi.org/10.1016/0010-4485(86)90040-0

[25] J. Pitman, *Probability*, ser. Springer Texts in Statistics. Springer, 1993. ISBN 9780387979748. [Online]. Available: http://books.google.pl/books?id=L6IWgaCuilwC

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 553–560
DOI: 10.15439/2014F190
ACSIS, Vol. 2

Accuracy Evaluation of Classical Integer Order and Direct Non-integer Order Based Numerical Algorithms of Non-integer Order Derivatives and Integrals Computations

Dariusz W. Brzeziński
Institute of Applied Computer Science
Lodz University of Technology
18/22 Stefanowskiego St., 90-924 Łódź, Poland
Email:dbrzezinski@kis.p.lodz.pl

Piotr Ostalczyk
Institute of Applied Computer Science
Lodz University of Technology
18/22 Stefanowskiego St., 90-924 Łódź, Poland
Email:piotr.ostalczyk@p.lodz.pl

***Abstract*—In this paper the authors evaluate in context of numerical calculations accuracy classical integer order and direct non-integer based order numerical algorithms of non-integer orders derivatives and integrals computations. Classical integer order based algorithm involves integer and fractional order differentiation and integration operators concatenation to obtain non-integer order. Riemann-Liouville and Caputo formulas are applied to obtain directly derivatives and integrals of non-integer orders. The following accuracy comparison analysis enables to answer the question, which algorithm of the two is burdened with lower computational error. The accuracy is estimated applying non-integer order derivatives and integrals computational formulas of some elementary functions available in the literature of the subject.**

I. Introduction

THERE are many formulas which can be applied to compute directly derivatives and integrals of non-integer orders [8], [9], [10], [11], [12], [14]. They include Riemann-Liouville non-integer (fractional) order integral/derivative and Caputo non-integer (fractional) derivative. However, non-integer order derivatives and integrals can be also computed applying integer and fractional order differentiation and integration operators concatenation. Classical integer order derivatives and integrals can be obtained applying well known numerical techniques. To calculate fractional order derivatives and integrals, Riemann-Liouville/Caputo formulas can be applied. The question, the authors want to answer is, which of the algorithms enables to obtain more accurate values of example non-integer derivatives and integrals of some elementary functions.

The paper is divided into the following parts: at the beginning the authors present details of the applied operators of non-integer (fractional) order differentiation and integration, followed by detailed description of the integer and fractional order differentiation and integration operators concatenation and the explanation of their practical numerical implementations. The final part of the paper include accuracy comparison analysis of example non-integer order derivatives and integrals of some elementary functions. The values assumed as exact for the accuracy comparison are calculated applying formulas of non-integer order derivatives and integrals available in the literature of the subject [11], [14].

II. Mathematical Preliminaries

Non-integer (fractional) order integration and differentiation operators include:

- Riemann-Liouville Fractional Order Integral

$$ {}_{t_0}^{RL}\mathrm{I}_t^{(\nu)} = \frac{1}{\Gamma(\nu)} \int_{t_0}^{t} \frac{f(\tau)}{(t-\tau)^{1-\nu}} d\tau \tag{1} $$

- Riemann-Liouville Fractional Order Derivative

$$ {}_{t_0}^{RL}\mathrm{D}_t^{(\nu)} f(t) = \frac{1}{\Gamma(n-\nu)} \left(\frac{d}{dt}\right)^n \int_{t_0}^{t} \frac{f(\tau)}{(t-\tau)^{1-\nu}} d\tau \tag{2} $$

- Caputo Fractional Derivative

$$ {}_0^{C}\mathrm{D}_t^{(\nu)} f(t) = \frac{1}{\Gamma(n-\nu)} \int_{0}^{t} \frac{f^{(n)}(\tau)}{(t-\tau)^{1-\nu}} d\tau \tag{3} $$

Formulas (2) and (3) are related by

$$ {}_{t_0}^{RL}D_t^{(\nu)} f(t) = {}_0^{C}D_t^{(\nu)} f(t) + \sum_{k=0}^{n-1} \frac{t^{k-\nu}}{\Gamma(k-\nu+1)} f^{(k)}(0) \tag{4} $$

where: $n-1 < \nu < n \in N = \{1, 2, ...\}$, $\nu \in R$ is the order of fractional integral/derivative, $f^{(n)}(\tau) = \frac{d^n f(\tau)}{d\tau^n}$ is the classical derivative of integer order, $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$ is the gamma function and t satisfies the following conditions $-\infty < t_0 < t < \infty$, $\lfloor n \rfloor = n + \nu$.

978-83-60810-58-3/$25.00

III. Applied Methods of Integration and Differentiation

A. Applied Methods for Classical Integer Orders Derivatives and Integrals Computations

- Classical integer order derivatives are calculated applying well known numerical methods involving 5 point stencil Central Diffidences.
- Classical integrals of the integer orders are obtained applying well known and efficient methods of numerical integration:
  - Gauss-Kronrod Quadrature (denoted as **GKr**)
    - The method is a very efficient modification of well known Gauss-Legendre Rule. It uses so called the Gauss-Kronrod Pairs. For example the G30/K61 pair includes the nodes of the 30-point Gauss-Legendre Quadrature + 31 new ones and all 61 different coefficients.
    - The advantages of the method is its efficiency and high accuracy obtained with only some dozens of sampling points if applied to precisely selected type of integrand. The main disadvantages are: a complex method of nodes and weight calculations, high precision input data requirement and indept knowledge about the method application to actually obtain high accuracy results.
  - Midpoint Rule (denoted as **NCm**)
    - The method is a very efficient modification of the rectangular rule. In the midpoint rule the sample point is taken from the middle of each subinterval. This feature enable the application of the method to the integrands with singularities at the end point of the integration range.
    - The method can be applied to any integrand, because the weight function equals 1 and the nodes of quadrature are of equal width. The accuracy of calculations should theoretically increase while increasing the amount of subintervals, for which the integration range is divided into. However mechanical increase of the amount of subintervals often leads to small accuracy increase and big increase in computational complexity.

All further details regarding integer order derivatives and integrals methods of computations can be found in available literature of the subject [1], [2], [3], [4], [5]. There is only presented, on Figs 2-3 indicative accuracy of the applied methods.

B. Applied New Efficient High-accuracy Methods of Integration for Non-integer and Fractional Orders Derivatives and Integrals Computations

The integrands in formulas (1)-(3) are difficult to integrate due to singularity at the end of integration range. The difficulty rises as the order of calculated fractional order integral nears 0 and the order of fractional order derivative nears 1. This feature makes inefficient widely known methods of numerical integration in context of accuracy of calculations. Additionally some of the fractional orders derivatives and integrals, mentioned in last sentence are not possible to compute with satisfactory accuracy at all (relative error exceeding 90-100 %).

Generally the problems associated with singularities in numerical integration are the most difficult to solve. In such cases high accuracy results can only be obtained by application of dedicated methods, as for example weighted type quadratures. This type of quadratures, however can only be applied to precisely selected types of functions in unmodified form. It is forced by their association with the corresponding weights. To obtain high accuracy results, either the integrand must satisfy conditions of a particular quadrature application requirements or the quadrature must be adopted to a particular integrand. Choosing the second method, the authors of the paper developed precise modifications to the existing numerical methods of integration: Gauss-Jacobi Quadrature and Double Exponential Quadrature, which initially were developed for integer order integration only.

Modified version of Gauss-Jacobi Quadrature have adopted weight function for fractional order derivatives and integrals computations (the method is denoted in the paper as **GJ**). Double Exponential Quadrature (denoted as **DE**) [16], [17] on the other hand involves hyperbolic functions substitution in independent variable transformation in integrand and trapezoidal rule applied to the transformed integrand. Application of the methods enables to obtain high accuracy computations results of fractional order derivatives and integrals [6], [7].

Important remarks:

- In the case of the non-integer (fractional) order differentiation and integration operators (1)-(3), regardless if there is integral or derivative to calculate, we always apply the integration operator.
- The non-integer (fractional) order differentiation operators (1)-(3) unlike integer order differentiation operators, are non-local operators. They are not calculated applying the values of the neighbor function points, but from the whole range of differentiation. This can be beneficial in case of a physical process analysis, because we take into consideration its history from the beginning, however this feature increases significantly the complexity of numerical calculations. Higher complexity of calculations influences negatively the accuracy of input data for each part of calculations. Inaccuracies in input data are the classical cause for numerical calculations accuracy decrease.

C. Concatenation of Operators of Integer and Fractional Order Differentiation and Integration as a Obtain Method of Non-integer Order

Scientifically interesting calculation method of non-integer order derivatives and integrals computations is application of concatenation of integer and fractional order differentiation and integration operators.

The practical operations of concatenations are related with fractional and integer order operators. Some of them are trivial,

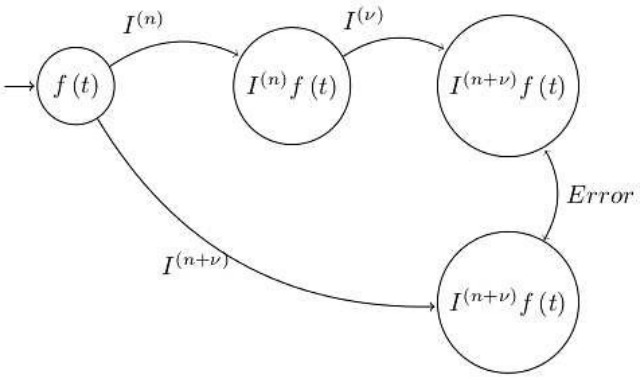


Fig. 1. Accuracy Estimation Based on Integer and Fractional Orders Integration Operators Concatenation

some require additional assumptions. There is always assumed the existence of the corresponding derivatives and integrals and that the operations of concatenation are always performed on the same definitive integration range.

Interesting concatenations combination of integer and fractional orders differentiation and integration in context of the paper's subject include:

- Fractional order derivative of integer order derivative

$$\mathrm{D}^{(\nu)}\left(\mathrm{D}^{(n)}f(t)\right)=\mathrm{D}^{(n+\nu)}f(t) \tag{5}$$

- Fractional order integral of integer order integral

$$\mathrm{I}^{(\nu)}\left(\mathrm{I}^{(n)}f(t)\right)=\mathrm{I}^{(n+\nu)}f(t) \tag{6}$$

where n is classical integer order derivative or integral, $\nu \geq 0$ and $t \geq 0$ (see Figure 1 for visualization).

Comprehensive mathematical calculations to the subject can be found in [14].

In the works of [9] and [10] discussed properties (5) and (6) are named *Law of Exponents*, because they relate to the operations performed on the exponents of the fractional and integer order operators (1)-(3). The *Law of Exponents* is generally true for the fractional integration operators. It is also valid for fractional order differentiation operators [14]. Still it is well known [10] and [9] in $\S IV.6$ *Law of Exponents*, that there exist functions for which fractional order differentiation operator (3) do not satisfy (5). However, the authors of the following paper focus only on functions, for which concatenations (5)-(6) are possible, in this case $f(t_0)=0$.

IV. Evaluation Details

The description of a general method of the fractional order differentiation and integration operators concatenation is presented in [10], [9]. It is expressed by the following formula

$$I^{(\nu)}\left[I^{(n)}f(t)\right]=\frac{1}{\Gamma(\nu)\cdot\Gamma(n)}\int_0^t(t-x)^{\nu-1}\cdot \cdot\left[\int_0^x f(y)\,dy\right]dx \tag{7}$$

We substitute $x=k*h$ in inner integral, where k is the amount of subintervals, which the integration range is divided into, and h is the width of each interval. As a result we obtain the values of the *inner* integral in h-spaced intervals. This values serve as input function $f(x)$ for the *outside* integral.

The integration methods applied in the following research require precisely sampled function's values in the nodes points and their synchronization with the corresponding weights. Sampling the function in equally spaced points and then interpolating the values of the function in the nodes points does not comply with its function sufficiently. There must be applied a different approach, which is presented as Algorithm 1.

Algorithm 1 Concatenation of integer and fractional order differentiation and integration operators. Nodes of a quadrature for $I^{(\nu)}$ as an integration step of $I^{(n)}$.

Step 1. Select fractional orders to concatenate $I^{(n)}$ and $I^{(\nu)}$. The $I^{(n)}$ is the *inner* integral and the $I^{(\nu)}$ is the *outer* one.

Step 2. Calculate quadrature nodes and weights for the $I^{(\nu)}$.

Step 3. Calculate $I^{(n)}$ applying as step the points of the nodes from *Step 2*.

Step 4. Calculate $I^{(\nu)}$ applying as inputting as $f(x)$ the values obtained in *Step 3*.

1) Arbitrary Precision in Numerical Calculations: To overcome the bottlenecks of the double precision computer arithmetic and to increase overall computations accuracy [18], two arbitrary precision programming libraries together with a C++ wrapper [22] are applied:

MPFR [21] is an arbitrary precision package for C language and is based on GMP [20].

MPFR supports arbitrary precision floating point variables. It also provides an exact rounding of all implemented operations and mathematical functions [19].

A. Testing Functions

The integrand in formulas (1)-(3) consists of two factors: the first factor, so called *core* and the second one, which is the actual function, of which there is non-integer (fractional) order derivative or integral to calculate.

The core has the biggest influence on the shape of the integrand and on the difficulty level of integration. In this respect, the actual function to integrate contributes only to a minimal extent.

Due to this, the authors decide it is enough to select two functions for testing purposes:

- Power function

$$f(t)=(t-t_0)^p,\ t_0=0,\ p=0.5\ ,t\in(0,1) \tag{8}$$

- Exponential function

$$f(t)=e^{at}\mathbf{I}(t),\ a=0.5,t\in(0,1) \tag{9}$$

B. Accuracy Estimation of Computations

Accuracy estimation is performed on the basis of the non-integer (fractional) order derivatives and integrals formulas (10)-(13) available in the literature of the subject [11], [14]. Due to the fact, that they are in fact computational formulas, there must be taken into consideration some calculations error, although very small.

Assuming $D^{(-\nu)} = I^{(\nu)}$.

- Power Function (8):
  - Fractional integral

$$ {}_{t_0}\mathrm{D}_t^{(-\nu)} f(t) = \frac{\Gamma(p+1)}{\Gamma(p+\nu+1)} (t-t_0)^{p+v}. \quad (10)$$

  - Fractional derivative

$$ {}_{t_0}\mathrm{D}_t^{(\nu)} f(t) = \frac{\Gamma(p+1)}{\Gamma(p+1-\nu)} (t-t_0)^{p-v}. \quad (11)$$

- Exponential Function (9):
  - Fractional integral

$$ {}_{t_0}\mathrm{D}_t^{(-\nu)} f(t) = t^{\nu} \sum_{k=0}^{N} \frac{(at)^k}{\Gamma(k+\nu+1)}. \quad (12)$$

  - Fractional derivative

$$ {}_{t_0}\mathrm{D}_t^{(\nu)} f(t) = t^{-\nu} \sum_{k=0}^{N} \frac{(at)^k}{\Gamma(k+1-\nu)}. \quad (13)$$

C. Accuracy Definition

In the whole paper the accuracy is expressed as relative error in % in context of integration range

$$ e_r^{(t)} = \left(1 - \frac{v_c}{v_e}\right) \cdot 100\% \quad (14)$$

where v_c denotes calculated value, v_e a value *assumed* as exact one and t_0, t is integration range.

V. Comparison Analysis

A. General remarks

The methods of numerical integration developed by the authors of the paper, to their best knowledge, are the only numerical methods of integration available at the moment, applying which it is possible to obtain high accuracy results calculating non-integer (fractional) derivatives and integrals:

- **GJ** method delivers high accuracy results with average accuracy above 10^{-80} mark. The order of the derivative and integral and integration range has no impact on accuracy. Actually, the method increases offered accuracy in cases, in which traditionally methods decrease it. The method requires 4-64 sampling points to reach average accuracy abilities. The type of integrated function influences only slightly the final results.
- **DE** is able to deliver 10^{-50} average accuracy level for fractional integrals of orders greater than 0.5 and fractional derivatives smaller that 0.5 with $600 - 1000$ sampling points. The methods is in general more dependable on range and type of the integrated function.

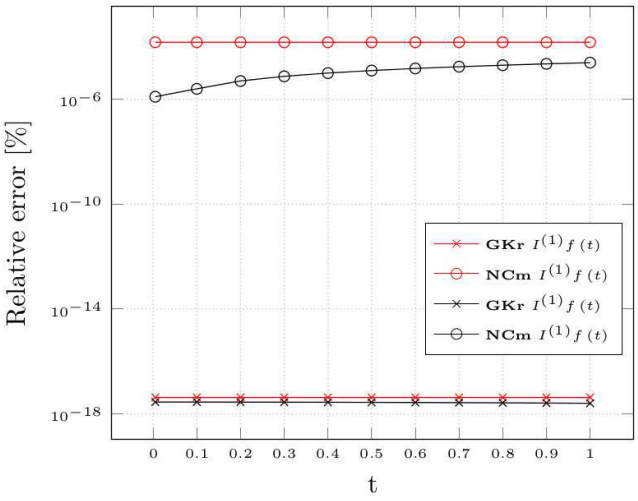


Fig. 2. Integer Order Integration results for function (8) in red and function (9) in black

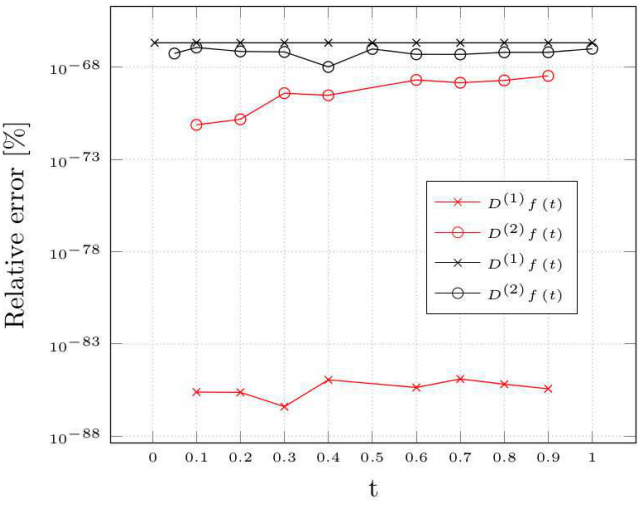


Fig. 3. Integer Order Differentiation results for function (8) in red and function (9) in black

B. Accuracy of integer order integration and differentiation

As it is presented on Fig. 2, GKr ensures enormous advantage over NCm in context of accuracy of integration. GKr achieves triple greater accuracy with only 61 sampling points. NCm was tested with 10000 sampling points. Further sampling points increase in case of GKr did not enable to obtain better results; actually, the accuracy deteriorated. In case of NCm, further accuracy increase was very slow with each 10000 sampling points added, that it did not justified the further experimentation.

Application of GKr is the optimal choice for integration operators concatenation.

Application of 5-point Central Differences to obtain 1^{st} and 2^{nd} derivative resulted with high accuracy, almost error free results (See Fig. 3). Further points increase did not bring any

accuracy increase.

Application of 5-point Central Differences is the optimal choice for differentiation operators concatenation.

C. Accuracy of non-integer order integration and differentiation

As it is presented on Figs 4-7 application of both developed methods of numerical integration GJ and DE to compute directly non-integer order derivatives and integrals enables to obtain highest possible results.

The accuracy of the integer and non-integer orders integration operators concatenation is limited to the accuracy possible to obtain applying integer order numerical integration methods, because it determines the accuracy of the input data for fractional order integration. For differentiation operators concatenation, the deteriorating amount of information about the integrated function during the operators concatenation affects the final accuracy. The loss of information understood in the sense of the input data accuracy decrease: during integer operator application the function is known in the entire range, i.e. it is available in the continuous form; during the fractional operator application the input function is available only in some earlier pre-calculated points, i.e. it is available in the discrete form only.

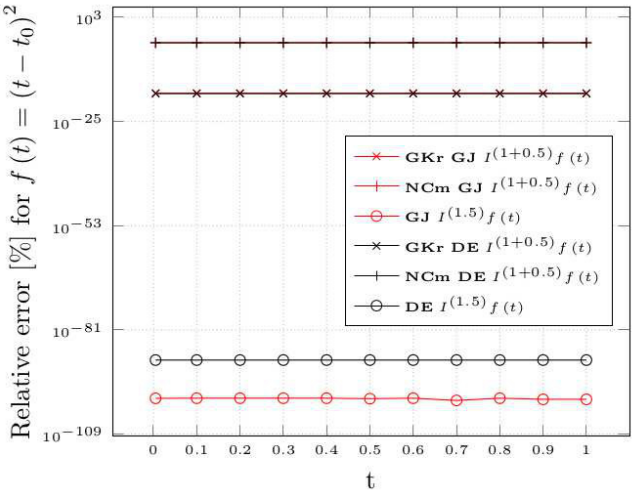


Fig. 4. Non-integer Order Integration Results for function (8). GKr and NCm applied to compute integer orders, GJ and DE to compute fractional and non-integer order

VI. Conclusion

The purpose of the following research was to evaluate in context of computations accuracy two algorithms of non-integer order derivatives and integrals computations: direct numerical calculation of non-integer order derivatives and integrals and the non-integer orders of derivatives and integrals obtained by the application of concatenation of the integer and fractional orders operators of integration and differentiation.

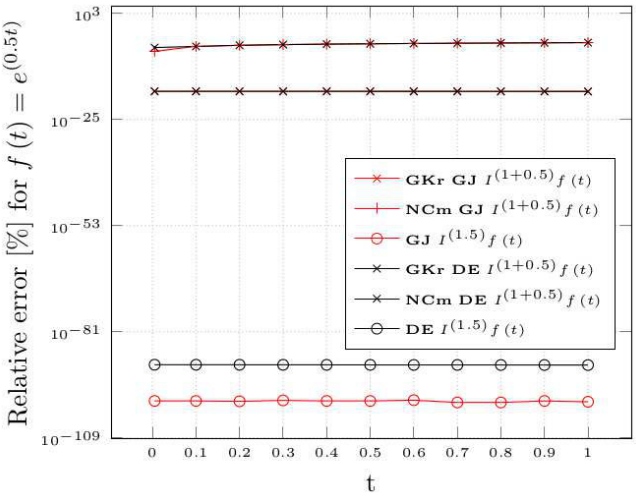


Fig. 5. Non-integer Order Integration Results for function (9) GKr and NCm applied to compute integer orders, GJ and DE to compute fractional and non-integer order

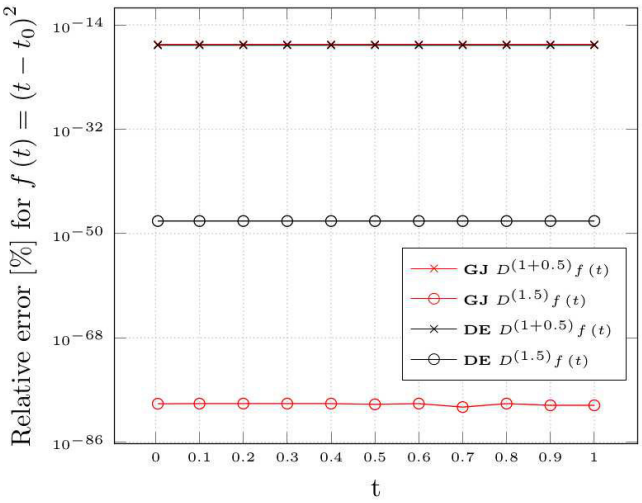


Fig. 6. Non-integer Order Differentiation Results for function (8)

There is no doubt, that direct calculation of non-integer orders derivatives and integrals applying both numerical methods, GJ and DE, developed by the authors of the paper for non-integer and fractional order derivatives and integrals computations applying formulas (1)-(3) are the methods to favorite if one want to obtain high accuracy results. The methods are efficient and their accuracy does not depend on order of the calculated derivative and integral or the integration range.

References

[1] R.L. Burden, J.D. Faires, "Numerical Analysis", 5 th. Ed., Brooks/Cole Cengage Learning, Boston, 2003.

[2] R.A. Krommer, Ch.W. Ueberhuber, "Computational Integration", SIAM, Philadelphia, 1986.

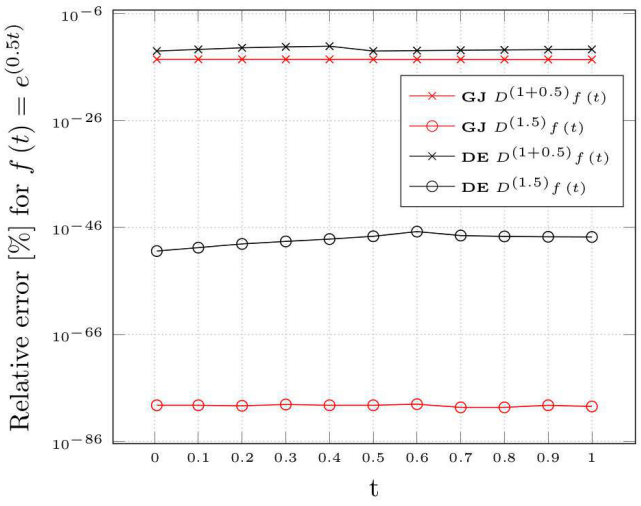


Fig. 7. Non-integer Order Differentiation Results for function (9)

[3] A.H. Stroud,D. Secrest, "Gaussian Quadrature Formulas", Prentice-Hall, Englewood Cliffs, NJ., 1966.

[4] P.K.Kythe, M.R.Schäferkotter, "Handbook of Computational Methods For Integration", Chapman & Hall/CRC, 2005.

[5] M. Hjort-Jensen, "Computational Physics", University of Oslo, 2009.

[6] D.W. Brzeziński and P. Ostalczyk, "High-accuracy Numerical Integration Methods for Fractional Order Derivatives and Integrals Computations", *After reviews, awaits publication in* Bulletin of the Polish Academy of Sciences Technical Sciences, vol. 4, 2014.

[7] D.W. Brzeziński and P. Ostalczyk, "Evaluation of Efficient Methods of Fractional Order Derivatives and Integrals Numerical Calculations", *in Proceedings of the XIV Conference on System Modelling and Control, September 23-24 2013, Łódz, Poland*, 2013.

[8] D. Baleanu, K. Diethelm, E. Scalas, and J.J. Trujillo, "Fractional Calculus Models and Numerical Methods," World Scientific Publishing Co.Pte. Ltd., Singapore, 2012.

[9] K.S. Miller and B. Ross, "An Introduction To The Fractional Calculus and Fractional Differential Equations", John Willey and Sons INC., New York,NY, 1993.

[10] R. Gorenflo and F. Mainardi, "Fractional Calculus: Integral and Differential Equations of Fractional Order", *in Fractals and Fractional Calculus in Continuum Mechanics*, Springer-Verlag, Wien and New York, 1997.

[11] I. Podlubny, Fractional Differential Equations, Academic Press, San Diego, CA, 1999.

[12] K.B. Oldham, J. Spanier, "The Fractional Calculus. Theory and Applications of Differentiation and Integration to Arbitrary Order", Academic Press, 1974.

[13] A.A. Kilbas, H.M. Srivastava, J.J. Trujillo, "Theory and Applications of Fractional Differential Equations", North Holland Mathematics Studies 204, Elsevier, 2006.

[14] P. Ostalczyk, Zarys rachunku różniczkowego i całkowego ułamkowych rzędów, Komitet Automatyki i Robotyki Polskiej Akademii Nauk, Wydawnictwo Politechniki Łódzkiej, Łódź, Poland, 2008.

[15] C. Schwartz, "Numerical Integration of Analytic Functions", *in Journal of Computational Physics*, vol. 4, pp. 19-29, 1969.

[16] H. Takahasi, Quadrature Formulas Obtained by Variable Transformation, Numerische Mathematik, nr 21, 1973.

[17] M. Mori, "Discovery of The Double Exponential Transformation and Its Developments", publ. RIMS, Kyoto Univ., 41, pp. 897-935, 2005.

[18] J.M. Muller, N. Brisebarre, F. De Dinechin, C.P. Jeannerod, V. Lefevre, G. Melquiond, N.Revol, D. Stehle, D. and S. Torres, "Handbook of Floating-Point Arithmetic", Birkhauser Boston, New York,NY, 2010.

[19] K.R. Ghazi, V.Lefevre, P.Theveny and P.Zimmermann, "Why and how to use arbitrary precision" IEEE Computer Society, vol.12, nr 3, 2010, DOI Bookmark: http://doi.ieeecomputersociety.org/10.1109/MCSE.2010.73.

[20] The GNU Multiple Precision Arithmetic Library, https://gmplib.org/.

[21] The GNU Multiple Precision Floating-Point Reliable Library, https://mpfr.org/.

[22] C++ wrapper for the GNU Multiple Precision Floating-Point Reliable Library, http://www.holoborodko.com/pavel/mpfr/.

[23] J. Waldvogel, "Towards A General Error Theory of the Trapezoidal Rule", *in Approximation and Computation*, pp 267-282, Springer Verlag, W.Gautschi, G.Mastroianni and Th.M.Rassias (Eds.), 2011.

APPENDIX

Below there are presented the main ideas behind two new numerical methods of integration developed by the authors of the paper for non-integer (fractional) order derivatives and integrals computations mentioned in the paper.

A. Double Exponential Formula

The Double Exponential (**DE**) formula joins two applied techniques: the double exponential transformation applied to the initial integrand and the trapezoidal rule applied to the transformed integrand.

General idea standing behind the DE transformation which was proposed by Schwartz [15] and become known as the Tanh rule (since $x = tanh(t)$) is as follows:

Let us consider the integral

$$I = \int_a^b f(x)\, dx$$

where $f(x)$ is integrable on interval (a, b). The function $f(x)$ may have singularity $x = a$, $x = b$ or at both.

First we apply the following variable transformation

$$x = \phi(t),\ \phi(-\infty) = a,\ \phi(\infty) = b.$$

We obtain

$$I = \int_{-\infty}^{\infty} f(\phi(t))\, \phi'(t)\, dt.$$

It is important that $\phi(t)$ possess the property such as $\phi'(t)$ decreases its values to 0 at at least double exponential as $t \to \pm\infty$, i.e.

$$\left|\phi'(t)\right| \to exp(-c\, exp(|t|)) \tag{15}$$

where c is some constant.

After that, it is best to apply the trapezoidal formula with an equal mesh size to the transformed integrand expression [23], i.e.

$$I = h \sum_{n=-\infty}^{\infty} f(\phi(nh))\, \phi'(nh)$$

where nh is sampling step.

Due to the property (15) truncation of the summation process can be done at some arbitrary chosen $n = -N_-$ and $n = +N_+$, i.e.

$$I = h \sum_{n=-N_-}^{N_+} f(\phi(nh))\, \phi'(nh), \tag{16}$$

$N = N_- + N_+ + 1$, where N states the amount of sampling points of the function.

Since $\phi'(nh)$ as well as the whole expression $f(\phi(nh))\, \phi'(nh)$ converges to 0 at exponential rate at

large $|n|$, the quadrature formula (16) is called the Double Exponential [16], [17].

Due to truncation of the summation process (16) at some arbitrary chosen $n = N_-,\ n = N_+$, function $f(x)$ can have singularities at $x = a$ and/or $x = b$ as long as it is integrable over the integration range.

There should be taken two kinds of errors into consideration when implementing the DE formula: discretization error, because we use the trapezoidal rule to approximate an integral and truncation error, because we truncate infinite sum at some N.The optimal strategy is to make both errors equal [17].

The subinterval width h, which defines the evaluation step and the number of sample points are key values in such strategy. The source [17] suggest the following value of h for DE formula

$$h \sim \frac{log\,(2\pi N\omega/c)}{N},$$

where c is some constant to be taken, usually 1 or $\pi/2$ and ω is the distance to the nearest singularity of the integrand.

Correct selection of a function (17)-(19) with optimal properties enables to control the level of convergence of the whole transformed expression (16). The rate of convergence has enormous impact on accuracy, i.e. to rapid convergence decreases the accuracy [16].

The authors test three different transformations and selected (18) because of it optimal convergence rate for the purpose of the research, which is also suggested by the literature of the subject [17], [16].

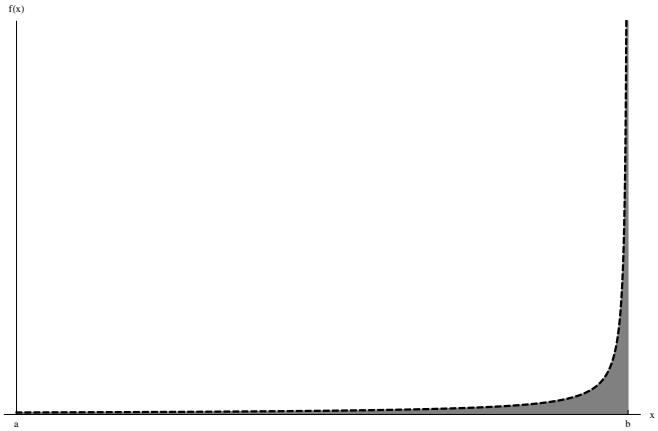


Fig. 8. Graph of the original *core* integrand in the formulas (1)-(3)

The transformations expressions are as follows:

$$x = \phi(t) = \\ = tanh\, t^p,\ \phi^{'}(t) = \frac{pt^{p-1}}{cosh^2 t^p},\ p = 1, 3, 5, \ldots \quad (17)$$

$$x = \phi(t) = \\ = tanh\,(\phi/2\ sinh\,(t)),\ \phi^{'}(t) = \frac{\phi/2\ cosh\ t}{cosh^2 sinh\,(t)} \quad (18)$$

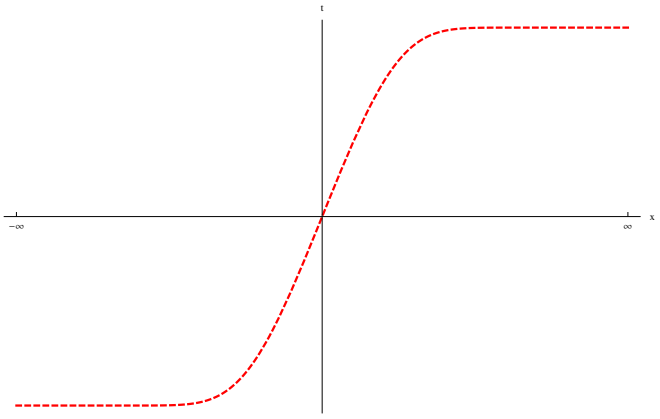


Fig. 9. Graph of the transforming expression (18)

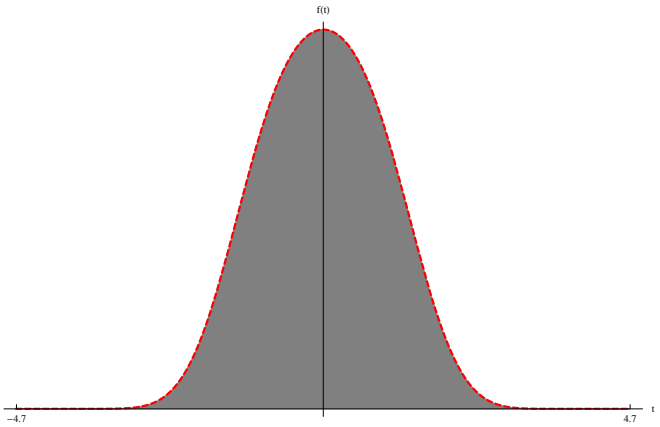


Fig. 10. Graph of the transformed *core* integrand (20) and the range applied in computations

$$x = \phi(t) = tanh\left(\phi/2\ sinh\left(t^3\right)\right),$$
$$\phi^{'}(t) = \frac{3\pi/2\ t^2\ \phi/2\ cosh\ t^3}{cosh^2 sinh\,(t^2)} \quad (19)$$

Applying the transformation (18) to the formulas (1)-(3) according the formula (16), we obtain the following trapezoidal form

$$S = h\sum_{i=1}^{N} f\left(\frac{b-a}{2}\ x_i + \frac{b+a}{2}\right) w_i \quad (20)$$

where

$$x_i = f\,(tanh\,(\pi/2\ sinh\,(t_i)))$$

are the nodes and

$$w_i = \frac{cosh\,(t_i)}{cosh^2\ \pi/2\ sinh\ t_i} \cdot \frac{b-a}{2}$$

are the weights of the Double Exponential Quadrature.

Additionally $t_i = -t_a + (i-1)\cdot h,\ i = 0, 1, 2, 3, \cdots N - 1, h = \frac{2t_a}{N-1}$ are the new integration range and the width of one trapezoidal subinterval. The selection of the value t_a parameter decides how near the singularity we integrate.

There is presented the visualization of the DE transformation on Figs 8-10.

B. Gauss-Jacobi Quadrature with Adopted Weight Function

A weight function, which enables to eliminate definite integration range endpoints singularities is Jacobi weight (21) [1], [2], [3], [4].

$$p(x) = (1-x)^{\lambda}(1+x)^{\beta},\ \lambda,\beta > -1. \tag{21}$$

A quadrature formula with the weight (21) assumes the form

$$\int_{-1}^{1}(1-x)^{\lambda}(1+x)^{\beta}\cdot f(x)\,dx \cong \cong \sum_{k=1}^{n} A_k \cdot f(x_k). \tag{22}$$

The nodes x_i are zeros of the Jacobi polynomial $J_n(x;\lambda,\beta)$.

The Jacobi Polynomial can be determined by applying Rodrigues formula

$$J_n(x;\lambda,\beta) = = \frac{(-1)^n}{2n\cdot n!}(1-x)^{-\lambda}(1+x)^{-\beta}\cdot \cdot \frac{d^n}{dx^n}\left[(1-x)^{\lambda+n}(1+x)^{\beta+n}\right] \tag{23}$$

The weights A_k can be computed applying the following formula

$$A_k = 2^{\lambda+\beta+1}\frac{\Gamma(\lambda+n+1)\Gamma(\beta+n+1)}{n!\Gamma(\lambda+\beta+n+1)}\cdot \cdot \frac{1}{(1-x_k^2)\left[J_n^{(\lambda,\beta)'}x_k\right]^2} \tag{24}$$

The remainder of the Gauss-Jacobi Quadrature is expressed as

$$R = \frac{2^{\lambda+\beta+2n+1}}{\lambda+\beta+2n+1}\cdot \cdot \frac{\Gamma(\lambda+n+1)\Gamma(\beta+n+1)\Gamma(\lambda+\beta+n+1)}{\Gamma^2(\lambda+\beta+2n+1)}\cdot \cdot \frac{n!}{2n}\cdot f^{2n}(\xi), \xi\in\langle -1,1\rangle \tag{25}$$

Now the transformation of the weight function. Substituting $\lambda = 1-\alpha, \beta = 0$ in (22) we obtain

$$\int_{-1}^{1}\frac{\phi(x)}{(1-x)^{1-\alpha}}dx \tag{26}$$

which coincides with *the core* integrand in the formulas (1)–(3).

To change the integration range from $[-1,1]$ to arbitrary chosen $[t_0,t]$ formula (26) must be transformed as follows

$$\left(\frac{t-t_0}{2}\right)^{\nu}\cdot\int_{-1}^{1}\frac{\phi(u)}{(1-u)^{1-\alpha}}du \tag{27}$$

where

$$\phi(u) = f\left(\left(\frac{t-t_0}{2}\right)u + \left(\frac{t-t_0}{2}\right)\right)$$

Applying formulas (26)-(27) we can express the formula (1) as

$${}^{RL}\mathrm{I}^{(\nu)} = \frac{1}{\Gamma(\nu)}\left(\frac{t-t_0}{2}\right)^{\nu}\int_{t_0}^{t}\frac{f(u)}{(t-u)^{1-\alpha}}du.$$

To calculate non-integer (fractional) order derivatives applying formula (3) we proceed the similar way

$$\left(\frac{t-t_0}{2}\right)^{n-\nu}\int_{-1}^{1}\frac{\phi(u)}{(1-u)^{1-\alpha}}du$$

where

$$\phi(u) = f\left(\left(\frac{t-t_0}{2}\right)u + \left(\frac{t-t_0}{2}\right)\right).$$

The formula (3) assumes the following form

$${}^{C}\mathrm{D}^{(\nu)} = \frac{1}{\Gamma(n-\nu)}\left(\frac{t-t_0}{2}\right)^{n-\nu}\int_{t_0}^{t}\frac{\left(\frac{d}{dt}\right)^n}{(t-u)^{n-\alpha-1}}du. \tag{28}$$

The formula (28) seems to have similar form as (3). The most difficult part in context of numerical integration, however is calculated applying a method which guarantees multiple times higher accuracy, applying the Jacobi polynomials.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 561–568
DOI: 10.15439/2014F315
ACSIS, Vol. 2

The WZ factorization in MATLAB

Beata Bylina, Jarosław Bylina
Marie Curie-Skłodowska University,
Institute of Mathematics,
Pl. M. Curie-Skłodowskiej 5,
20-031 Lublin, Poland
Email: {beata.bylina,jaroslaw.bylina}@umcs.pl

***Abstract*—In the paper the authors present the WZ factorization in MATLAB. MATLAB is an environment for matrix computations, therefore in the paper there are presented both the sequential WZ factorization and a block-wise version of the WZ factorization (called here VWZ). Both the algorithms were implemented and their performance was investigated. For random dense square matrices with the dominant diagonal we report the execution time of the WZ factorization in MATLAB and we investigate the accuracy of such solutions. Additionally, the results (time and accuracy) for our WZ implementations were compared to the similar ones based on the LU factorization.**

Keywords: linear system, WZ factorization, LU factorization, matrix factorization, matrix computations

I. Introduction

IN THE international academic circles MATLAB is accepted as a reliable and convinient software for numerical computations. Particularly, it is used for linear algebra computations. Nowadays, there are a lot of papers devoted to the use of MATLAB in mathematics (linear systems [7], least-squares problems [9]; function approximation [12]; eigenvalues [2], [11] — and many others). In this paper we use MATLAB to solve linear systems.

Solution of linear systems of the form:

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \qquad \text{where} \quad \mathbf{A} \in \mathbb{R}^{n\times n}, \quad \mathbf{b} \in \mathbb{R}^{n}, \qquad (1)$$

is an important and common problem in engineering and scientific computations. One of the direct methods of solving a dense linear system (1) is to factorize the matrix $\mathbf{A}$ into some simpler matrices — it is its decomposition into factor matrices (that is, factorization) of a simpler structure — and then solving simpler linear systems. The most known factorization is the LU factorization. MATLAB provides many ways to solve linear systems, one of them is based on the LU factorization: `[L,U]=lu(A)`. This method is powerful and simple to use.

In [7] an object-oriented method is presented, which is is a meta-algorithm that selects the best factorization method for a particular matrix, whether sparse or dense — allowing the reuse of its factorization for subsequent systems.

In this work we study another form of the factorization, namely the WZ factorization and investigate both the accuracy of the computations and their time. In [4], [5] we showed that there are matrices for which applying the incomplete WZ preconditioning gives better results than the incomplete LU factorization.

The aim of the paper is to analyze the potential of implementations of the WZ factorization in a high-level language (as it is the case of MATLAB). We implement the WZ factorization and compare its performance to a MATLAB function implementing the LU factorization, namely: `[L,U]=lu(A)` — and to the authors' own MATLAB implementation of the LU factorization.

The content of the paper is following. In Section II we describe the idea of the WZ factorization [8], [13] and the way the matrix $\mathbf{A}$ is factorized to a product of matrices $\mathbf{W}$ and $\mathbf{Z}$ — such a factorization exists for every nonsingular matrix (with pivoting) what was shown in [8]. Section III provides information about some modifications of the original WZ algorithm — in a way to decrease the number of loops and to make as much as possible computations in blocks — and this will allow us to use MATLAB efficiently. In Section IV we present the results of our experiments. We analyzed the time of WZ factorization. We study the influence of the size of the matrix on the achieved numerical accuracy. We compare the WZ factorization to the LU factorization. Section V is a summary of our experiments.

II. WZ Factorization (WZ)

Here we describe shortly the WZ factorization usage to solve (1). The WZ factorization is described in [8], [10]. Let us assume that the $\mathbf{A}$ is a square nonsingular matrix of an even size (it is somewhat easier to obtain formulas for even sizes than for odd ones). We are to find matrices $\mathbf{W}$ and $\mathbf{Z}$ that fulfill $\mathbf{WZ} = \mathbf{A}$ and the matrices $\mathbf{W}$ and $\mathbf{Z}$ consist of the rows $\mathbf{w}_i^T$ and $\mathbf{z}_i^T$ shown in Figure 1, respectively.

After the factorization we can solve two linear systems:

$$\mathbf{W}\mathbf{y} = \mathbf{b},$$

$$\mathbf{Z}\mathbf{x} = \mathbf{y}$$

(where $\mathbf{c}$ is an auxiliary intermediate vector) instead of one (1).

Figure 2 shows an example of a matrix nad its WZ factors.

In this paper we are interested only in obtaining the matrices $\mathbf{Z}$ and $\mathbf{W}$. The first part of the algorithm consists in setting

This work was partially supported within the project N N516 479640 of the Ministry of Science and Higher Education of the Polish Republic (MNiSW) "Modele dynamiki transmisji, sterowania, zatłoczeniem i jakością usług w Internecie".

$$
\begin{aligned}
\mathbf{w}_1^T &= (1, \underbrace{0, \ldots, 0}_{n-1}) \\
\mathbf{w}_i^T &= (w_{i1}, \ldots, w_{i,i-1}, 1, \underbrace{0, \ldots, 0}_{n-2i+1}, w_{i,n-i+2}, \ldots, w_{in}) \qquad \text{for} \quad i = 2, \ldots, \tfrac{n}{2}, \\
\mathbf{w}_i^T &= (w_{i1}, \ldots, w_{i,n-i}, \underbrace{0, \ldots, 0}_{2i-n-1}, 1, w_{i,i+1}, \ldots, w_{in}) \qquad \text{for} \quad i = \tfrac{n}{2}+1, \ldots, n-1, \\
\mathbf{w}_n^T &= (\underbrace{0, \ldots, 0}_{n-1}, 1) \\
\mathbf{z}_i^T &= (\underbrace{0, \ldots, 0}_{i-1}, z_{ii}, \ldots, z_{i,n-i+1}, 0, \ldots, 0) \qquad \text{for} \quad i = 1, \ldots, \tfrac{n}{2}, \\
\mathbf{z}_i^T &= (\underbrace{0, \ldots, 0}_{n-i}, z_{i,n-i+1}, \ldots, z_{ii}, 0, \ldots, 0) \qquad \text{for} \quad i = \tfrac{n}{2}+1, \ldots, n.
\end{aligned}
$$

Fig. 1. Rows of the matrices $\mathbf{W}$ and $\mathbf{Z}$

$$
\mathbf{A} = \begin{bmatrix} 2 & 1 & 3 & -6 & 3 & 3 \\ 10 & 6 & 9 & -13 & 10 & 14 \\ 12 & 13 & 12 & -13 & 19 & 17 \\ 8 & 10 & 11 & -4 & 12 & 11 \\ 12 & 8 & 13 & -20 & 14 & 17 \\ 3 & 1 & 1 & -1 & 1 & 4 \end{bmatrix}
$$

$$
\mathbf{W} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 0 & 2 \\ 3 & 1 & 1 & 0 & 2 & 2 \\ 1 & 2 & 0 & 1 & 1 & 2 \\ 3 & 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{Z} = \begin{bmatrix} 2 & 1 & 3 & -6 & 3 & 3 \\ 0 & 2 & 1 & 1 & 2 & 0 \\ 0 & 0 & -4 & 6 & 0 & 0 \\ 0 & 0 & 2 & 2 & 0 & 0 \\ 0 & 3 & 2 & 0 & 3 & 0 \\ 3 & 1 & 1 & -1 & 1 & 4 \end{bmatrix}
$$

Fig. 2. A matrix $\mathbf{A}$ and its factors $\mathbf{W}$ and $\mathbf{Z}$

succesive parts of columns of the matrix $\mathbf{A}$ to zeros. In the first step we do that with the elements in columns 1st and nth — from the 2nd row to the $n-1$st row. Next we update the matrix $\mathbf{A}$.

More formally we can describe the first step of the algoritm the following way.

1) For every $i = 2, \ldots, n-1$ we compute w_{i1} and w_{in} from the system:

$$
\begin{cases} a_{11}w_{i1} + a_{n1}w_{in} = -a_{i1} \\ a_{1n}w_{i1} + a_{nn}w_{in} = -a_{in} \end{cases}
$$

and we put them in a matrix of the form:

$$
\mathbf{W}^{(1)} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ w_{21} & 1 & \ddots & \vdots & w_{2n} \\ \vdots & 0 & \ddots & 0 & \vdots \\ w_{n-1,1} & \vdots & \ddots & 1 & w_{n-1,n} \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.
$$

2) We compute:

$$
\mathbf{A}^{(1)} = \mathbf{W}^{(1)}\mathbf{A}.
$$

After the first step we get a matrix of the form:

$$
\mathbf{A}^{(1)} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1,n-1} & a_{1n} \\ 0 & a_{22}^{(1)} & \cdots & a_{2,n-1}^{(1)} & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & a_{n-1,2}^{(1)} & \cdots & a_{n-1,n-1}^{(1)} & 0 \\ a_{n1} & a_{n2} & \cdots & a_{n,n-1} & a_{nn} \end{bmatrix},
$$

where (for $i, j = 2, \ldots, n-1$):

$$
a_{ij}^{(1)} = a_{ij} + w_{i1}a_{1j} + w_{in}a_{nj}.
$$

Then, we proceed analogously — but for the inner square matrices — $\mathbf{A}^{(1)}$ of size $n-2$ and so on.

So, the whole algorithm is following.

For $k = 1, 2, \ldots, \frac{n}{2} - 1$:

1) For every $i = k+1, \ldots, n-k$ we compute w_{ik} and $w_{i,n-k+1}$ from the system:

$$\begin{cases} a_{kk}^{(k-1)} w_{ik} & + & a_{n-k+1,k}^{(k-1)} w_{i,n-k+1} \\ & = & -a_{ik}^{(k-1)} \\ a_{k,n-k+1}^{(k-1)} w_{ik} & + & a_{n-k+1,n-k+1}^{(k-1)} w_{i,n-k+1} \\ & = & -a_{i,n-k+1}^{(k-1)} \end{cases}$$

and we put them in a matrix of the form shown in Figure 3.

2) We compute:

$$\mathbf{A}^{(k)} = \mathbf{W}^{(k)} \mathbf{A}^{(k-1)}$$

(where obviously $\mathbf{A}^{(0)} = \mathbf{A}$).

After $\frac{n}{2} - 1$ such steps we get the matrix

$$\mathbf{Z} = \mathbf{A}^{(\frac{n}{2}-1)}$$

.

Moreover, we know that:

$$\mathbf{W}^{(\frac{n}{2}-1)} \cdot \ldots \cdot \mathbf{W}^{(1)} \cdot \mathbf{A} = \mathbf{Z}$$

so we get:

$$\mathbf{A} = \left(\mathbf{W}^{(1)}\right)^{-1} \cdot \ldots \cdot \left(\mathbf{W}^{(\frac{n}{2}-1)}\right)^{-1} \cdot \mathbf{Z} = \mathbf{W}\mathbf{Z}$$

Figure 4 shows the same algorithm implemented in MATLAB.

III. Vector WZ Factorization (VWZ)

Now we describe a matrix-vector algorithm for the WZ factorization of the matrix $\mathbf{A}$ which is originally presented in [6].

This is a sequential algorithm where we grouped and ordered the scalar operations anew, into matrix-vector operations. We are showing the algorithm without pivoting, working only for matrices for which such a WZ factorization is executable (for example, for matrices with a dominant diagonal.

Let us write the matrices $\mathbf{A}$, $\mathbf{W}$, $\mathbf{Z}$ as block matrices. We can get equations presented in Figure 5.

In Figure 5, $\widehat{\mathbf{W}}$ and $\widehat{\mathbf{Z}}$ are square matrices of the same structure as the matrices $\mathbf{W}$ and $\mathbf{Z}$, respectively; $\widehat{\mathbf{A}}$ is a full square matrix; $\widehat{\mathbf{W}}$, $\widehat{\mathbf{Z}}$ and $\widehat{\mathbf{A}}$ are of the size 2 less than the size of $\mathbf{W}$, $\mathbf{Z}$ and $\mathbf{A}$; vectors $\mathbf{a}_{1*}^T$, $\mathbf{a}_{n*}^T$, $\mathbf{z}_{1*}^T$, $\mathbf{z}_{n*}^T$ are row vectors; vectors $\mathbf{a}_{*1}$, $\mathbf{a}_{*n}$, $\mathbf{w}_{*1}$, $\mathbf{w}_{*n}$ are column vectors.

From the comparison of the corresponding elements in Figure 5 we get:

$$\begin{array}{ll} a_{11} = z_{11}; & a_{1n} = z_{1n}; \\ a_{n1} = z_{n1}; & a_{nn} = z_{nn} \end{array}$$

$$\mathbf{a}_{1*}^T = \mathbf{z}_{1*}^T; \quad \mathbf{a}_{n*}^T = \mathbf{z}_{n*}^T$$

$$\begin{cases} \mathbf{a}_{*1} & = & \mathbf{w}_{*1} z_{11} & + & \mathbf{w}_{*n} z_{n1} \\ \mathbf{a}_{*n} & = & \mathbf{w}_{*1} z_{1n} & + & \mathbf{w}_{*n} z_{nn} \end{cases} \tag{2}$$

$$\widehat{\mathbf{A}} = \mathbf{w}_{*1}\mathbf{z}_{1*}^T + \widehat{\mathbf{W}}\widehat{\mathbf{Z}} + \mathbf{w}_{*n}\mathbf{z}_{n*}^T$$

From (2) we can describe our algorithm for finding $\mathbf{W}$ and $\mathbf{Z}$ as following:

1) let the first row of the matrix $\mathbf{Z}$ be the first row of the matrix $\mathbf{A}$;
2) let the last row of the matrix $\mathbf{Z}$ be the last row of the matrix $\mathbf{A}$;
3) compute the vectors $\mathbf{w}_{*1}$ and $\mathbf{w}_{*n}$ from:

$$\begin{aligned} \mathbf{w}_{*1} &= \alpha \mathbf{a}_{*n} - \beta \mathbf{a}_{*1}, \\ \mathbf{w}_{*n} &= \gamma \mathbf{a}_{*1} - \delta \mathbf{a}_{*n}, \end{aligned}$$

where

$$\begin{aligned} \alpha &= \frac{z_{n1}}{z_{1n}z_{n1} - z_{11}z_{nn}}, \\ \beta &= \frac{z_{nn}}{z_{1n}z_{n1} - z_{11}z_{nn}}, \\ \gamma &= \frac{z_{1n}}{z_{1n}z_{n1} - z_{11}z_{nn}}, \\ \delta &= \frac{z_{11}}{z_{1n}z_{n1} - z_{11}z_{nn}}; \end{aligned}$$

4) update the inner part of the matrix $\mathbf{A}$ (the matrix without its first and last row and column):

$$\widehat{\mathbf{A}}^{\text{new}} = \widehat{\mathbf{W}}\widehat{\mathbf{Z}} = \widehat{\mathbf{A}} - \mathbf{w}_{*1}\mathbf{z}_{1*}^T - \mathbf{w}_{*n}\mathbf{z}_{n*}^T;$$

5) if the size of the matrix $\widehat{\mathbf{A}}^{\text{new}}$ is 3 or more, then start over from 1., but with $\mathbf{A} = \widehat{\mathbf{A}}^{\text{new}}$, $\mathbf{W} = \widehat{\mathbf{W}}$ and $\mathbf{Z} = \widehat{\mathbf{Z}}$ (so all three matrices become smaller and smaller and the algorithm comes eventually to the end).

Figure 6 shows this algorithm implemented in MATLAB.

IV. Numerical Experiments

In this section we tested the time and the absolute accuracy of the WZ factorization. Our intention was to investigate the WZ factorization and compare our implementation (done in MATLAB, which is a high-level language) of the WZ factorization with the LU factorization.

The input matrices are generated (by the authors). They are random matrices with a dominant diagonal of even sizes (500, 1000, 1500 and so on, up to 3000 or 4000).

The MATLAB implementation was compiled and tested under MS Windows on workstations with an AMD processor and an Intel processor. Table I shows details of specification of the hardware used in the experiment. To measure the performance time standard MATLAB functions were used — `tic` and `toc`. We measured the difference between $\mathbf{A}$ and $\mathbf{WZ}$ by absolute error $||\mathbf{A} - \mathbf{WZ}||_2$ — to compute the norm there was used a standard MTLAB function: `norm`. The implementations were tested in MATLAB R2008 and MATLAB R2010.

Figures 8, 9, 10 and 11 show the performance time (in seconds) of the WZ factorization on the AMD processor and the Intel processor, implemented in MATLAB R2008 and MATLAB R2010. Additionally, we compared the WZ

$$\mathbf{W}^{(k)} = \begin{bmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & w_{k+1,k} & \ddots & & w_{k+1,n-k+1} & & \\ & & \vdots & & \ddots & \vdots & & \\ & & w_{n-k,k} & & & w_{n-k,n-k+1} & & \\ & & & & & 1 & & \\ & & & & & & \ddots & \\ & & & & & & & 1 \end{bmatrix}$$

Fig. 3. The form of the matrix $\mathbf{W}^{(k)}$

```
% steps of elimination — from A to Z
for k = 0 : n/2-2
  k2 = n-k-1;
  det = A(k,k)*A(k2,k2)-A(k2,k)*A(k,k2);
% finding elements of W
  for i = k+1 : k2-1
    w(i,k) = (A(k2,k)*A(i,k2)-A(k2,k2)*A(i,k))/det;
    w(i,k2) = (A(k,k2)*A(i,k)-A(k,k)*A(i,k2))/det;
% updating A
    for j = k+1 : k2-1
      A(i,j) = A(i,j)+wk1*A(k,j)+wk2*A(k2,j);
```

Fig. 4. The sequential implementation of the WZ factorization for solving linear systems

.

$$\mathbf{A} = \begin{bmatrix} a_{11} & \mathbf{a}_{1*}^T & a_{1n} \\ \mathbf{a}_{*1} & \widehat{\mathbf{A}} & \mathbf{a}_{*n} \\ a_{n1} & \mathbf{a}_{n*}^T & a_{nn} \end{bmatrix} = \mathbf{WZ} = \begin{bmatrix} 1 & 0 & 0 \\ \mathbf{w}_{*1} & \widehat{\mathbf{W}} & \mathbf{w}_{*n} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} z_{11} & \mathbf{z}_{1*}^T & z_{1n} \\ 0 & \widehat{\mathbf{Z}} & 0 \\ z_{n1} & \mathbf{z}_{n*}^T & z_{nn} \end{bmatrix} =$$

$$= \begin{bmatrix} z_{11} & \mathbf{z}_{1*}^T & z_{1n} \\ \mathbf{w}_{*1}z_{11} + \mathbf{w}_{*n}z_{n1} & \mathbf{w}_{*1}\mathbf{z}_{1*}^T + \widehat{\mathbf{W}}\widehat{\mathbf{Z}} + \mathbf{w}_{*n}\mathbf{z}_{n*}^T & \mathbf{w}_{*1}z_{1n} + \mathbf{w}_{*n}z_{nn} \\ z_{n1} & \mathbf{z}_{n*}^T & z_{nn} \end{bmatrix}$$

Fig. 5. The WZ factorization written as blocks of vector

```
% steps of elimination — from A to Z
for k = 0 : n/2-2
  k2 = n-k-1;
  det =A(k,k)*A(k2,k2)-A(k2,k)*A(k,k2);
% finding elements of W
    W(k+1:k2-1,k) = (A(k2,k)*A(k+1:k2-1,k2)-A(k2,k2)*A(k+1:k2-1,k))/det;
    W(k+1:k2-1,k2) = (A(k,k2)*A(k+1:k2-1,k)-A(k,k)*A(k+1:k2-1,k2))/det;
% updating A
    A(k+1:k2-1,k+1:k2-1) = A(k+1:k2-1,k+1:k2-1)+ W(k+1:k2-1,k)*A(k,k+1:k2-1)
                                               +W(k+1:k2-1,k2)*A(k2,k+1:k2-1);
```

Fig. 6. The vector implementation of the WZ factorization for solving linear systems

```
% steps of elimination — from A to U
for k = 0 : n-2
% finding elements of L
  for i = k+1 : n-1
    l(i,k) =-a(i,k)/a(k,k);
% updating A
    for j = k+1 : n
      A(i,j) = A(i,j) + l(i,k)*A(k,j);
```

Fig. 7. The sequential implementation of the LU factorization for solving linear systems

TABLE I
HARDWARE USED IN EXPERIMENTS

#	CPU	Memory
1	AMD FX-8120 3.1 GHz	16 GB
2	Intel Core 2 Duo 2.53 GHz	4 GB

factorization (both in its sequential — Figure 4 — and vector — Figure 6 — versions) with a standard MATLAB LU factorization function, namely `lu`, which uses the subroutine DGETRF from LAPACK [1], and also with the simple LU implementation (shown in Figure 7).

Results show that the vector WZ factorization (VWZ) is much faster than the sequential WZ factorization in both tested MATLAB versions and on both architectures.

However, on the older processor (Intel Core is here the case) the sequentials algorithms perform better than on the newer (AMD) — and the block algorithms (VWA and the standard MATLAB function `lu`) perform better on the newer one. It is caused by the differences in architectues — newer ones prefer block algorithms because of their stronger inner parallelism.

Tables II, III, IV and V illustrate the accuracy (given as the norms $||\mathbf{A} - \mathbf{WZ}||_2$ and $||\mathbf{A} - \mathbf{LU}||_2$) of the WZ and LU factorizations in MATLAB. The first column shows the norm for the sequential WZ factorization (from Figure 4); the second — the vector WZ factorization (VWZ, from Figure 6); the third presents the norm for the sequential LU factorization (from Figure 7); the fourth — the norm for the standard MATLAB function `lu`.

Based on the results, we can state that different implementations give quite similar accuracies. However, the sizes of the matrix influences the accuracy (it worsens when the size grows).

Tables VI, VII, VIII, IX, illustrate the speedup for the VWZ and LU factorizations in MATLAB (both R2008 and R2010) relative to the sequential WZ factorization. The first column shows the speedup of VWZ, the second — the speedup of the LU factorization and the third — the speed of the standard MATLAB function `lu` — all relative to the sequential WZ factorization.

Based on these results, we can conclude that various implementations of the WZ factorization give different performance. Namely, VWZ is even about 4 times faster then the sequential WZ (on the AMD processor; on the Intel processor the speedup is only about 2). The LU factorization implemented by the authors is the slowest of all the tested implementations. However, the standard MATLAB function `lu` is the fastest — this function implements a block LU factorization, which makes the processor architecture is better utilized.

V. CONCLUSION

In this paper we did some performance analysis of a MATLAB implementations of the WZ factorization. We examined a sequential implementation of the WZ factorization. We also implemented in MATLAB a vector version of the WZ factorization (VWZ) — to avoid loops. We compared these implementations with two versions of the LU factorization — our MATLAB implementation and a standard MATLAB function `[L,U]=lu(A)`.

From the results we can conclude that the reduction of the number of nested loops in the original WZ factorization increased the speed even four times. The sequential WZ factorization is faster than the sequential LU factorization. Of course, the fastest of the implementation is the built-in MATLAB function `lu` — which utilizes LAPACK block factorization [1].

The implementation and the architecture had no impact on the accuracy of the factorization — the accuracy depended only on the size of the matrix what is quite self-evident.

The version of MATLAB has no significant influence on neither the performance time nor the speedup — only the architecture and the size of the matrix count.

VI. FUTURE WORK

To accelerate the WZ factorization, it would be desirable to build a block algorithm for the WZ factorization and to utilize parallelism — especially for the machines with many processing units.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, D. Sorensen, *LAPACK Users' Guide (Third ed.)*, SIAM, Philadelphia 1999.

[2] T. Betcke, N. J. Higham, V. Mehrmann,Ch. Schröder, F. Tisseur, NLEVP: A Collection of Nonlinear Eigenvalue Problems, *ACM Trans. Math. Softw.*, Volume 39 Issue 2, February 2013, Article No. 7.

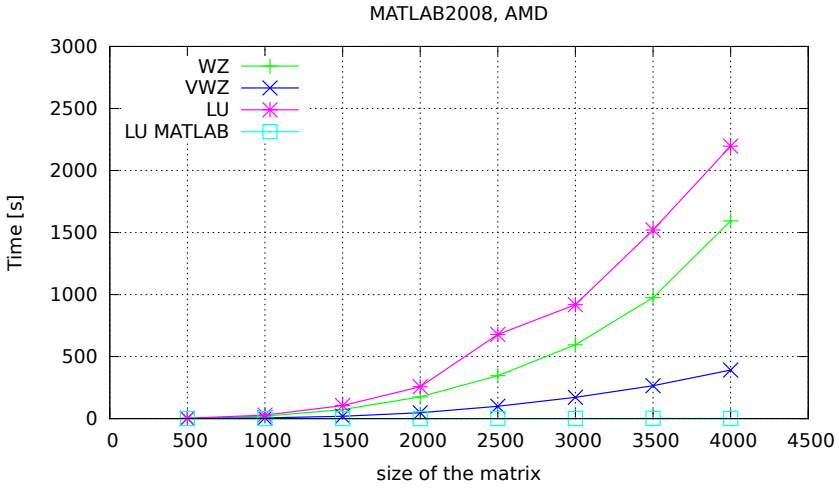


Fig. 8. The WZ factorization performance time (in seconds) on the AMD processor, in MATLAB R2008

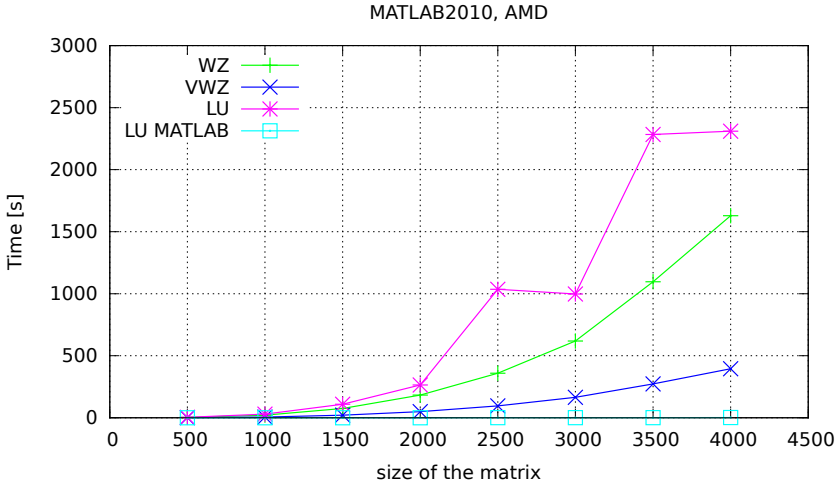


Fig. 9. The WZ factorization performance time (in seconds) on the AMD processor, in MATLAB R2010

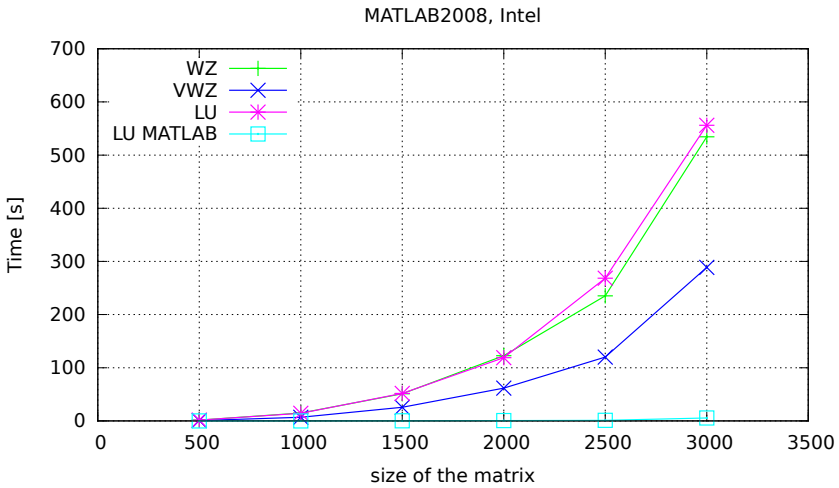


Fig. 10. The WZ factorization performance time (in seconds) on the Intel processor, in MATLAB R2008

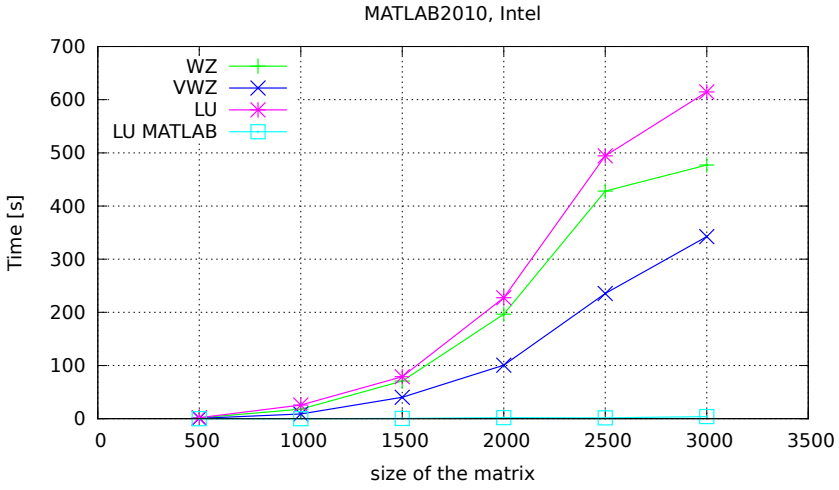


Fig. 11. The WZ factorization performance time (in seconds) on the Intel processor, in MATLAB R2010

TABLE II
THE NORMS FOR THE WZ AND LU FACTORIZATIONS IN MATLAB R2008 ON THE AMD PROCESSOR

matrix size	WZ	VWZ	LU	`lu`
500	$1.08 \cdot 10^{-12}$	$4.00 \cdot 10^{-13}$	$8.53 \cdot 10^{-13}$	$5.68 \cdot 10^{-13}$
1000	$2.84 \cdot 10^{-12}$	$1.02 \cdot 10^{-12}$	$2.61 \cdot 10^{-12}$	$7.96 \cdot 10^{-13}$
1500	$8.18 \cdot 10^{-12}$	$3.86 \cdot 10^{-12}$	$7.50 \cdot 10^{-12}$	$5.00 \cdot 10^{-12}$
2000	$7.96 \cdot 10^{-12}$	$2.73 \cdot 10^{-12}$	$7.73 \cdot 10^{-12}$	$2.95 \cdot 10^{-12}$
2500	$2.50 \cdot 10^{-11}$	$5.91 \cdot 10^{-12}$	$2.09 \cdot 10^{-11}$	$3.18 \cdot 10^{-12}$
3000	$2.51 \cdot 10^{-11}$	$7.28 \cdot 10^{-12}$	$2.09 \cdot 10^{-11}$	$2.09 \cdot 10^{-11}$
3500	$2.13 \cdot 10^{-11}$	$7.73 \cdot 10^{-12}$	$2.50 \cdot 10^{-11}$	$5.91 \cdot 10^{-12}$
4000	$2.54 \cdot 10^{-11}$	$9.09 \cdot 10^{-12}$	$2.64 \cdot 10^{-11}$	$4.55 \cdot 10^{-12}$

TABLE III
THE NORMS FOR THE WZ AND LU FACTORIZATIONS IN MATLAB R2010 ON THE AMD PROCESSOR

matrix size	WZ	VWZ	LU	`lu`
500	$9.08 \cdot 10^{-13}$	$4.00 \cdot 10^{-13}$	$8.53 \cdot 10^{-13}$	$5.68 \cdot 10^{-13}$
1000	$2.84 \cdot 10^{-12}$	$1.02 \cdot 10^{-12}$	$2.61 \cdot 10^{-12}$	$7.96 \cdot 10^{-13}$
1500	$7.27 \cdot 10^{-12}$	$3.86 \cdot 10^{-12}$	$7.50 \cdot 10^{-12}$	$5.00 \cdot 10^{-12}$
2000	$7.96 \cdot 10^{-12}$	$2.73 \cdot 10^{-12}$	$7.73 \cdot 10^{-12}$	$2.95 \cdot 10^{-12}$
2500	$2.09 \cdot 10^{-11}$	$5.91 \cdot 10^{-12}$	$2.09 \cdot 10^{-11}$	$3.18 \cdot 10^{-12}$
3000	$2.32 \cdot 10^{-11}$	$7.28 \cdot 10^{-12}$	$2.09 \cdot 10^{-11}$	$2.09 \cdot 10^{-11}$
3500	$2.58 \cdot 10^{-11}$	$7.30 \cdot 10^{-12}$	$2.50 \cdot 10^{-11}$	$5.91 \cdot 10^{-12}$
4000	$2.45 \cdot 10^{-11}$	$9.09 \cdot 10^{-12}$	$2.64 \cdot 10^{-11}$	$4.55 \cdot 10^{-12}$

TABLE IV
THE NORMS FOR THE WZ AND LU FACTORIZATIONS IN MATLAB R2008 ON THE INTEL PROCESSOR

matrix size	WZ	VWZ	LU	`lu`
500	$9.09 \cdot 10^{-13}$	$9.09 \cdot 10^{-13}$	$8.52 \cdot 10^{-13}$	$5.68 \cdot 10^{-13}$
1000	$2.84 \cdot 10^{-12}$	$2.84 \cdot 10^{-12}$	$2.61 \cdot 10^{-12}$	$6.82 \cdot 10^{-13}$
1500	$7.27 \cdot 10^{-12}$	$7.27 \cdot 10^{-12}$	$7.73 \cdot 10^{-12}$	$5.00 \cdot 10^{-12}$
2000	$8.40 \cdot 10^{-12}$	$8.40 \cdot 10^{-12}$	$7.95 \cdot 10^{-12}$	$1.82 \cdot 10^{-12}$
2500	$2.09 \cdot 10^{-11}$	$2.09 \cdot 10^{-12}$	$2.09 \cdot 10^{-11}$	$1.36 \cdot 10^{-12}$
3000	$2.27 \cdot 10^{-11}$	$2.27 \cdot 10^{-12}$	$2.09 \cdot 10^{-11}$	$3.63 \cdot 10^{-11}$

TABLE V
THE NORMS FOR THE WZ AND LU FACTORIZATIONS IN MATLAB R2010 ON THE INTEL PROCESSOR

matrix size	WZ	VWZ	LU	`lu`
500	$6.83 \cdot 10^{-13}$	$6.83 \cdot 10^{-13}$	$1.19 \cdot 10^{-12}$	$3.98 \cdot 10^{-13}$
1000	$2.39 \cdot 10^{-12}$	$2.39 \cdot 10^{-12}$	$2.50 \cdot 10^{-12}$	$7.96 \cdot 10^{-13}$
1500	$7.96 \cdot 10^{-12}$	$7.96 \cdot 10^{-12}$	$8.18 \cdot 10^{-12}$	$1.59 \cdot 10^{-12}$
2000	$9.78 \cdot 10^{-12}$	$9.78 \cdot 10^{-12}$	$1.00 \cdot 10^{-11}$	$1.82 \cdot 10^{-12}$
2500	$2.18 \cdot 10^{-11}$	$2.18 \cdot 10^{-11}$	$2.36 \cdot 10^{-11}$	$3.64 \cdot 10^{-12}$
3000	$2.36 \cdot 10^{-11}$	$2.36 \cdot 10^{-11}$	$2.41 \cdot 10^{-11}$	$4.55 \cdot 10^{-12}$

TABLE VI
THE SPEEDUP OF VWZ, OF THE LU FACTORIZATION AND OF THE STANDARD MATLAB FUNCTION `LU` — RELATIVE TO THE SEQUENTIAL WZ FACTORIZATION (MATLAB R2008 ON THE AMD PROCESSOR)

matrix size	VWZ	LU	`lu`
500	3.63	0.73	225.00
1000	3.73	0.72	288.57
1500	3.78	0.68	426.41
2000	3.68	0.68	486.44
2500	3.47	0.51	628.85
3000	3.47	0.64	646.47
3500	3.63	0.64	601.82
4000	4.07	0.72	870.95

TABLE VII
THE SPEEDUP OF VWZ, OF THE LU FACTORIZATION AND OF THE STANDARD MATLAB FUNCTION `LU` — RELATIVE TO THE SEQUENTIAL WZ FACTORIZATION (MATLAB R2010 ON THE AMD PROCESSOR)

matrix size	VWZ	LU	`lu`
500	4.09	0.70	237.00
1000	3.69	0.71	519.25
1500	3.63	0.68	574.92
2000	3.72	0.69	536.50
2500	3.75	0.35	641.50
3000	3.76	0.62	703.00
3500	4.02	0.48	856.25
4000	4.12	0.71	920.52

TABLE VIII
THE SPEEDUP OF VWZ, OF THE LU FACTORIZATION AND OF THE STANDARD MATLAB FUNCTION `LU` — RELATIVE TO THE SEQUENTIAL WZ FACTORIZATION (MATLAB R2008 ON THE INTEL PROCESSOR)

matrix size	VWZ	LU	`lu`
500	2.56	0.99	159.00
1000	2.14	1.03	165.44
1500	2.00	0.99	189.30
2000	1.99	1.03	204.50
2500	1.96	0.88	206.35
3000	1.85	0.96	96.11

TABLE IX
THE SPEEDUP OF VWZ, OF THE LU FACTORIZATION AND OF THE STANDARD MATLAB FUNCTION `LU` — RELATIVE TO THE SEQUENTIAL WZ FACTORIZATION (MATLAB R2010 ON THE INTEL PROCESSOR)

matrix size	VWZ	LU	`lu`
500	2.01	0.92	95.01
1000	2.01	0.70	175.04
1500	1.75	0.90	144.39
2000	1.95	0.86	101.59
2500	1.82	0.86	265.22
3000	1.39	0.78	119.58

[3] B. Bylina, J. Bylina: Analysis and Comparison of Reordering for Two Factorization Methods (LU and WZ) for Sparse Matrices, *Lecture Notes in Computer Science* **5101**, Springer-Verlag Berlin Heidelberg 2008, pp. 983–992.

[4] B. Bylina, J. Bylina: Incomplete WZ Factorization as an Alternative Method of Preconditioning for Solving Markov Chains, *Lecture Notes in Computer Science* **4967**, Springer-Verlag Berlin Heidelberg 2008, 99–107.

[5] B. Bylina, J. Bylina: Influence of preconditioning and blocking on accuracy in solving Markovian models, *International Journal of Applied Mathematics and Computer Science* 19 (2) (2009), pp. 207–217.

[6] B. Bylina, J. Bylina: The Vectorized and Parallelized Solving of Markovian Models for Optical Networks, *Lecture Notes in Computer Science* **3037**, Springer-Verlag Berlin Heidelberg 2004, 578–581.

[7] T. A. Davis, Algorithm 930: FACTORIZE: An Object-oriented Linear System Solver for MATLAB, *ACM Trans. Math. Softw.*, Volume 39 Issue 4, July 2013, Article No. 28. pages = 28:1–28:18

[8] S. Chandra Sekhara Rao: Existence and uniqueness of WZ factorization, *Parallel Computing* **23** (1997), pp. 1129–1139.

[9] Choi, T. Sou-Cheng, M. A. Saunders, Algorithm 937: MINRES-QLP for Symmetric and Hermitian Linear Equations and Least-squares Problems, *ACM Trans. Math. Softw.*, Volume 40 Issue 2, February 2014, Article No. 16. pages = 16:1–16:12,

[10] D. J. Evans, M. Hatzopoulos: The parallel solution of linear system, *Int. J. Comp. Math.* **7** (1979), pp. 227–238.

[11] X. Ji, J. Sun, T. Turner, Algorithm 922: A Mixed Finite Element Method for Helmholtz Transmission Eigenvalues, *ACM Trans. Math. Softw.*, Volume 38 Issue 4, August 2012, Article No. 29. pages = 29:1–29:8,

[12] K. Poppe, R. Cools, CHEBINT: A MATLAB/Octave Toolbox for Fast Multivariate Integration and Interpolation Based on Chebyshev Approximations over Hypercubes, *ACM Trans. Math. Softw.*, Volume 40 Issue 1, September 2013, Article No. 2. pages = 2:1–2:13,

[13] P. Yalamov, D. J. Evans: The WZ matrix factorization method, *Parallel Computing* **21** (1995), pp. 1111–1120.

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 569–576

DOI: 10.15439/2014F313
ACSIS, Vol. 2

Performance Analysis of Multicore and Multinodal Implementation of SpMV Operation

Beata Bylina[1], Jarosław Bylina[2],
Przemysław Stpiczyński[3], Dominik Szałkowski[4]
Institute of Mathematics
Maria Curie-Sklodowska University
Lublin, Poland
Email: beata.bylina@umcs.pl[1], jaroslaw.bylina@umcs.pl[2]
przemyslaw.stpiczynski@umcs.pl[3], dominik.szalkowski@umcs.pl[4]

***Abstract*—In this paper we present two algorithms for performing sparse matrix-dense vector multiplication (known as SpMV operation). We show parallel (*multicore*) version of algorithm, which can be efficiently implemented on the contemporary multicore architectures. Next, we show distributed (so-called *multinodal*) version targeted at high performance clusters. Both versions are thoroughly tested using different architectures, compiler tools and sparse matrices of different sizes. Considered matrices comes from The University of Florida Sparse Matrix Collection. The performance of the algorithms is compared to the performance of SpMV routine from widely used Intel Math Kernel Library.**

Keywords: sparse matrix-dense vector multiplication, SpMV operation, parallel matrix-vector multiplication, multicore platforms, computer cluster.

I. Introduction

IN THIS paper we consider multiplication of a sparse matrix by a dense vector, which is called SpMV operation. This operation is fundamental part of many numerical algorithms [4], [8]. In particular SpMV is used for iterative solving of systems of linear equations, e.g. in projective GMRES method or CG method.

Given a $n \times n$ square, sparse matrix $\mathbf{A}$ and a dense vector $\mathbf{x}$ of dimension n we define operation SpMV as

$$\mathbf{y} \leftarrow \mathbf{A}\mathbf{x}.$$

Let us denote ith row of matrix $\mathbf{A}$ by $\mathbf{A}(i, 1 : n)$. Then, to compute ith element of vector $\mathbf{y}$, we have to compute the dot product of $\mathbf{A}(i, 1 : n)$ and $\mathbf{x}$ vectors. So the whole operation of computing $\mathbf{y}$ vector can be easy parallelized, since the computations of all resulting elements are independent. Hence SpMV operation can be treated as n distinct tasks, which have ith row of $\mathbf{A}$ and $\mathbf{x}$ as a input data and produce ith element of $\mathbf{y}$. Note that $\mathbf{x}$ is shared between all computing tasks.

In the paper [11] authors focus on SpMV operation in the case of multicore platforms. They survey some low level optimization techniques related to hardware properties, while using CSR format for storing sparse matrices. All techniques are then benchmarked on a few multicore environments.

In the article [12] authors show a new format suitable for multicore architectures, which they call *Compressed Sparse Block* (*CSB*). It allows effective storage and efficient computations. It also uses special optimizations in the case of multiplication of banded matrices.

The aim of this paper is to present our research on the efficient implementation of a sparse matrix by a dense vector multiplication with the use of contemporary parallel multicore and distributed computer architectures to gain high performance at low cost.

We propose a parallel SpMV algorithm based on modified SPARSKIT library routine [9] targeted at multicore platforms. We investigate the performance of this algorithm using various architectures, compilers and a few sparse matrices, which arises in real life problems. These matrices comes from The University of Florida Sparse Matrix Collection [3]. We include optimized SpMV routine from Intel Math Kernel Library [5] in the comparison.

Next we introduce a distributed algorithm for computing SpMV on computer clusters consisting of multiple nodes. Our universal approach allows to use any existing SpMV implementation locally within one node. For performance comparison we use the same set of sparse matrices as previously.

The paper is structured as follows. Section II describes data structures suitable for representing sparse matrices and their usability for the implementation of SpMV operation. Next section contains short description of standard, sequential SpMV algorithm. In Section IV we present multicore version of existing SPARSKIT SpMV routine. The description of SpMV algorithm for distributed environments is included in Section V. Then we present some numerical results and concluding remarks in sections VI and VII respectively.

II. Storage formats for sparse matrices

Special data structures and algorithms are used for storing sparse matrices (for efficient memory usage) and performing basic mathematical operations. The survey of many storage formats can be found in [8]. Note, that the same formats are used in algorithms designed for sequential and parallel architectures. However, due to different properties of these architectures, different formats may be preferred in each case.

978-83-60810-58-3/$25.00

$$A = \begin{bmatrix} -4 & 0 & 0 & 0 & 1 \\ 0 & -1 & 0 & 8 & 0 \\ 0 & 0 & 0 & 5 & 0 \\ -1 & 31 & 0 & 21 & -1 \\ 0 & 0 & 0 & 0 & -8 \end{bmatrix}$$

Fig. 1. Sparse matrix stored in dense format

$$data = \begin{bmatrix} -4 & 1 & -1 & 8 & 5 & -1 & 31 & 21 & -1 & -8 \end{bmatrix}$$
$$col = \begin{bmatrix} 0 & 4 & 1 & 3 & 3 & 0 & 1 & 3 & 4 & 4 \end{bmatrix}$$
$$row = \begin{bmatrix} 0 & 0 & 1 & 1 & 2 & 3 & 3 & 3 & 3 & 4 \end{bmatrix}$$

Fig. 2. Matrix from Fig. 1 stored in COO format

Below we shortly present three widely used formats for storage of sparse matrices. Fig. 1 shows square, sparse matrix of dimension 5 stored in dense format, which, from the programmers point of view, is equivalent to using one two-dimensional array.

A. Coordinate Format (COO)

The simplest and the most flexible format for storing any sparse matrix is so-called *Coordinate Format* or *COO* for short. In this format, only nonzero values are stored, together with rows and columns indexes. Technically, it uses three one-dimensional arrays:

- $data$ for storing nonzero elements,
- col for storing indexes of columns of nonzero elements in the original matrix,
- row for storing indexes of rows of nonzero elements in the original matrix.

On Fig. 2 we see COO storage scheme for the matrix from Fig. 1. Unfortunately, there are some disadvantages of this format, namely it is not memory and computationally efficient (especially in the case of SpMV operation). Note that MATLAB software uses this format [6].

B. Matrix Market Format (MM)

The University of Florida Sparse Matrix Collection [3] is large repository of sparse matrices, which comes from real life applications. It uses Matrix Market format (*MM*) for storing sparse matrices. This format is based on COO with some optimizations added, e.g. it can store only the half of the matrix, in case it is symmetric [7].

C. Compressed Sparse Row Format (CSR)

Another way to store sparse matrix is to use *Compressed Sparse Row* format (*CSR*). As in the case of COO, only the nonzero elements are stored and their columns indexes, while the rows indexes are kept in somewhat different way. There are also three one-dimensional arrays used:

- $data$ which keeps nonzero elements,
- col which keeps indexes of columns of nonzero elements in the original matrix,

$$data = \begin{bmatrix} -4 & 1 & -1 & 8 & 5 & -1 & 31 & 21 & -1 & -8 \end{bmatrix}$$
$$col = \begin{bmatrix} 0 & 4 & 1 & 3 & 3 & 0 & 1 & 3 & 4 & 4 \end{bmatrix}$$
$$ptr = \begin{bmatrix} 0 & 2 & 4 & 5 & 9 & 10 \end{bmatrix}$$

Fig. 3. Matrix from Fig. 1 stored in CSR format

```
      do 100 i = 1,n
         t = 0.0d0
         do 99 k=ptr(i), ptr(i+1)-1
            t = t + data(k)*x(col(k))
 99      continue
         y(i) = t
100   continue
```

Fig. 4. Standard implementation of SpMV for CSR storage

- ptr which keeps indexes of the beginnings of the consecutive rows in $data$ array.

Fig. 3 shows sparse matrix stored in CSR format.

III. Sequential SpMV algorithm

CSR is the most common format used, when dealing with applications containing many SpMV operations. Basic, sequential implementation of SpMV is presented on Fig. 4. We assume that `data`, `col` and `ptr` arrays keeps a sparse matrix in CSR format, while `x` is given vector and `y` is the result of the operation.

IV. Multicore SpMV algorithm

SPARSKIT [9] is Fortran library for dealing with sparse matrices. It provides several formats for storing matrices (including CSR) and routines for performing fundamental mathematical operations. There is SpMV routine in this library for matrices stored in the CSR format, however it is strictly sequential, hence it doesn't take advantage of contemporary parallel architectures. We used OpenMP [10] directives for simple and effective parallelization (use of all present CPU cores) of available source code. The modified source code using `omp parallel do` directive is presented on Fig. 5. We will refer to this algorithm as the *multicore algorithm*.

V. Multinodal SpMV algorithm

In this section we present distributed version of SpMV for clusters consisting of multicore nodes, which we will call the *multinodal algorithm*.

Assume that $\mathbf{A} \in \mathbf{R}^{n \times n}$ matrix is divided into p^2 blocks (with possible different dimensions)

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{00} & \cdots & \mathbf{A}_{0,p-1} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{p-1,0} & \cdots & \mathbf{A}_{p-1,p-1} \end{bmatrix},$$

```
      subroutine pamux (n, x, y, a,ja,ia)
      real*8  x(*), y(*), a(*)
      integer n, ja(*), ia(*)
      real*8 t
      integer i, k
!$omp parallel do private(t,k)
      do 100 i = 1,n
         t = 0.0d0
         do 99 k=ia(i), ia(i+1)-1
            t = t + a(k)*x(ja(k))
 99      continue
         y(i) = t
100   continue
!$omp end parallel do
      return
      end subroutine pamux
```

Fig. 5. Implementation of SPARSKIT SpMV routine using OpenMP

where $\mathbf{A}_{ij} \in \mathbf{R}^{n_i \times n_j}$ and $\sum_{i=0}^{p-1} n_i = n$. Vectors $\mathbf{x}$ and $\mathbf{y}$ are also divided into p blocks, where $\mathbf{x}_i, \mathbf{y}_i \in \mathbf{R}^{n_i}$. Then

$$\begin{bmatrix} \mathbf{y}_0 \\ \vdots \\ \mathbf{y}_{p-1} \end{bmatrix} \leftarrow \begin{bmatrix} \mathbf{A}_{00} & \dots & \mathbf{A}_{0,p-1} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{p-1,0} & \dots & \mathbf{A}_{p-1,p-1} \end{bmatrix} \begin{bmatrix} \mathbf{x}_0 \\ \vdots \\ \mathbf{x}_{p-1} \end{bmatrix}$$

and

$$\mathbf{y}_i \leftarrow \sum_{j=0}^{p-1} \mathbf{A}_{ij}\mathbf{x}_j, \qquad i = 0, \dots, p-1. \tag{1}$$

Algorithm 1 describes the multiplication of sparse matrix by dense vector using computer cluster, which has at least p^2 distinct CPUs. The matrix is appropriately distributed between the grid of $p \times p$ tasks denoted by P_{ij}, $0 \le i, j < p$. Algorithm comprises the following steps:

1) sending the data from the first column of tasks to "diagonal" tasks,
2) broadcasting the data columnwise,
3) performing actual computations,
4) performing global reduction.

We assume that each computing task is running on different CPU, however there can be more than one CPU installed in one cluster node.

On figure 6 we see the grid of tasks in the case of dimension 2×2. The grid in the case of 4×4 dimension, together with communication scheme is presented on Fig. 7.

To implement Algorithm 1 we used routines from BLACS (Basic Linear Algebra Communication Subprograms) [1] and MKL (Intel Math Kernel Library) [5] libraries. BLACS routines were used for organizing task grid and transferring data between tasks, while optimized multicore `mkl_dcsrgem` from MKL was used for performing SpMV. The most important part of Fortran implementation of this algorithm is presented on Fig. 8. We used the following variables in the implementation:

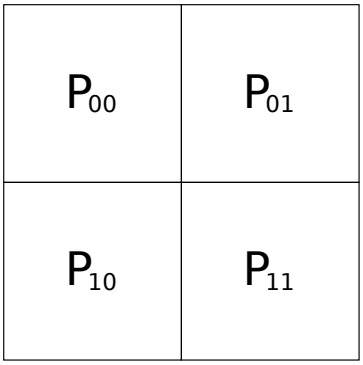


Fig. 6. Task grid of dimension 2×2

- `proc_row`, `proc_col` are coordinates of current task in the task grid,
- arrays `data`, `col`, `ptr` store block $\mathbf{A}_{\texttt{proc\_row,proc\_col}}$ of sparse matrix in CSR format,
- arrays `x` and `y` keep vectors $\mathbf{x}$ and $\mathbf{y}$ respectively,
- `nrows` and `ncols` denotes the number of rows and the number of columns of the $\mathbf{A}_{\texttt{proc\_row,proc\_col}}$ block,
- `context` describes appropriate BLACS context.

Note, that this algorithm is not tied to `mkl_dcsrgem` routine. Instead, it can use implementation from Section IV or any other available code.

VI. Numerical experiments

In this section we review the tests of our SpMV implementations for multicore (Section IV) and distributed (Section V) systems.

A. Data

We used several sparse matrices from The University of Florida Sparse Matrix Collection [3]. All matrices were downloaded in the Matrix Market format and then were converted to CSR format, which was used in all numerical experiments. We present the results for 4 matrices: parabolic_fem, bmw3_2, torso1, nd24k. Detailed parameters are shown in Table I, where we have:

- n is the number of rows and columns,
- nz is the number of nonzero elements,
- $d = nz/n$ denotes the density of the matrix.

Fig. 9 shows the sparsity patterns of these matrices.

Considered matrices were first read from the files and then distributed between all running MPI tasks using BLACS `dgesd2d` routine. We used simple distribution scheme, in which we divided the matrices into the blocks of almost the same sizes.

Note, that instead of reading data from files it is possible to generate matrices locally in each node.

B. Test environment

We used two hardware platforms for testing: E5-2660 and X5650. Their specifications are presented in Table II.

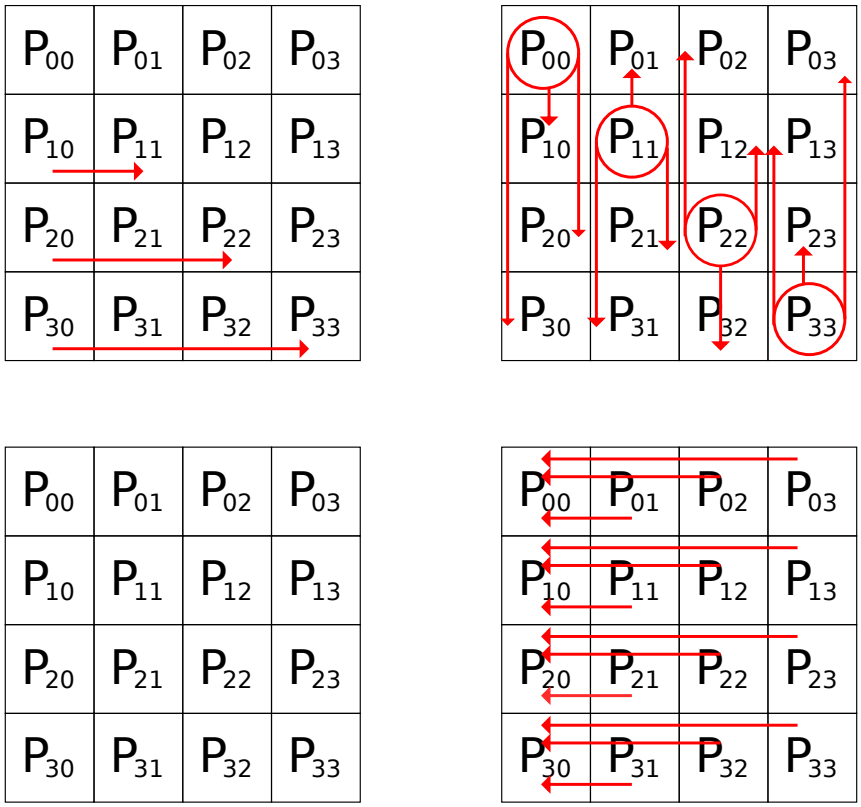


Fig. 7. Communication scheme for the 4×4 task grid: 1) sending the data from the first column of tasks to "diagonal" tasks (top left), 2) broadcasting the data columnwise (top right), 3) performing actual computations (no communication, bottom left), 4) performing global reduction (bottom right)

```
! step 1) sending the data to "diagonal" tasks
   if ((proc_col.eq.0).and.(proc_row.ne.0)) then
      call dgesd2d(context, nrows, 1, y, 1, proc_row, proc_row)
   else
      if ((proc_row.eq.proc_col).and.(proc_row.ne.0)) then
         call dgerv2d(context, nrows, 1, x, 1, proc_row, 0)
      end if
   end if

! step 2) broadcasting the data columnwise
   if (proc_row.eq.pcol) then
      call dgebs2d(context, 'C', ' ', ncols, 1, x, 1)
   else
      call dgebr2d(context, 'C', ' ', ncols, 1, x, 1, proc_col, proc_col)
   end if

! step 3) performing actual computations
   call mkl_dcsrgemv('N', nrows, data, col, ptr, x, y)

! step 4) performing global reduction
   call dgsum2d(context, 'R', ' ', nrows, 1, y, 1, proc_row, 0)
```

Fig. 8. Main part of Fortran implementation of Algorithm 1

Algorithm 1 Outline of the multinodal SpMV algorithm

Require: Each P_{ij} task holds $\mathbf{A}_{ij}$ matrix, each P_{i0}, $0 \leq i < p$, task stores $\mathbf{x}_i$ and $\mathbf{y}_i$
Ensure: Each P_{i0}, $0 \leq i < p$, task holds resulting $\mathbf{y}_i$ vector obtained using equation (1)
1: Each P_{i0}, $0 < i < p$, task sends $\mathbf{x}_i$ to P_{ii}
2: Each P_{jj}, $0 \leq j < p$, task broadcasts $\mathbf{x}_j$ to P_{ij}, $0 \leq i < p$
3: Each P_{ij}, $0 \leq i, j < p$, task performs $\mathbf{t}_{ij} \leftarrow \mathbf{A}_{ij}\mathbf{x}_j$
4: Global reduction $\mathbf{y}_i \leftarrow \sum_{j=0}^{p-1} \mathbf{t}_{ij}$ is performed by P_{ij}, $0 \leq j < p$ tasks {$\mathbf{y}_i$ vector is held by P_{i0}, $0 \leq i < p$}

TABLE I
PARAMETERS OF CONSIDERED SPARSE MATRICES

name	*n*	*nz*	*d*	*symmetricity*
parabolic_fem	525825	3674625	6.98	symmetric
bmw3_2	227632	11288630	49.59	symmetric
torso1	116158	8516500	73.32	symmetric
nd24k	72000	28715634	398.83	non-symmetric

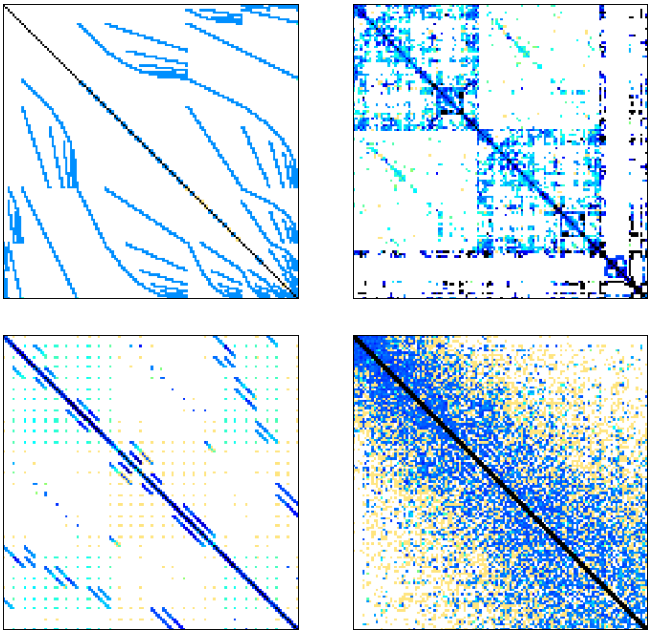


Fig. 9. Sparsity patterns of parabolic_fem (top left), bmw3_2 (top right), torso1 (bottom left) and nd24k (bottom right) matrices

Our algorithms were implemented in Fortran 95 using appropriate parallel and numerical libraries (OpenMP, MPI, SPARSKIT, MKL). Two Fortran compilers, namely `ifort` by Intel and `pgfortran` by The Portland Group, were used for compiling source codes with compiler flags, which are shown in Table III.

For time measurement we used the following routines:

- `omp_get_wtime()` for the multicore version,
- `MPI_Wtime()` in the multinodal case.

C. Results for multicore algorithm

On figures 10, 11, 12 and 13 we see the performance (in Gflops) of the multicore version of SpMV multiplication. The performance is shown for two platforms (X5650, E5-2660), two compilers (Intel, pgi), and we also include the performance chart of SpMV routine from Intel MKL library optimized for multicore CPUs (denoted by mkl).

Using obtained results we conclude that:

- For small number of running threads the performance is similar in each case.
- For growing number of threads E5-2660 architecture outperforms X5650, due to its older architecture. We were expecting this result.
- Compiler version has negligible impact on the performance of the algorithms, however there is a drop in the performance in the case of `pgfortran` dealing with large number of threads.
- Simple SPARSKIT implementation with OpenMP directives (Fig. 5) gives as good performance as the optimized MKL version of SpMV.

TABLE IV
PERFORMANCE (GFLOPS) OF THE MULTINODAL SPMV ALGORITHM

matrix	1 task	4 tasks	16 tasks
parabolic_fem	1.91	0.31	0.41
bmw3_2	2.69	2.02	3.05
torso1	3.12	4.49	8.25
nd24k	3.31	4.39	12.00

D. Results for multinodal algorithm

Table IV shows the performance of our multinodal SpMV implementation (Algorithm 1). In the column denoted by "1 task" we see the performance of the multicore version compiled by `ifort` with MKL support and running on X5650 system. Multinodal version was compiled using `mpiifort` compiler and was running on cluster consisting of 2 or 8 X5650 nodes, connected using 40Gbit/s Infiniband. To optimize the workload of each node we used the following number of MPI tasks:

- 4 tasks were running on 2 nodes with 4 CPUs (as in Fig. 6),
- 16 tasks were running on 8 nodes with 16 CPUs (as in Fig. 7).

Each MPI task was using multithreaded version of SpMV from MKL.

Looking at Table IV we see, that:

- there are cases, when the algorithm achieves very high scalability (e.g. nd24k matrix)
- for some matrices (e.g. parabolic_fem), the performance decreases,

TABLE II
SOFTWARE AND HARDWARE PROPERTIES OF E5-2660 AND X5650 SYSTEMS

	E5-2660 System	X5650 System
CPU	2x Intel E5-2660 (20M Cache, 2.20 GHz, 8 cores with HT)	2x Intel Xeon X5650 (12M Cache, 2.66 GHz, 6 cores with HT)
CPU memory	48GB DDR3	48GB DDR3
Operating system	CentOS 5.5 (Linux 2.6.18-164.el5)	Debian (GNU/Linux 7.0)
Libraries	OpenMP, SPARSKIT, Intel Composer XE 2013	OpenMP, MPI, SPARSKIT, Intel Composer XE 2013
Compilers	The Portland Group, Intel	The Portland Group, Intel

TABLE III
COMPILER FLAGS

Algorithm version	Compiler	Compiler flags
Multicore for E5-2660	`ifort` by Intel	`-O3 -openmp -xAVX`
Multicore for X5650	`ifort` by Intel	`-O3 -openmp -xSSE4.2`
Multicore for E5-2660 and MKL	`ifort` by Intel	`-O3 -openmp -mkl=parallel -xAVX`
Multicore for X5650 and MKL	`ifort` by Intel	`-O3 -openmp -mkl=parallel -xSSE4.2`
Multicore (both systems)	`pgfortran` by The Portland Group	`-O3 -mp -fastsse`
Multinodal for X5650 and MKL	`mpifort` by Intel	`-O3 -openmp -mkl=parallel -xSSE4.2`

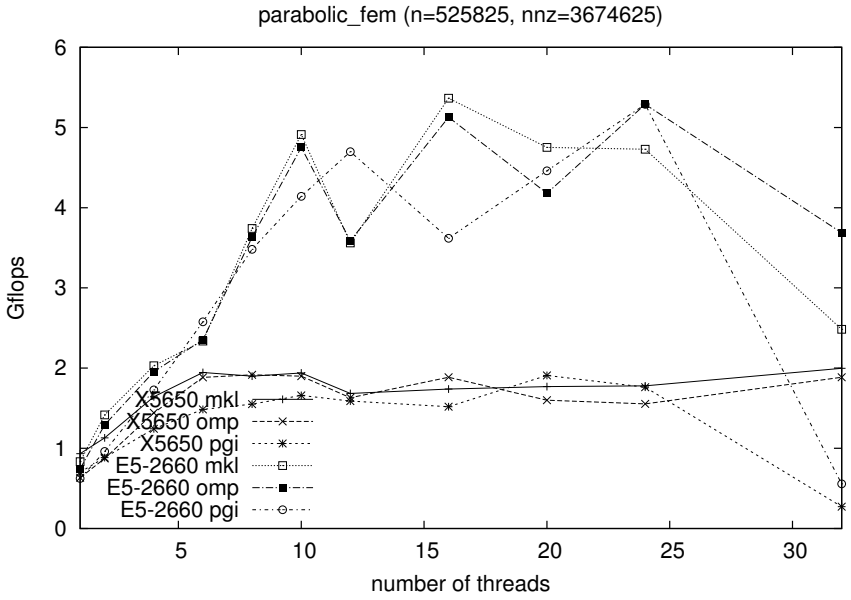


Fig. 10. The performance of SpMV operation for parabolic_fem matrix

TABLE V
DISTRIBUTION OF TORSO1 AND PARABOLIC_FEM MATRICES BETWEEN 4 TASKS

	$torso1$			$parabolic\_fem$		
$name$	n	nz	d	n	nz	d
A_{00}	58079	2407469	41.45	262912	656124	2.50
A_{01}	58079	1693760	29.16	262912	0	0
A_{10}	58079	1941287	33.42	262913	787970	3.00
A_{11}	58079	2473984	42.59	262913	656131	2.50

TABLE VI
DISTRIBUTION OF TORSO1 AND PARABOLIC_FEM MATRICES BETWEEN 16 TASKS

	$torso1$			$parabolic\_fem$		
$name$	n	nz	d	n	nz	d
A_{00}	29039	756119	26.04	131456	131456	1.00
A_{01}	29039	401235	13.82	131456	0	0.00
A_{02}	29039	75227	2.59	131456	0	0.00
A_{03}	29039	589153	20.29	131456	0	0.00
A_{10}	29039	401361	13.82	131456	262142	1.99
A_{11}	29039	848745	29.22	131456	262142	1.99
A_{12}	29039	444083	15.29	131456	0	0.0
A_{13}	29039	585297	20.15	131456	0	0.0
A_{20}	29039	75417	2.59	131456	261886	1.99
A_{21}	29039	444272	15.30	131456	121560	0.92
A_{22}	29039	806323	27.77	131456	241257	1.83
A_{23}	29039	254204	8.75	131456	0	0.00
A_{30}	29041	737556	25.40	131457	262144	1.99
A_{31}	29041	684042	23.55	131457	142380	1.08
A_{32}	29041	254204	8.75	131457	185688	1.41
A_{33}	29041	1159262	39.92	131457	229186	1.74

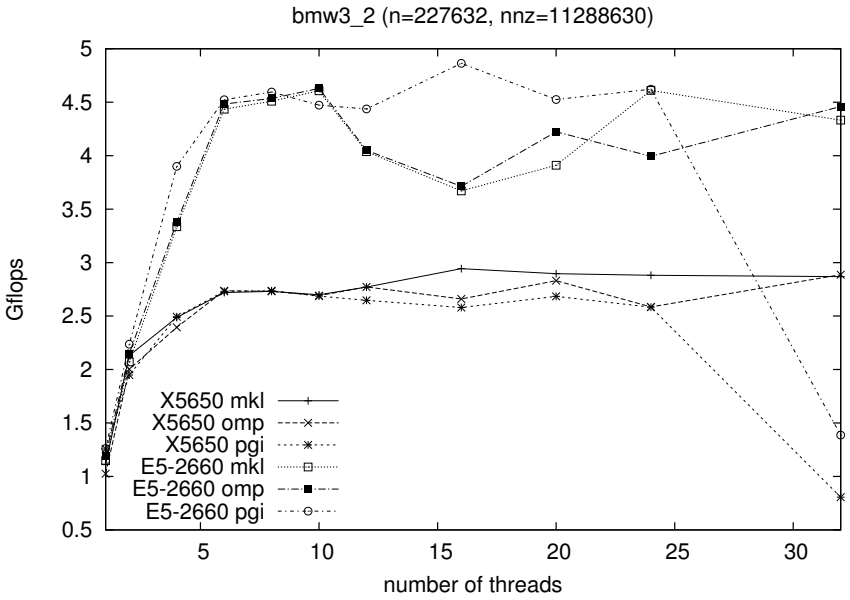


Fig. 11. The performance of SpMV operation for bmw3_2 matrix

- the scalability of the algorithm is related to the distribution scheme and to the matrix properties, especially to its density — sparser the matrix, worse the scalability. Table V shows the division of torso1 and parabolic_fem matrices in the case of 4 tasks. We see, that the densities of the resulting blocks are smaller than the densities of whole matrices (showed in the Table I). Notice, that for parabolic_fem matrix there is a block with no nonzero elements, hence one of the nodes stay idle. The situation is even worse, when we consider 16 tasks division. Looking at Table VI we see, that there are 6 empty blocks for parabolic_fem matrix and the densities of the rest of them are very low.

VII. Conclusion and future work

In this work we investigated the parallelization of SpMV operation, an important and highly-demanding numerical kernel used in many numerical methods. We compared various implementations, namely routine from MKL library, OpenMP parallelized SPARSKIT version compiled using two compilers and two architectures and our own distributed version. The results show, that the most important factor of achieving high performance is hardware architecture together with the distribution pattern and the properties of sparse matrices. It is worth to note, that it is possible to further optimize multinodal implementation (Algorithm 1) by distributing blocks of matrices between nodes taking into account the original matrix density to obtain balanced workload of each MPI task.

Another way to speed up computations is to use GPU cards or MIC architecture accelerators instead of CPUs.

Acknowledgements

This work was prepared using the supercomputer resources provided by the Institute of Mathematics of the Maria Curie-Skłodowska University in Lublin.

References

[1] *Basic Linear Algebra Communication Subprograms*, http://www.netlib.org/blacs/

[2] B. Bylina, J. Bylina, M. Karwacki: *Computational Aspects of GPU-accelerated Sparse Matrix-Vector Multiplication for Solving Markov Models*; Theoretical and Applied Informatics, 23 (2011), no. 2, ISSN 1896-5334, pp. 127–145.

[3] T. A. Davis, Y. Hu, *The University of Florida Sparse Matrix Collection*, ACM Transactions on Mathematical Software, Vol 38, 2011, pp.1-25, http://www.cise.ufl.edu/research/sparse/matrices.

[4] G. H. Golub, C. F. van Van Loan: *Matrix Computations*, Johns Hopkins Studies in Mathematical Sciences, 3rd Edition, 2013.

[5] *Intel Math Kernel Library*, http://software.intel.com/en-us/articles/intel-mkl/

[6] *MATLAB. The Language of Technical Computing*, http://www.mathworks.com/products/matlab/

[7] *Matrix Market Exchange Formats*, http://math.nist.gov/MatrixMarket/formats.html

[8] Y. Saad, *Iterative Methods for Sparse Linear Systems: Second Edition*, SIAM, 2003.

[9] Y. Saad, *SPARSKIT: A basic tool kit for sparse computations; Version 2*, June 1994.

[10] *The OpenMP API specification for parallel programming*, http://openmp.org/

[11] S. Williams, L. Oliker, R. Vuduc, J. Shalf, K. Yelick, J. Demmel, *Optimization of sparse matrix-vector multiplication on emerging multicore platforms*, Parallel Computing 35 (2009), pp. 178-194.

[12] Yang, B., Gu, S., Gu, T.-X., Zheng, C. and Liu, X.-P. (2014) Parallel Multicore CSB Format and Its Sparse Matrix Vector Multiplication. *Advances in Linear Algebra & Matrix Theory, 4, 1-8.*

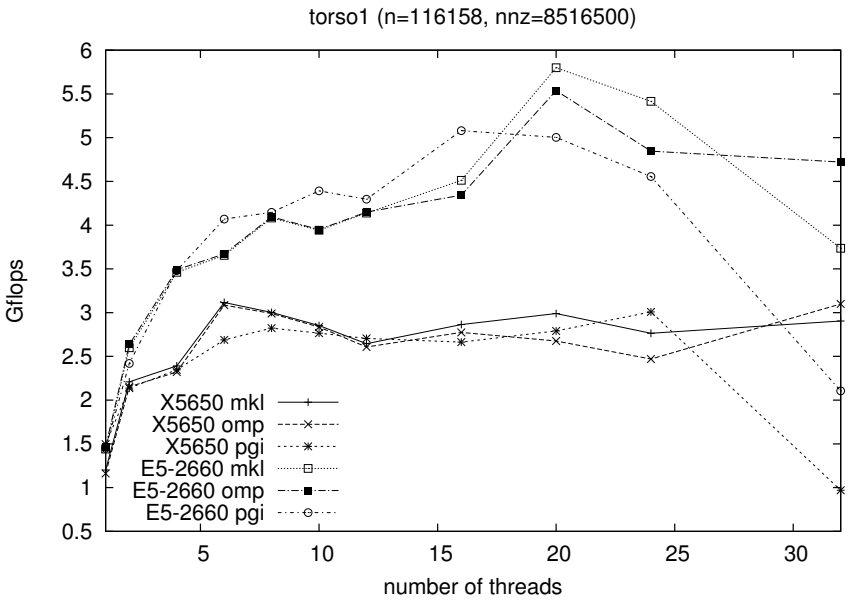


Fig. 12. The performance of SpMV operation for torso1 matrix

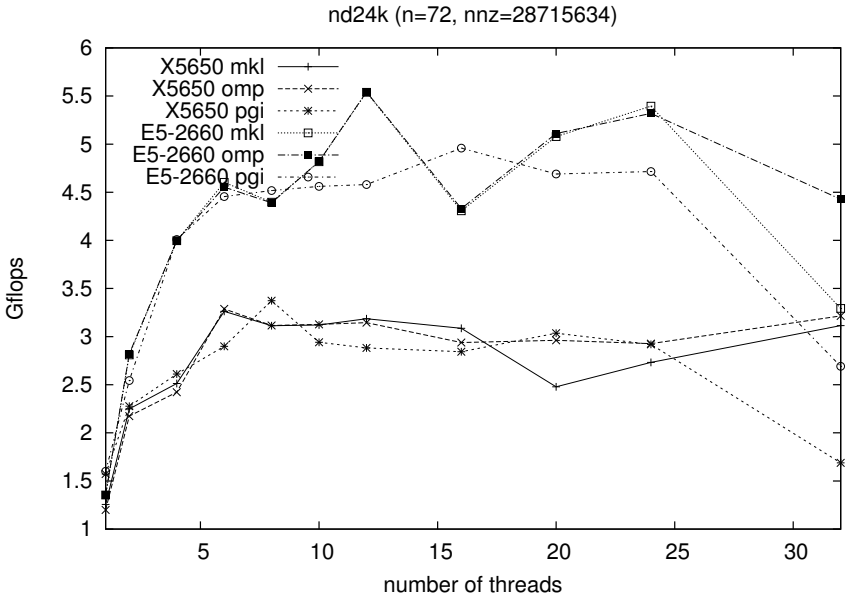


Fig. 13. The performance of SpMV operation for nd24k matrix

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 577–586
DOI: 10.15439/2014F340
ACSIS, Vol. 2

Implementation of a distributed parallel in time scheme using PETSc for a Parabolic Optimal Control Problem

Juan José Cáceres Silva*, Benjamín Barán*†‡ and Christian E. Schaerer†
*Science and Technology School, Catholic University, Asunción, Paraguay
†Polythechnic School, National University of Asunción, San Lorenzo, Paraguay, P.O.Box: 2111 SL.
‡Polythechnic School, East National University, Ciudad del Este, Paraguay

***Abstract*—This work presents a parallel implementation of the Parareal method using Portable Extensible Toolkit for Scientific Computation (PETSc). An optimal control problem of a parabolic partial differential equation with known boundary conditions and initial state is solved, where the minimized cost function relates the controller v usage and the approximation of the solution y to an optimal known function y^*, measured by $\|y\|$ and $\|y^*\|$, respectively. The equations that model the process are discretized in space using Finite Elements and in time using Finite Differences. After the discretizations, the problem is transformed to a large linear system of algebraic equations, that is solved by the Conjugate Gradient method. A Parareal preconditioner is implemented to speed up the convergence of the Conjugate Gradient.**

The main advantage in using the Parareal approach is to speed up the resolution time, when comparing to implementations that use only the Conjugate Gradient or GMRES methods. The implementation developed in this work offers a parallelization relative efficiency for the strong scaling of approximately 70% each time the process count doubles. For weak scaling, 75% each time the process count doubles for a constant solution size per process and 96% each time the process count doubles for a constant data size per process.

I. Introduction

MANY challenges of engineering design, such as heat dissipation, electromagnetic inversion, diffraction tomography among others, can be modeled as a parabolic optimal control problem [1]. The problem to be solved is [2, 3]:

$$\begin{cases} minimize & J(y,v) \\ s.t. & \nabla_t y = \Delta_x y + v\,, \end{cases} \tag{1}$$

with [3]

$$J(y,v) = \frac{\alpha}{2}\int_\Omega\int_{t_0}^{t_f} \|y-y^*\|_2^2 dtdx + \frac{\beta}{2}\int_\Omega \|y(t_f)-y^*(t_f)\|_2^2 dx + \frac{\gamma}{2}\int_\Omega\int_{t_0}^{t_f}\|v\|_2^2 dtdx\,, \tag{2}$$

where y^* is the optimal condition for the function y, α is the weight for the general approximation of the function y, β is the weight for the approximation at the final instant of the function y and γ gives the cost of the controller usage.

The finite elements discretization using the Galerkin method yields the following state equation [4, 5]:

$$M\dot{\underline{z}} = K\underline{z} + B\underline{u} \tag{3}$$

where $\underline{z} \in \mathbb{R}^{\hat{q}}$ is the nodal representation of y, $\underline{u} \in \mathbb{R}^{\hat{p}}$ is the nodal representation of v, M is the mass matrix, K is the stiffness matrix and B is the coupling matrix. Using this discretization, the cost function (2) becomes:

$$J_h(\underline{z},\underline{u}) = \frac{\alpha}{2}\int_{t_0}^{t_f}(\underline{z}-\underline{z}^*)^T M(\underline{z}-\underline{z}^*)\,dt + \frac{\beta}{2}\left\{[\underline{z}(t_f)-\underline{z}^*(t_f)]^T M[\underline{z}(t_f)-\underline{z}^*(t_f)]\right\} + \frac{\gamma}{2}\int_{t_0}^{t_f}\underline{u}^T R\underline{u}\,dt \tag{4}$$

where $\underline{z}^*$ is the nodal representation of y^* and R is the controller's mass matrix.

The finite differences discretization, using a time interval τ with $\hat{l}$ time instants, is based on equation [4]:

$$F_1\underline{z}(i+1) = F_0\underline{z}(i) + \tau B\underline{u}(i+1);\ \text{for } 0 < i < \hat{l} \quad \text{and } \underline{z}(0) = y_0 \tag{5}$$

where $F_0, F_1 \in \mathbb{R}^{\hat{q}\times\hat{q}}$ are matrices defined by $F_0 = M$ and $F_1 = M + \tau K$. The arrangement of equation (3) for all times yield:

$$\mathbf{Ez} + \mathbf{Nu} = \mathbf{f}_3 \tag{6}$$

where $\mathbf{z} \in R^{\hat{l}\hat{q}}$ and $\mathbf{u} \in R^{\hat{l}\hat{p}}$. With a similar argument, equation (4) has the following form:

$$J_h^\tau(\mathbf{z},\mathbf{u}) = \frac{1}{2}(\mathbf{z}-\mathbf{z}^*)^T\mathbf{Q}(\mathbf{z}-\mathbf{z}^*) + \frac{1}{2}\mathbf{u}^T\mathbf{G}\mathbf{u} + (\mathbf{z}-\mathbf{z}^*)^T\mathbf{g}. \tag{7}$$

Using Lagrange multipliers [6] for minimizing equation (7) subject to equality constraint (6) and imposing first order optimality conditions [7, 8, 9], the following KKT system [3] with saddle point form [10] is obtained [2, 4, 11]:

$$\begin{bmatrix} \mathbf{Q} & 0 & \mathbf{E}^T \\ 0 & \mathbf{G} & \mathbf{N}^T \\ \mathbf{E} & \mathbf{N} & 0 \end{bmatrix}\begin{bmatrix} \mathbf{z} \\ \mathbf{u} \\ \mathbf{q} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{0} \\ \mathbf{f}_3 \end{bmatrix} \tag{8}$$

978-83-60810-58-3/$25.00

A. Schur's Equation

In order to simplify the linear system (8), the variables $\mathbf{z}$ and $\mathbf{q}$ are solved in terms of the control variable $\mathbf{u}$ [12, 4, 10, 13]. Making $\mathbf{z} = \mathbf{E}^{-1}\mathbf{f}_3 - \mathbf{E}^{-1}\mathbf{N}\mathbf{u}$, $\mathbf{q} = \mathbf{E}^{-T}\mathbf{f}_1 - \mathbf{E}^{-T}\mathbf{Q}\mathbf{z}$, and substituting on the final equation from (8), gives [13, 14, 15]:

$$\mathbf{H}\mathbf{u} = \mathbf{f} \tag{9}$$

where $\mathbf{H} = \mathbf{G} + \mathbf{N}^T\mathbf{E}^{-T}\mathbf{Q}\mathbf{E}^{-1}\mathbf{N}$ and $\mathbf{f} = \mathbf{N}^T\mathbf{E}^{-T}(\mathbf{Q}\mathbf{E}^{-1}\mathbf{f}_3 - \mathbf{f}_1)$. [1]

Doing this, the reduced Schur Complement [16, 14] Doing this, the equation system (8) is reduced. This expression is known as the Schur complement for equation (8) [16].

From this point on, the problem to solve is (9), a linear equation system, lets say $Ax = b$ for a general form, where the matrix A (in this case matrix $\mathbf{H}$ from equation (9)) is symmetric positive definite [10, 7, 4].

II. Mathematical Solution Formulation

A. Conjugate Gradient

The Conjugate Gradient method is used to solve a generic equation $Ax = b$ where $A \in \mathbb{R}^{\tilde{n}\times\tilde{n}}$ is symmetric positive definite and $b \in \mathbb{R}^{\tilde{n}}$.

On this work, the iterative algorithm defined in [17] is applied to the input matrix A and vector b, with error tolerance ε, initial guess x_0 and iteration limit for convergence max_i, as follows:

Algorithm 1 Conjugate Gradient

Input: $A, b, \varepsilon, x_0, max_i$
Output: x

1: $r_0 \leftarrow b - Ax_0$
2: $p_0 \leftarrow r_0$
3: $i \leftarrow 0$
4: **while** $r_{i+1} \geq \varepsilon \wedge i < max_i$ **do**
5: $\quad \alpha_i \leftarrow \frac{r_i^T r_i}{p_i^T A p_i}$ ▷ In our implementation, Ap_i is calculated by Algorithm 2
6: $\quad x_{i+1} \leftarrow x_i + \alpha_i p_i$
7: $\quad r_{i+1} \leftarrow r_i - \alpha_i A p_i$ ▷ In our implementation, Ap_i is calculated by Algorithm 2
8: $\quad \beta_i \leftarrow \frac{r_{i+1}^T r_i}{r_i^T r_i}$
9: $\quad p_{i+1} \leftarrow r_{i+1} + \beta_i p_i$
10: $\quad i \leftarrow i + 1$
11: **end while**
12: **if** $r_{i+1} < \epsilon$ **then**
13: $\quad$ **return** x_i ▷ Convergent
14: **else**
15: $\quad$ **return** $n.c.$ ▷ Not convergent
16: **end if**

[1] Recall that $\mathbf{u}, \mathbf{b} \in \mathbb{R}^{\hat{l}\hat{p}}$, $\mathbf{H}, \mathbf{G} \in \mathbb{R}^{\hat{l}\hat{p}\times\hat{l}\hat{p}}$, $\mathbf{N} \in \mathbb{R}^{\hat{l}\hat{q}\times\hat{l}\hat{p}}$ and $\mathbf{E}, \mathbf{Q} \in \mathbb{R}^{\hat{l}\hat{q}\times\hat{l}\hat{q}}$.

B. Using the Conjugate Gradient

In order to use Algorithm 1, the input matrix $\mathbf{H}$ must be previously computed, which requires a great computational work [5]. To avoid building matrix $\mathbf{H}$, steps 5) and 7) from Algorithm 1 are performed using Algorithm 2.

Let $\mathbf{s}$ be a generic input vector, and matrices $\mathbf{G}, \mathbf{N}, \mathbf{E}$ and $\mathbf{Q}$ as defined in Section I. The value of the product $\mathbf{H}\mathbf{s}$ is found using only matrix-vector operations, to avoid matrix-matrix operations that require more computational resources [18]. Algorithm 2 describes these matrix-vector operations [7].

Algorithm 2 Matrix-Vector Product $\mathbf{H}\mathbf{s}$

Input: $\mathbf{G}, \mathbf{N}, \mathbf{E}, \mathbf{Q}, \mathbf{s}$
Output: $\mathbf{x}$ ▷ $\mathbf{x} = \mathbf{H}\mathbf{s} = \mathbf{G}\mathbf{s} + \mathbf{N}^T\mathbf{E}^{-T}\mathbf{Q}\mathbf{E}^{-1}\mathbf{N}\mathbf{s}$

1: $\mathbf{s}_1 \leftarrow \mathbf{G}\mathbf{s}$
2: $\mathbf{s}_2 \leftarrow \mathbf{N}\mathbf{s}$
3: $\mathbf{s}_3 \leftarrow \mathbf{E}^{-1}\mathbf{s}_2$ ▷ $\mathbf{E}\mathbf{s}_3 = \mathbf{N}\mathbf{s}$, in our implementation, solved by Algorithm 4
4: $\mathbf{s}_4 \leftarrow \mathbf{Q}\mathbf{s}_3$ ▷ $\mathbf{s}_4 = \mathbf{Q}\mathbf{E}^{-1}\mathbf{s}_2$
5: $\mathbf{s}_5 \leftarrow \mathbf{E}^{-T}\mathbf{s}_4$ ▷ $\mathbf{E}^T\mathbf{s}_5 = \mathbf{Q}\mathbf{s}_3$, in our implementation, solved by Algorithm 4
6: $\mathbf{x} \leftarrow \mathbf{s}_1 + \mathbf{N}^T\mathbf{s}_5$ ▷ $\mathbf{x} = \mathbf{G}\mathbf{s} + \mathbf{N}^T\mathbf{s}_5$

The direct implementation of Algorithm 2 can be unviable since steps 3) and 5) require inverse matrices [7]. To avoid this, the steps 3) and 5) from the Algorithm 2 can be solved using an inner Conjugate Gradient. This step will have a high computational cost because it will be done for each iteration of the outer Conjugate Gradient.

The idea is to replace steps 3) and 5) from Algorithm 2 using the Parareal method [7].

C. Parareal

The Parareal method [19, 9] is an iterative method used to solve a time dependant equation, based on a time domain decomposition $[t_0, t_f]$ in $\hat{k}$ *coarse* time intervals, each of size $\Delta T = (t_f - t_0)/\hat{k}$, with $T_0 = t_0$ and $T_k = t_0 + k\Delta T$ for $1 \leq k \leq \hat{k}$. This sets the solution for each instant T_k with $1 \leq k \leq \hat{k}$ using the *multiple-shooting* technique [20, 21] that requires the parallel resolution of the equation $\mathbf{z} = \mathbf{E}^{-1}\mathbf{b}$ for each (T_{k-1}, T_k) subinterval. To accelerate each multiple-shooting iteration, the residual equations are preconditioned by a coarse time grid discretization, with a time interval ΔT [7].

An approximation $\mathbf{E}_n^{-1}$ for $\mathbf{E}^{-1}$, is based on n Richardson's iterations [22], through the Parareal algorithm, where the Richardson's algorithm is used as an external iteration for a Schur's complement problem [7, 16, 23]. The matrix $\mathbf{E}_n$ is used to approximate the solution $\mathbf{z}$ by $\mathbf{z}_n = \mathbf{E}_n^{-1}\mathbf{b}$, and the main interest is that $\mathbf{z}_n = \mathbf{E}_n^{-1}\mathbf{b}$ and $\mathbf{z}_n \to \mathbf{z}$ as $n \to \infty$, in practical situations n is bounded [4].

Let $\hat{m} = (T_k - T_{k-1})/\tau$, $j_{k-1} = (T_{k-1} - T_0)/\tau$ and Z_k be the solution for the instant T_k, defined by solving from time T_{k-1} to T_k using the Finite Difference discretization scheme on the fine grid [24] (for each time instant, of size τ) with initial values Z_{k-1} in T_{k-1} and right hand side vector

$\mathbf{b} = [\underline{b}(j_{k-1}+1)^T, \ldots, \underline{b}(j_{k-1}+\hat{m})^T]^T$. The solution of each coarse interval is given by:

$$F_1 \otimes Z_k = F_0^{\Delta} \otimes Z_{k-1} + S_k \tag{10}$$

where, $\otimes$ represents the Kronecker product [25], $F_0^{\Delta} = (F_0 F_1^{-1})^{\hat{m}-1} F_0 \in \mathbb{R}^{\hat{q}\times\hat{q}}$, $Z_0 = 0$, the matrices F_0 y F_1 as set in (5) and

$$S_k = \sum_{m=1}^{\hat{m}} (F_1^{-1}F_0)^{\hat{m}-m}[F_0 Z_{k-1} - \underline{b}(j_{k-1}+m)] \tag{11}$$

Imposing continuity, $F_1 \otimes Z_k - F_0^{\Delta} \otimes Z_{k-1} - S_k = 0$ on the instants T_k, for $1 \le k \le \hat{k}$, the system $\mathbf{CZ} = \mathbf{S}$ is obtained [9, 7]:

$$\underbrace{\begin{bmatrix} F_1 & & & \\ -F_0^{\Delta} & F_1 & & \\ & \ddots & \ddots & \\ & & -F_0^{\Delta} & F_1 \end{bmatrix}}_{\mathbf{C}} \underbrace{\begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_{\hat{k}} \end{bmatrix}}_{\mathbf{Z}} = \underbrace{\begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_{\hat{k}} \end{bmatrix}}_{\mathbf{S}} \tag{12}$$

The case where the coarse solution in T_k with initial data $Z_{k-1} \in \mathbb{R}^{\hat{q}}$ in T_{k-1} is obtained after using a Finite Differences step for the coarse time interval $G_1 Z_k = G_0 Z_{k-1}$ is considered, where the matrices $G_1 = (M + K\Delta T)$ and $G_0 = M \in \mathbb{R}^{\hat{q}\times\hat{q}}$ are defined.

A coarse grid propagator based on G_0 and G_1 is used in the Parareal algorithm to precondition (12) [9]. The coarse grid propagation system $\mathbf{Z}^{i+1} = \mathbf{Z}^i + \mathbf{E}_g^{-1}\mathbf{R}^i$ is defined as:

$$\underbrace{\begin{bmatrix} Z_1^{i+1} \\ Z_2^{i+1} \\ \vdots \\ Z_{\hat{k}}^{i+1} \end{bmatrix}}_{\mathbf{Z}^{i+1}} = \underbrace{\begin{bmatrix} Z_1^{i} \\ Z_2^{i} \\ \vdots \\ Z_{\hat{k}}^{i} \end{bmatrix}}_{\mathbf{Z}^{i}} + \underbrace{\begin{bmatrix} G_1 & & & \\ -G_0 & G_1 & & \\ & \ddots & \ddots & \\ & & -G_0 & G_1 \end{bmatrix}^{-1}}_{\mathbf{E}_g^{-1}} \underbrace{\begin{bmatrix} R_1^{i} \\ R_2^{i} \\ \vdots \\ R_{\hat{k}}^{i} \end{bmatrix}}_{\mathbf{R}^{i}} \tag{13}$$

for $0 \le i \le (n-1)$, where the residue $\mathbf{R}^i = \left[{R_1^i}^T, \ldots, {R_{\hat{k}}^i}^T\right] \in \mathbb{R}^{\hat{k}\hat{q}}$ in (13) is defined as $\mathbf{R}^i = \mathbf{S} - \mathbf{CZ}^i$, where $\mathbf{Z}^i = \left[{Z_1^i}^T, \ldots, {Z_{\hat{k}}^i}^T\right]^T \in \mathbb{R}^{\hat{k}\hat{q}}$ and $\mathbf{Z}^0 = [0^T, \ldots, 0^T]^T$. Each R_j^i vector stands for the i-th iteration of the residue, on the T_j time instant. Likewise, each Z_j^i vector stands for the i-th iteration of the solution, on the T_j time instant.

Now, $\mathbf{z}_n = \mathbf{E}_n^{-1}\mathbf{b}$ is defined. Let $\mathbf{z}_n$ be the nodal representation of a piecewise linear function $\underline{z}^n$ in the time dimension with respect to the fine space discretization parameterized by τ in $[t_0, t_f]$. Because $\mathbf{z}_n \in \mathbb{R}^{(\hat{l}+\hat{k}-1)\hat{q}}$ is continuous in each coarse subinterval $[T_{k-1}, T_k]$, the function $\underline{z}^n$ can be discontinuous on the points T_k, with $1 \le k \le \hat{k}-1$. On each $[T_{k-1}, T_k]$ subinterval, $\underline{z}^n$ is defined by solving from the instant T_{k-1} to the instant T_k using the Finite Differences scheme with fine time intervals τ and initial data Z_{k-1}^n in T_{k-1}. The equation that describes the solution on the fine intervals, starting from a coarse interval is:

$$F_1 \underline{z}^n(i+1) = F_0 \underline{z}^n(i) - \underline{b}(i+1); \text{ for } T_{k-1} \le t < T_k; \text{ y } \underline{z}^n(T_{k-1}) = Z_{k-1}^n. \tag{14}$$

The vector $\mathbf{z}_n$ is obtained computing (14) for $2 \le k \le \hat{k}$. With this algorithm the steps 3) and 5) from Algorithm 2 can be solved, and therefore it can find the product $\mathbf{Hs}$. In the program, the input vector $\mathbf{s}$ for Algorithm 2 will be each vector p_k from Algorithm 1, used on the outer iteration of the Conjugate Gradient.

III. Implementation

The user defines the spatial discretization size $\hat{q}$, the fine time discretization size $\hat{l}$, the coarse time discretization size $\hat{k}$, the initial condition y_0 and the target solution y^*.

To help the convergence rate of the outer Conjugate Gradient, an initial guess $\mathbf{u}_0$ is found through:

$$\mathbf{u}_0 = \mathbf{N}\mathbf{E}^{-T}\mathbf{Q}\mathbf{E}^{-1}\mathbf{f}_3 - \mathbf{N}^T\mathbf{E}^{-T}\mathbf{f}_1\,. \tag{15}$$

The program implemented on this work uses the `main` structure of Algorithm 3.

Algorithm 3 Main

Input: $\hat{q}, \hat{l}, \hat{k}, \hat{m}, y_0, y^*$
Output: $\mathbf{u}$

1: $[M, K, B] = finiteElements(\hat{q}, y_0, y^*)$ ▷ Call to the function that does the space discretization, as described in Section I
2: $[\mathbf{E}, \mathbf{Q}, \mathbf{N}, \mathbf{b}] = fineGrid(M, K, B, \hat{l})$ ▷ Call to the function that does the fine time discretization, as described in Section I
3: $[\mathbf{C}, \mathbf{E}_g] = coarseGrid(M, K, \hat{k}, \hat{m})$ ▷ Call to the function that does the coarse time discretization, as described in Section II-C
4: $[\mathbf{u}_0] = preconditioner(\mathbf{G}, \mathbf{E}, \mathbf{Q}, \mathbf{N}, \mathbf{b})$ ▷ Call to a function that calculates (15)
5: $[\mathbf{u}] = cg(\varepsilon, \mathbf{u}_0, max_i, \hat{k}, \hat{m}, \mathbf{G}, \mathbf{E}, \mathbf{Q}, \mathbf{N}, \mathbf{b}, \mathbf{C}, \mathbf{E}_g)$ ▷ Call to Algorithm 1

With the defined problem data, the finite elements matrices are generated. Next, the matrices of the time discretization are built, and the system (8) can be formulated. The matrices from the time discretization are considered as the fine grid matrices, because they have every time instant from the problem. Afterward, the coarse grid matrices are generated from the finite elements matrices. The coarse grid matrices are needed for the application of the Parareal method, as shown in equations (12) and (13).

With all the matrices created, the Conjugate Gradient method is executed to resolve $\mathbf{Hu} = \mathbf{b}$. On each Conjugate Gradient's iteration i, the product $\mathbf{H}p_i$ must be computed as described in Algorithm 1. To perform the product of matrix $\mathbf{H}$ by a vector, the Algorithm 2 is called. The steps 3) and 5) form Algorithm 2 are solved using the Parareal method. When the product $\mathbf{H}p_i$ is computed, the outer Conjugated Gradient's execution resumes.

For instance, to appreciate the benefits of the Parareal method, consider a space discretization grid with $\hat{q} = 2{\times}2 = 4$ elements, a fine time discretization with $\hat{l} = 10000$ time instants and a coarse time discretization with $\hat{k} = 10$ time instants. As a consequence, each process gets $\hat{m} = 1000$ time instants of the fine grid. For this example, the solution of the system $\mathbf{Ez} = \mathbf{b}$ involves $4 \cdot 10 \cdot 1000 \times 4 \cdot 10 \cdot 1000$ equations and variables, while the approximation $\mathbf{E}_n\mathbf{z}_n = \mathbf{b}$ is a linear system of only $4 \cdot 10 \times 4 \cdot 10$ equations and variables.

IV. Algorithms

The pseudocode of the function used for the Parareal method (Algorithm 4) and its dependences are presented next, given that its implementation is the main contribution of this work. Besides, Algorithm 4 shows how the message passing is managed to maintain a low communication cost among the parallel processes.

The first pseudocode presented corresponds to the Parareal method. The same naming conventions as in Section II-C are used. The input parameters for the `Parareal` function are: the vector $b = \mathbf{b}$, the matrix $E_g = \mathbf{E}_g$, the matrix $C = \mathbf{C}$, the vector of the initial approximated solution Z_0, the coarse intervals count $\hat{k}$, the fine intervals by coarse interval count $\hat{m}$ and the error tolerance ε.

The output of the `Parareal` function is an approximation to $z = \mathbf{z} \leftarrow \mathbf{E}^{-1}\mathbf{b}$ as described on Section II-C. Algorithm 4 calls the functions `fineSolver` (Algorithm 5) and `marching` (Algorithm 6) to be next described in this section.

Algorithm 4 Parareal

Input: $b, E_g, C, \hat{k}, \hat{m}, Z_0, \varepsilon$
Output: y
1: $S \leftarrow fineSolver(b, \hat{k}, \hat{m})$ ▷ Call to Algorithm 5
2: $Z \leftarrow Z_0$
3: $R \leftarrow S$ ▷ $\mathbf{R}^1 \leftarrow \mathbf{S} - \mathbf{CZ}^0$, communication of S^k to the next process
4: **while** $\|r_i\| > \varepsilon$ **do**
5: $coarse \leftarrow E_g^{-1}R$ ▷ $aux \leftarrow \mathbf{E}_g^{-1}\mathbf{R}^i$
6: $Z \leftarrow Z + coarse$ ▷ $\mathbf{Z}^{i+1} \leftarrow \mathbf{Z}^i + aux$
7: $R \leftarrow S - C \times Z$ ▷ $\mathbf{R}^{i+1} \leftarrow \mathbf{S} - \mathbf{CZ}^i$
8: **end while**
9: $y \leftarrow marching(b, x_{i-1}, \hat{k}, \hat{m})$ ▷ Call to Algorithm 6
10: **return** y

Algorithm 5 shows how the jumps vector $\mathbf{S}$ is generated according to equation (11), that saves only the final elements of the coarse time interval.

Algorithm 5 fineSolver

Input: $b, \hat{k}, \hat{m}$
Output: S
1: **for all** $k < \hat{k}$ **do** ▷ parallel loop, distributed in $\hat{k}$ processes
2: $s \leftarrow \vec{0}$
3: **for all** $i < \hat{m}$ **do** ▷ local loop, calculated on each process
4: $s \leftarrow F_1^{-1}(F_0 s - b(k,i))$ ▷ Equation (10)
5: **end for**
6: $S(k) = s$
7: **end for**
8: **return** S

With this information the iterative loop of the Parareal algorithm is performed, the loop computes vector $\mathbf{Z}^i$ iteratively, as indicated in equation (13), until the solution of the coarse grid $\mathbf{Z}^n$ is found, when the required tolerance is reached.

After finding the coarse solution, the function `marching` is called, so that each process can extend its initial coarse solution to their own fine time intervals $\underline{z}^n$. Joining the solution of every process, the approximated general solution $\mathbf{z}_n$ is found.

Algorithm 6 marching

Input: $b, coarse, \hat{k}, \hat{m}$
Output: y
1: **for all** $k < \hat{k}$ **do** ▷ parallel loop, distributed in $\hat{k}$ processes
2: $z \leftarrow coarse(k)$
3: **for all** $i < \hat{m}$ **do** ▷ local loop, calculated on each process
4: $z \leftarrow F_1^{-1}(F_0 z - b(i,k))$ ▷ Equation (14)
5: $y(k,i) = z_i$
6: **end for**
7: **end for**
8: **return** y

The functions `fineSolver` and `marching` are similar, because both solve the problem on the coarse time intervals. The `fineSolver` function finds its fine grid solution to calculate the final coarse instants (used as a preconditioner for the Parareal). The function `marching` finds the fine grid solution given an initial condition $\mathbf{Z}^n$ (the coarse grid solution), to complete the global solution.

As it was mentioned previously, some special attention is needed when a process requires some data that belongs to another process. The algorithms were designed to reduce the data communication between processes. With the proposed solution, the data communication between processes is needed only on the `Parareal` function, when the coarse data grid is propagated according to equation (13). Next, the experimental results of the implementation are presented.

V. Numerical experiments

The results of the experiments using the implementation of the Parareal method are presented in this Section. The experiments are based on the reference paper [7], used for validation.

The hardware used is a cluster of four Dell PowerEdge R710 nodes, with 2 processors of 4 cores Intel Xeon E5530 of 2.4GHz, Intel 5530 chipset, 8GB DDR3 of 1066 MHz RAM memory, connected in a Giga-Ethernet (1Gbps) LAN, as shown in Figure 1.

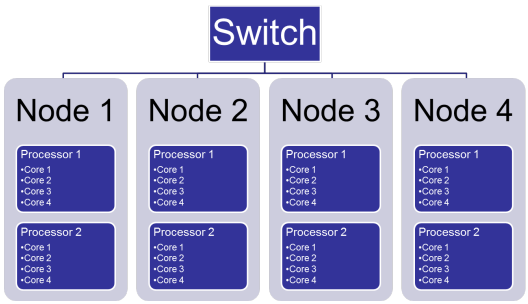


Fig. 1. Experimental platform

In this set of experiments, there are at most 4 nodes used and, for simplicity, only the first processor of each node is used. Therefore, only 4 cores per node are used.

A. Definitions

The optimal control problem to be solved for the experiments is given by the following 2D heat equation:

$$\begin{cases} z_t - z_{xx} = v, & x \in \Omega, \quad 0 < t \\ z(t,0) = 0, & x \in \partial\Omega, \quad 0 \leq t \\ z(0,x) = 0, & x \in \partial\Omega\,, \end{cases} \tag{16}$$

where $\Omega = [0,1] \times [0,1]$. The selected target function is $y^*(x_1,x_2) = x_1(1-x_1)e^{-x_1}x_2(1-x_2)e^{-x_2}$ for $t \in [0,1]$. The selected problem, the problem sizes and considerations are used as in the reference paper [7].

As naming conventions for the experiments, pCG(n,p) represents the execution of the modified Conjugate Gradient, using the Parareal method, for n nodes, each with p parallel processes. In all the experiments, a single process is run in each core. Times values are shown in seconds for all tables.

For each comparative table all tests are remade, so there may be some time differences in different tables that run the same setup of pCG(n,p). Those differences are unavoidable [26], but the variations are small in general, so they are acceptable anyway.

It should be mentioned that in the conducted experiments, the peak FLOPS / average FLOPS ratio [27] was not larger than 1.06 in any experiment.

B. Validation

The implementation of the pCG(n,p) of this work is compared to the reference work (IFOM) [7]. As [7] is a theoretical work about Parareal, there is no execution time; on the contrary, there are only data about the required iterations needed for the resolution of the problems. The values of Table I are given in the format $iter_e(iter_i)$, where $iter_e$ is the external iteration count (Conjugate Gradient) and $iter_i$ is the inner iteration count (Parareal).

TABLE I
Comparision of iterations of [7]'s implementation and our implementation pCG(n,p), for inner tolerance values ε_i, outer tolerance $\varepsilon_o = 10^{-6}$, $\alpha = 1$, $\beta = 12$, $\gamma = 10^{-5}$, inner grid size 13×13, $\tau = 1/512$, $\hat{k} = 32$, $\Delta T/\tau = 16$ and $n.c.$ means that the system does not converge in 100 iterations.

ε_i	IFOM	pCG$(1,p)$		
		$p=1$	$p=2$	$p=4$
10^{-12}	16(586)	16(586)	16(580)	16(578)
10^{-10}	17(510)	17(510)	17(502)	17(504)
10^{-8}	18(442)	18(442)	18(414)	18(424)
10^{-7}	18(362)	18(362)	18(364)	20(404)
10^{-6}	21(338)	21(338)	21(342)	19(340)
10^{-5}	24(274)	24(274)	24(280)	n.c.
10^{-4}	28(220)	28(220)	63(312)	43(258)

Table I shows that the iteration count obtained through the implementation of pCG$(1,1)$ is consistent with the theoretical results expected from [7]. Some other tests were also made, which compare the solutions u_{IFOM} of IFOM and u_{pCG} of pCG(n,p), and the error $\epsilon = \|u_{IFOM} - u_{pCG}\|/\|u_{IFOM}\|$ was smaller than 10^{-6} on all the cases. Furthermore, checking the execution of the Parareal solver on pCG$(1,1)$, the error of each iteration's result was less than 10^{-6}.

C. Efficiency

The following concepts are used for the efficiency analysis of the implementation. *Strong Scaling* is defined as the variation of the resolution time as the number of processes changes, while having a fixed problem size[2] [28]. *Weak Scaling* is defined as the variation of the resolution time as the number of processes changes, while having a fixed problem size per process[3] and therefore, the problem size is proportional to the number of processes [28].

The product $c = n \cdot p$ represents the total processes used, remembering that n is the amount of nodes used and p is the amount of processes run per node. Considering the strong scaling, an increase in the number of processes c, decreases the problem size in each process. Conversely, considering the weak scaling, an increase in the number of processes c, increases the total problem size.

A well recognized metric used to describe the scaling of a program is the parallelism efficiency. The absolute efficiency of the parallelism is:

$$e_c = \frac{t_s}{ct_c} \tag{17}$$

[2]Related to the Amdahl's Law[29].
[3]Related to the Gustafson's Law [30].

where e_c is the absolute efficiency of the parallelism in c processes, t_s is the execution time of the best known serial solution and t_c is the execution time of the program using c processes [28]. In this analysis, the best known serial solution is the one developed on this work, so the equality $t_s = t_1$ is used.

Another metric proposed in this work is the relative efficiency when the number of processes doubles ϵ_c, this is defined as:

$$\epsilon_c = \begin{cases} (e_c)^{1/log_2 c}, & c > 1 \\ 1, & c = 1 \end{cases} \quad (18)$$

The only values used for the experiments were $\epsilon_1 = 1$, $\epsilon_2 = e_2$, $\epsilon_4 = (e_4)^{1/2}$, $\epsilon_8 = (e_8)^{1/3}$ and $\epsilon_{16} = (e_{16})^{1/4}$.

1) Strong scaling: The absolute efficiency for the strong scaling is calculated as:

$$e_c = \frac{time(\text{pCG}(1,1))}{c \cdot time(\text{pCG}(n,p))}. \quad (19)$$

The problem size for the first test is $\hat{q} = 13{\times}13$ and $\hat{l} = 512$, as it is used in [7]. To illustrate the problem size for this configuration, the dimension of matrix $\mathbf{G}$ is 199680×199680 and the dimension of matrix $\mathbf{E}$ is 86528×86528.

TABLE II
TIMES OF PCG$(1,p)$, FOR DIFFERENT INNER TOLERANCE ε_i VALUES, OUTER TOLERANCE $\varepsilon_e = 10^{-6}$, $\alpha = 1$, $\beta = 12$, $\gamma = 10^{-5}$, INNER GRID SIZE 13×13, $\tau = 1/512$, $\hat{k} = 32$ AND *n.c.* MEANS THAT THE SYSTEM DOES NOT CONVERGE IN 100 ITERATIONS.

ε_i	pCG$(1,p)$		
	$p=1$	$p=2$	$p=4$
10^{-12}	9.075293	5.867213	4.406441
10^{-10}	9.0619	5.964406	4.398958
10^{-8}	8.99508	5.988965	4.353884
10^{-7}	8.826598	5.693643	4.540562
10^{-6}	9.293346	6.345327	4.236906
10^{-5}	9.908053	6.485624	n.c.
10^{-4}	10.695609	14.248947	7.465788

Table II shows the execution times of the experiments presented in Table I. Fixing the values of p, it can be noticed that the time values do not differ greatly until an inner tolerance of 10^{-5}. When the required precision of the inner solver (Parareal) is increased, the inner iterations count increases, while the outer iterations count (Conjugate Gradient) decreases. Although there are less iterations needed for the low precision cases, the external iterations count increase leads to greater execution times. For a case with inner tolerance of 10^{-5}, the algorithm does not converge. At the same time, for lower precision cases the convergence rate of the outer Conjugate Gradient is lowered.

The inner tolerance value of 10^{-6} is used as a reference, because it is coherent with the outer tolerance value of 10^{-6}. Therefore the base tolerance value of 10^{-6} is chosen for the following tests.

To build Table III, the absolute efficiency e_c of the parallelism is calculated from the data of Table II, following (19).

TABLE III
EFFICIENCY OF PCG$(1,p)$, FOR DIFFERENT INNER TOLERANCE ε_i VALUES, OUTER TOLERANCE $\varepsilon_e = 10^{-6}$, $\alpha = 1$, $\beta = 12$, $\gamma = 10^{-5}$, INNER GRID SIZE 13×13, $\tau = 1/512$, $\hat{k} = 32$ AND *n.c.* MEANS THAT THE SYSTEM DOES NOT CONVERGE IN 100 ITERATIONS.

ε_i	pCG$(1,p)$		
	$p=1$	$p=2$	$p=4$
10^{-12}	1	0.77339	0.51489
10^{-10}	1	0.75966	0.51500
10^{-8}	1	0.75097	0.51650
10^{-7}	1	0.77513	0.48599
10^{-6}	1	0.73230	0.54836
10^{-5}	1	0.76384	n.c.
10^{-4}	1	0.37531	0.35815

The efficiency values from Table III show that the average relative efficiency for a single cluster node is approximately $\epsilon = 0.73$ each time the number of processes doubles.

The next step is to calculate the efficiency for multiple cluster nodes. In order to have a viable test, the spatial grid must be larger. The total execution times and the relative efficiency ϵ are calculated with a grid of $\hat{q} = 19{\times}19$. Figure 2 is obtained using the average efficiency of the different inner tolerances $\varepsilon_i = \{10^{-12}, 10^{-10}, 10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}\}$.

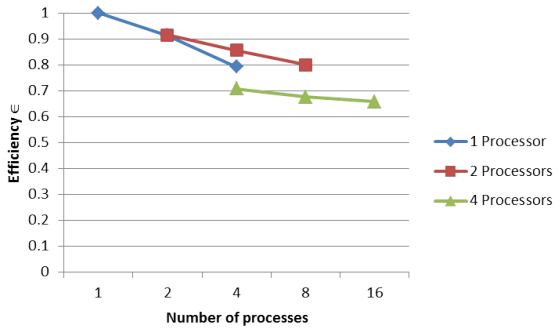


Fig. 2. Average strong scaling (Total time)

The average strong scaling shown in Figure 2 is approximately $\epsilon = 0.7$ each time the number of processes doubles. Considering only the times of the solver (without taking into account the time used to build the matrices and the execution of the external Conjugate Gradient preconditioner), the main contribution of this work can be noticed. The Figure 3 shows the times of the Parareal solver.

On this test, the average relative efficiency on a single cluster node is approximately $\epsilon = 0.79$, that is higher than the one calculated on Table III. This indicates that the strong scaling efficiency increases as the problem size grows.

The next test for the strong scaling is for a big sized problem, that cannot be solved in a single cluster node, and that is near the size limit that two nodes can solve.

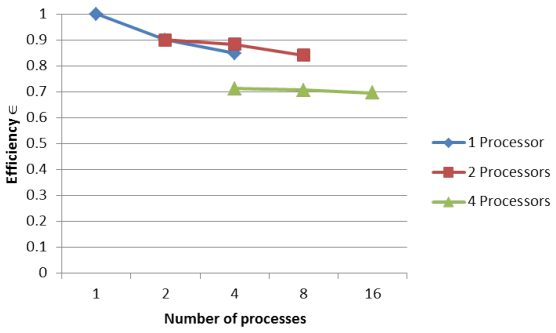


Fig. 3. Average strong scaling (considering only the solver)

TABLE V
EFFICIENCY OF PCG$(1,p)$, INNER TOLERANCE $\varepsilon_i = 10^{-6}$ VALUES, OUTER TOLERANCE $\varepsilon_e = 10^{-6}$, $\alpha = 1$, $\beta = 12$, $\gamma = 10^{-5}$, INNER GRID SIZE 19×19, $\hat{l} = 8192$, AND *o.o.m.r.* MEANS THAT THE SYSTEM RUNS OUT OF MEMORY DURING THE RESOLUTION OF THE PROBLEM.

$\hat{k}$	pCG(n,p)					
	$n=2$			$n=4$		
	$p=1$	$p=2$	$p=4$	$p=1$	$p=2$	$p=4$
512	o.o.m.r.	o.o.m.r.	o.o.m.r.	1	0.96257	0.55323
256	1	0.80291	0.49520	0.54950	0.53390	0.32647
128	1	0.97602	0.43423	0.53439	0.51296	0.31407
64	1	0.87475	0.43981	0.54046	0.50359	0.30965
32	1	0.89704	0.42091	0.51832	0.48340	0.29379
16	1	0.82382	0.44782	0.44917	0.42319	0.25103

The values $\hat{q} = 19 \times 19$ and $\hat{l} = 8192$ are chosen. With these values, the size of matrix $\mathbf{E}$ is 2957312×2957312 (approximately $3 \cdot 10^6 \times 3 \cdot 10^6 = 9 \cdot 10^{12}$ elements) and the size of matrix $\mathbf{G}$ is 6537216×6537216 (approximately $6.54 \cdot 10^6 \times 6.54 \cdot 10^6 \approx 4.3 \cdot 10^{13}$ elements).

TABLE IV
TIMES OF PCG$(1,p)$, INNER TOLERANCE $\varepsilon_i = 10^{-6}$ VALUES, OUTER TOLERANCE $\varepsilon_e = 10^{-6}$, $\alpha = 1$, $\beta = 12$, $\gamma = 10^{-5}$, INNER GRID SIZE 19×19, $\hat{l} = 8192$, AND *o.o.m.r.* MEANS THAT THE SYSTEM RUNS OUT OF MEMORY DURING THE RESOLUTION OF THE PROBLEM.

$\hat{k}$	pCG(n,p)					
	$n=2$			$n=4$		
	$p=1$	$p=2$	$p=4$	$p=1$	$p=2$	$p=4$
512	o.o.m.r.	o.o.m.r.	o.o.m.r.	506.84255	263.27448	229.03597
256	958.6446	596.9837	483.9645	436.1471	224.44206	183.52355
128	919.7547	471.1759	529.52744	430.2823	224.12969	183.02848
64	926.5926	529.6313	526.7008	428.6131	229.998	187.0255
32	939.2415	523.5229	557.8653	453.0177	242.8749	199.8105
16	922.1194	559.6586	514.7818	513.2409	272.3714	229.5808

It is not possible to compute the efficiency in a classic sense for Table IV, given that there is no serial solution that can solve the proposed problem cases as the system runs out of memory when the matrices are built in a single node. Therefore, to obtain the efficiency, the base execution time may be considered as the one that solves the problem with the least amount of cluster nodes, and the least amount of cores of each node. In the tests shown in Table IV, for $\hat{k} = 512$ and $\hat{m} = 16$ the base case will be pCG$(4,1)$ and for all the other tested conditions of $\hat{k}$ and $\hat{m}$, it will be pCG$(2,1)$.

Considering the data of Table V, the average relative efficiency can be established as $\epsilon = 0.7$ each time the number of processes doubles.

From the data presented, the general strong scaling obtained is approximately $\epsilon = 0.7$ each time the number of processes doubles.

As a noteworthy detail to keep in mind, in general, the efficiency decays faster when the number of nodes increases than when the number of processes per node increases. This occurs because the LAN's connection to the new nodes adds latency to the computations and has a lower data transfer rate than the local bus on each node. In general, the LAN's data transfer rate is not enough to keep the same efficiency in processes on different nodes as compared to processes on a same node. There is also a bus bandwidth from the central memory in each node that limits the efficiency of increasing the number of processes per node, as adding more processes will decrease the bandwidth available for each process after a certain point. The limit will depend on the equipment specifications.

2) Weak scaling: The efficiency for the weak scaling is computed as:

$$e_c = \frac{time(\text{pCG}(1,1))}{time(\text{pCG}(n,p))}. \tag{20}$$

Given that the total problem size grows as the number of used processes c does, there is no need to multiply the denominator by $c = n \cdot p$.

To compute the weak scaling, the problem size per process must be fixed. As a first experimental option, the number of elements from vector $\mathbf{u}$ can be fixed; this is, the size of the solution found by each process $\hat{q} \cdot \hat{m}$ is constant and the number of coarse intervals $\hat{k}$ is shifted to obtain several configurations. Let $\hat{q} = 19 \times 19$, $\hat{m} = 32$ be the fixed size per process, when the total execution times are measured, the relative efficiency ϵ is computed. Figure 4 shows the average of the relative efficiency for the coarse instants per node $\hat{k}/n = \{1, 2, 4, 8, 16\}$.

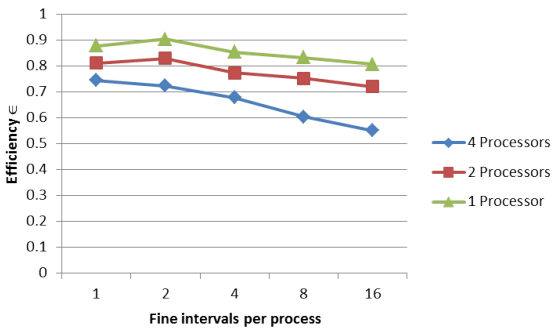


Fig. 4. Average weak scaling (Total time)

The average weak scaling form Figure 4 is approximately $\epsilon = 0.75$ each time the number of processes doubles. Considering only the execution time of the function pCG(n,p), the

weak scaling is approximately $\epsilon = 0.85$ each time the number of processes doubles, with an increasing efficiency as the problem size increases, as it can be observed in Figure 5 (this is a meaningful improvement with respect to the efficiency of the whole program, when the scaling efficiency drops to $\epsilon = 0.29$ for the pCG$(4, 4)$ with $\hat{k} = 256$).

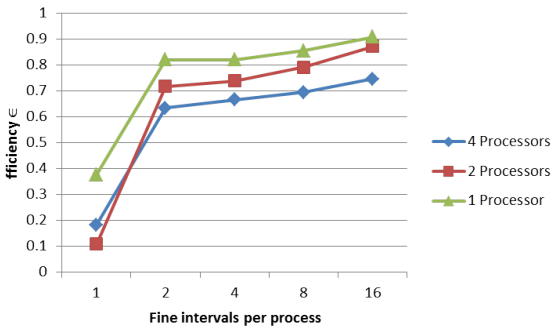


Fig. 5. Average weak scaling (considering only the solver)

The problem with the pCG's preconditioner is that it uses standard PETSc operations to solve the inverse of matrix $\mathbf{E}$. In particular for the test pCG$(4, 4)$ with $\hat{k} = 256$, the time to build the matrices is 9.21 seconds, 332 seconds for the preconditioner and 48.7 seconds for the pCG function. This clearly shows that solving a single time the inverse of matrix $\mathbf{E}$ by the standard means is much slower than the multiple solutions done by the pCG function.

The problem with the first consideration for the fixed problem size is that the matrices involved on the problem have their size squared compared to the solution vector. Then, as a second experimental option, the fixed data size per process is the size of the matrices stored on each process. With this approach, and setting the parameters $\hat{q} = 19 \times 19$ and $\hat{m} = 32$, each execution time is measured, then the relative efficiency ϵ is computed. The average of the relative efficiency for the coarse instants per node $\hat{k}/n = \{1, 2, 4, 8, 16\}$ is shown in Figure 6.

A detail to have in mind is that to allow the size of the matrices to be constant, on all the tests, the only possible combinations of n and p are those that make $c = n \cdot p$ a perfect square. This occurs because $\hat{k}$ must be divisible by c and the matrices's sizes are proportional to the square of $\hat{k}$. Therefore, the number of elementes of matrix $\mathbf{E}$, $size(\mathbf{E})$ will be set as reference.

Most iterative Krylov subspace methods have a computational complexity of $\mathcal{O}(sol_s \cdot iter)$ where sol_s is the solution size and $iter$ the number of iterations needed for the convergence of the algorithm [31]. The number of iterations using the Conjugate Gradient method is bounded by $1 \leq iter \leq sol_s$ because it can be used as a direct method [32]. The first experimental option tests the lower $iter$ bound and the second experimental option tests the upper $iter$ bound, because the number of iterations needed for the convergence is unknown before the execution of the solver.

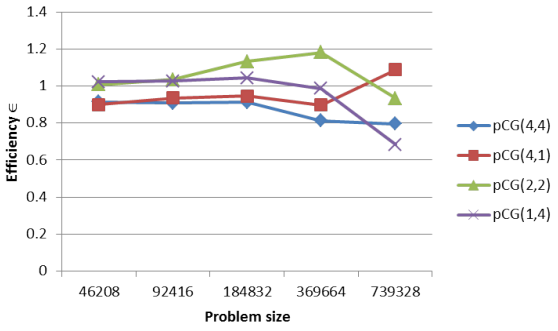


Fig. 6. Average weak scaling (Total time)

Because the number of iterations the solver used to converge in each experiment is greater than 1 and less than $\hat{lp}$ (equal to sol_s), the first experimental option will give a scaling less than 1, and the second experiment can give a scaling greater than 1. This can be seen in Figure 6, where in some cases the scaling is greater than 1.

On average, for the second experimental option, the relative efficiency of the weak scaling is approximately $\epsilon = 0.96$ each time the number of processes doubles.

The same process as in the first experimental option is used for the solver time, Figure 7 shows the weak scaling of the pCG function.

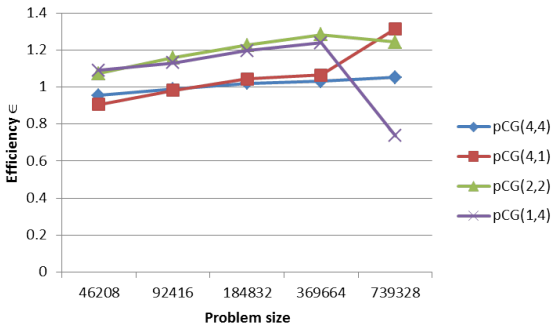


Fig. 7. Average weak scaling (considering only the solver)

It can be noticed in Figure 7 that most of the resolution times (not considering the preconditioner) scale on a supraliear way. In this context, the average relative efficiency is approximately $\epsilon = 1.09$ each time the number of processes doubles.

On both weak scaling considerations, the efficiency lost as the number of nodes increases is greater than the efficiency lost as the number of cores used per node increases. This is expected as it was analyzed on the strong scaling, due to the data transfer among cores on a single node is faster than the transfer among cores on different nodes, because the data has to travel through the switch holding the LAN.

The analysis continues comparing the problem resolution using or not the Parareal method.

D. Parareal vs No Parareal

To solve the external Conjugate Gradient, the steps 3) and 5) from Algorithm 2 must be solved by some iterative method. Three options to solve those steps are compared, *a*) the Parareal implementation done for this work *b*) the PETSc's implementation of the Conjugate Gradient, and *c*) the PETSc's implementation of GMRES.

The parameters used for the test are $\hat{q} = 9 \times 9$, $\hat{l} = 256$, $\hat{m} = 16$, while the values $\hat{k} = \{16, 32\}$ are considered. The Conjugate Gradient did not converge for any test case from this set, therefore only the Parareal and GMRES methods are compared.

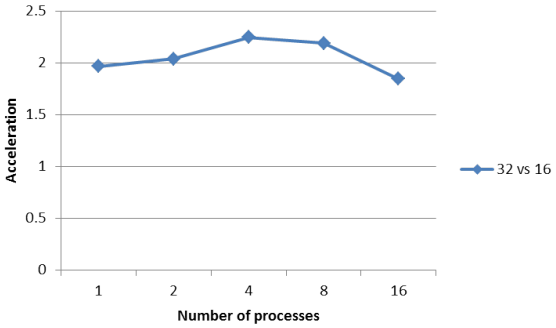


Fig. 8. Acceleration of Parareal vs GMRES (Total time)

The quotient from the resolution time of the GMRES and the the Parareal methods gives the speed-up for the cases $\hat{k} = 16$ and $\hat{k} = 32$. Then, the quotient from the cases $\hat{k} = 32$ and $\hat{k} = 16$ yield the acceleration obtained as the problem size doubles. This is presented in Figures 8 and 9.

Figures 8 and 9 show that the execution of the Parareal takes less time than the execution of the GMRES for every problem size tested, since the acceleration is higher than 1 for every case. Indeed, when the problem size increases the quotient from the execution time of the GMRES and the execution time from the Parareal increases with an average of 2 when the solution size doubles.

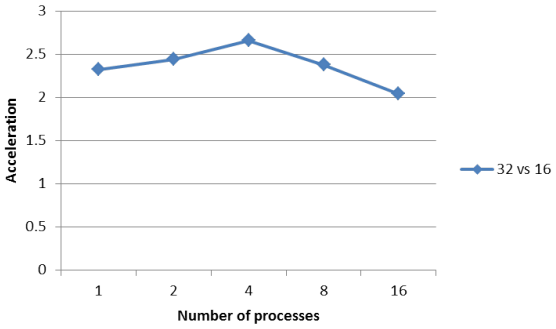


Fig. 9. Acceleration of the Parareal vs GMRES (Solver only)

As hinted, a considerable time difference was expected, after analysing the strong scaling tests, and for that reason the problem sizes for the test were small. The acceleration shown in Figures 8 and 9 are from linear systems where the size of matrix $\mathbf{E}$ are 20736×20736 (when $\hat{k} = 16$) and 41472×41472 (when $\hat{k} = 32$) respectively.

The main time load for the preconditioner of the external Conjugate Gradient is the solution of a system $\mathbf{E}x = b$. With the execution time of the preconditioner ($t_{precond}$), the resolution time of the system using GMRES (t_{GMRES}) can be estimaded by $t_{precond} \cdot iter_{cg} \approx t_{GMRES}$, where $iter_{cg}$ is the iterations needed for the convergence of the pCG (even if it uses the Parareal, as the iteration count for the external Conjugate Gradient is stable for every inner method tested).

The time approximation variation var (of the approximated value versus the measured value) found from the tests of this section, calculated by $var = |t_{GMRES} - t_{precond} \cdot iter_{cg}|/t_{GMRES}$, is at most 20%.

To emphasize the time reduction achieved using the Parareal method, the time of pCG$(4, 4)$ with $\hat{q} = 19 \times 19$ and $\hat{l} = 8196$ is analyzed. In this case, building the matrices takes 9.21 seconds, the external preconditioner takes 332 seconds and the pCG function takes 48.7 seconds (with 16 iterations). Solving this pCG$(4, 4)$ case with the GMRES solver would take $332 \cdot 16 - 48.7 = 5263$ seconds more (almost one hour and a half compared to 49 seconds when using the Parareal).

E. Parallelization suitability

When the strong scaling was analyzed, it was found that an increase in the problem size would give a better parallelization efficiency on each node.

Extending that idea, on a sufficiently small problem the cost of creating new processes and the communication costs among the processes would be higher than the time benefit obtained with the parallelization. The parallelization is convenient when the solution of the pCG(1,1) takes longer than every other analyzed solution of pCG(n, p). The tests are done with grid sizes that gives problem sizes of approximately twice as big each time.

Table VI is used to find the convenient problem size for the parallelization on a single node.

TABLE VI
TIMES OF PCG(n, p), OUTER TOLERANCE $\varepsilon_e = 10^{-6}$, INNER TOLERANCE $\varepsilon_i = 10^{-6}$, $\alpha = 1$, $\beta = 12$, $\gamma = 10^{-5}$, $\hat{l} = 512$ AND $\hat{k} = 32$.

$\hat{q}$	pCG(n, p)		
	$n = 1$		
	$p = 1$	$p = 2$	$p = 4$
4×4	1.6396871	2.6615668	2.7016241
7×7	2.434376	2.5144259	2.447761
9×9	3.673742	3.319042	2.524157

The convenient minimum problem size on a single node for matrix $\mathbf{E}$ is 41472×41472 elements ($\hat{q} = 9 \times 9$ and $\hat{l} = 512$). In a similar fasion, multiple nodes are analyzed on Table VII.

Table VII shows that, in general, the parallelization is suitable starting on a matrix $\mathbf{E}$ of 86528×86528 elements ($\hat{q} = 13 \times 13$ and $\hat{l} = 512$).

TABLE VII
TIMES OF PCG(n,p), OUTER TOLERANCE $\varepsilon_e = 10^{-6}$, INNER TOLERANCE $\varepsilon_i = 10^{-6}$, $\alpha = 1$, $\beta = 12$, $\gamma = 10^{-5}$, $\hat{l} = 512$ AND $\hat{k} = 32$.

$\hat{q}$	pCG(n,p)					
	$n=2$			$n=4$		
	$p=1$	$p=2$	$p=4$	$p=1$	$p=2$	$p=4$
7×7	3.012665	2.7350261	2.472511	3.914736	3.915797	3.9322839
9×9	3.723882	2.760471	2.400935	3.842674	3.819844	4.711632
13×13	6.70257	4.045551	2.946394	6.058765	4.413698	4.337244

VI. CONCLUSIONS

- The experiments of Section V-D shows that the execution of the program using the Parareal method is considerably faster than executions that use the CG or GMRES methods. Not only it does perform better, but it accelerates for larger problem sizes.
- The experiments of Section V-C present a relative efficiency of around $\epsilon = 0.7$ each time the number of processes doubles for the strong scaling. At the same time, for the weak scaling, the relative efficiency is $\epsilon = 0.75$ each time the number of processes doubles for a constant solution size per process, and $\epsilon = 0.96$ each time the number of processes doubles for a constant data size per process.
- The experiments of Section V-C find that for the used hardware (described in Section V), the parallelization begins to be convenient for solution size of 40000 elements.

In summary, this paper presented a parallel efficient alternative in PETSc to solve a parabolic optimal control problem using the Parareal mthod. Experimental results above summarized demonstrate the advantages of this proposal over classical methods as the Conjugate Gradient and GMRES ones in a computing cluster.

REFERENCES

[1] J. Carlsson, "Optimal Control of Partial Differential Equations in Optimal Design," PhD Dissertation, KTH, School of Computer Science and Communication (CSC), Numerical Analysis and Computer Science, NADA, 2008.

[2] G. Biros and O. Ghattas, "Parallel Lagrange-Newton-Krylov-Schur methods for PDE-Constrained optimization I. The Krylov-Schur solver," *SIAM Journal on Scientific Computing*, no. 27, p. 687–713, 2005.

[3] J. L. Lions, *Optimal control of systems governed by partial differential equations*. Springer, 1971.

[4] T. P. Mathew, M. Sarkis, and C. E. Schaerer, "Analysis of block parareal preconditioners for parabolic optimal control problems," *SIAM*, no. 32, pp. 1180–1200, 2010.

[5] J. J. C. Silva, "Implementación de un esquema de paralelización temporal distribuida usando petsc," Diploma thesis, Ciencias y Tecnología - Universidad Católica Nuestra Señora de la Asunción, feb 2014.

[6] G. Strang, *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 1998.

[7] X. Du, M. Sarkis, C. E. Schaerer, and D. Szyld, "Inexact and truncated Parareal-in-time Krylov subspace methods for parabolic optimal control problems," *Electronic Transactions on Numerical Analysis*, 2013.

[8] J. Fritz, *Partial Differential Equations*. Springer, 1982.

[9] J.-L. Lions, Y. Maday, and G. Turinici, "Résolution d'edp par un schéma en temps "pararéel"," *C. R. Acad. Sci. Paris Sér. I Math.*, vol. 332, no. 7, pp. 661–668, 2001.

[10] M. Benzi, G. H. Golub, and J. Liesen, "Numerical solution of saddle point problems," *Acta Numerica*, no. 14, pp. 1–137, 2005.

[11] T. Rees, M. Stoll, and A. Wathen, "All-at-once preconditioning in PDE-constrained optimization," *Kybernetika*, no. 46, p. 341–360, 2010.

[12] P. Neittaanmäki and D. Tiba, *Optimal Control of Nonlinear Parabolic Systems*. Marcel Dekker, Inc., 1994.

[13] O. Bashir, K. Willcox1, O. Ghattas, B. van Bloemen Waanders, and J. Hill, "Hessian-based model reduction for large-scale systems with initial condition inputs," *International Journal for Numerical Methods in Engineering*, vol. 73, no. 6, pp. 844–868, 2008.

[14] T. P. Mathew, *Domain Decomposition Methods for the Numerical Solution of Partial Differential Equations*. Springer, 2008.

[15] W. Samyono, "Hessian matrix-free Lagrange-Newton-Krylov-Schur-Schwarz methods for elliptic inverse problems," PhD Dissertation, Old Dominion University, 2006.

[16] J. Gallier, "The Schur Complement and Symmetric Positive Semidefnite (and Defnite) Matrices," *Penn Engineering*, 2010.

[17] Y. Saad, *Iterative methods for sparse linear systems*. SIAM, 2003.

[18] R. D. Falgout and U. M. Yang, "hypre: a library of high performance preconditioners," *Numerical Solution of Partial Differential Equations on Parallel Computers*, 2002.

[19] M. J. Gander and S. Vandewalle, "Analysis of the parareal time-parallel time-integration method," *SIAM Journal on Scientific Computing*, no. 29, pp. 556–578, 2007.

[20] C. E. Schaerer and E. Kaszkurewicz, "The shooting method for the solution of ordinary differential equations: A control-theoretical perspective," *International Journal of Systems Science*, vol. 32, no. 8, 2001.

[21] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, 3rd ed. Springer-Verlag, 2002.

[22] E. H.-L. Liu, "Fundamental Methods of Numerical Extrapolation With Applications," *MIT Open Course Ware*, 2006.

[23] F. Zhang, "The Schur Complement and Its Applications," *Springer*, 2005.

[24] C. E. Schaerer, E. Kaszkurewicz, and N. Mangiavacchi, "A Multilevel Schwarz Shooting Method for the solution of the Poisson Equation in Two Dimensional Incompressible Flow Simulations," *Applied Mathematics and Computation*, vol. 153, no. 3, pp. 803–831, 2004.

[25] M. S. Gockenbach, *Partial differential equations: analytical and numerical methods*. SIAM, 2002.

[26] S. Balay, W. D. Gropp, L. C. McInnes, and B. F. Smith, "Efficient management of parallelism in object oriented numerical software libraries," in *Modern Software Tools in Scientific Computing*, E. Arge, A. M. Bruaset, and H. P. Langtangen, Eds. Birkhäuser Press, 1997, pp. 163–202.

[27] S. Balay, J. Brown, K. Buschelman, V. Eijkhout, W. D. Gropp, D. Kaushik, M. G. Knepley, L. C. McInnes, B. F. Smith, and H. Zhang, "PETSc users manual," Argonne National Laboratory, Tech. Rep. ANL-95/11 - Revision 3.3, 2012.

[28] A. Kaminsky, *BIG CPU, BIG DATA: Solving the World's Toughest Computational Problems with Parallel Computing*. Rochester Institute of Technology, 2013.

[29] G. M. Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities," *AFIPS spring joint computer conference*, 1967.

[30] J. L. Gustafson, "Reevaluating Amdahl's Law," *Communications of the ACM*, 1988.

[31] C. Vogel, *Computational Methods for Inverse Problems*. SIAM, 2002.

[32] Y. Saad, "Krylov subspace methods for solving large unsymmetric linear systems," *Mathematics of Computation*, 1981.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 587–595
DOI: 10.15439/2014F279
ACSIS, Vol. 2

An error estimate of Gaussian Recursive Filter in 3Dvar problem

Salvatore Cuomo
University of Naples Federico II
Department of Mathematics and Applications "R. Caccioppoli", Italy
Email: salvatore.cuomo@unina.itl

Ardelio Galletti
University of Naples "Parthenope"
Department of Science and Technology, Italy
Email: ardelio.galletti@uniparthenope.it

Raffaele Farina
Centro Euro-Mediterraneo sui Cambiamenti Climatici
CMCC, Italy
Email: raffaele.farina@cmcc.it

Livia Marcellino
University of Naples "Parthenope"
Department of Science and Technology, Italy
Email: livia.marcellino@uniparthenope.it

***Abstract*—Computational kernel of the three-dimensional variational data assimilation (3D-Var) problem is a linear system, generally solved by means of an iterative method. The most costly part of each iterative step is a matrix-vector product with a very large covariance matrix having Gaussian correlation structure. This operation may be interpreted as a Gaussian convolution, that is a very expensive numerical kernel. Recursive Filters (RFs) are a well known way to approximate the Gaussian convolution and are intensively applied in the meteorology, in the oceanography and in forecast models. In this paper, we deal with an oceanographic 3D-Var data assimilation scheme, named OceanVar, where the linear system is solved by using the Conjugate Gradient (GC) method by replacing, at each step, the Gaussian convolution with RFs. Here we give theoretical issues on the discrete convolution approximation with a first order (1st-RF) and a third order (3rd-RF) recursive filters. Numerical experiments confirm given error bounds and show the benefits, in terms of accuracy and performance, of the 3-rd RF.**

I. Introduction

In recent years, Gaussian filters have assumed a central role in image filtering and techniques for accurate measurement [26]. The implementation of the Gaussian filter in one or more dimensions has typically been done as a convolution with a Gaussian kernel, that leads to a high computational cost in its practical application. Computational efforts to reduce the Gaussian convolution complexity are discussed in [16], [24]. More advantages may be gained by employing a *spatially recursive filter*, carefully constructed to mimic the Gaussian convolution operator.

Recursive filters (RFs) are an efficient way of achieving a long impulse response, without having to perform a long convolution. Initially developed in the context of time series analysis [5], they are extensively used as computational kernels for numerical weather analysis, forecasts [17], [20], [25], digital image processing [8], [23]. Recursive filters with higher order accuracy are very able to accurately approximate a Gaussian convolution, but they require more operations.

In this paper, we investigate how the RF mimics the Gaussian convolution in the context of variational data assimilation analysis. Variational data assimilation (Var-DA) is popularly used to combine observations with a model forecast in order to produce a *best* estimate of the current state of a system and enable accurate prediction of future states. Here we deal with the three-dimensional data assimilation scheme (3D-Var), where the estimate minimizes a weighted nonlinear least-squares measure of the error between the model forecast and the available observations. The numerical problem is to minimize a cost function by means of an iterative optimization algorithm. The most costly part of each step is the multiplication of some grid-space vector by a covariance matrix that defines the error on the forecast model and observations. More precisely, in 3D-Var problem this operation may be interpreted as the convolution of a covariance function of background error with the given forcing terms.

Here we deal with numerical aspects of an oceanographic 3D-Var scheme, in the real scenario of OceanVar. Ocean data assimilation is a crucial task in operational oceanography and the computational kernel of OceanVar software is a linear system resolution by means of the Conjugate Gradient (GC) method, where the iteration matrix is relate to an errors covariance matrix, having a Gaussian correlation structure.

In [9], it is shown that a computational advantage can be gained by employing a first order RF that mimics the required Gaussian convolution. Instead, we use the 3rd-RF to compute numerically the Gaussian convolution, as how far is only used in signal processing [27], but only recently used in the field of Var-DA problems.

In this paper we highlight the main sources of error, introduced by these new numerical operators. We also investigate the real benefits, obtained by using 1-st and 3rd-RFs, through a careful error analysis. Theoretical aspects are confirmed by some numerical experiments. Finally, we report results in the case study of the OceanVar software.

The rest of the paper is organized as follows. In the next section we recall the three-dimensional variational data assimilation problem and we remark some properties on the

978-83-60810-58-3/$25.00

conditioning for this problem. Besides, we describe our case study: the OceanVar problem and its numerical solution with CG method. In section III, we introduce the n-th order recursive filter and how it can be applied to approximate the discrete Gaussian convolution. In section IV, we estimate the effective error, introduced at each iteration of the CG method, by using 1st-RF and 3rd-RF instead of the Gaussian convolution. In section V, we report some experiments to confirm our theoretical study, while the section VI concludes the paper.

II. Mathematical Background

The aim of a generic variational problem (VAR problem) is to find a best estimate x, given a previous estimate x_b and a measured value y. With these notations, the VAR problem is based on the following regularized constrained least-squared problem:

$$\min_x J(x)$$

where x is defined in a grid domain D. The objective function $J(x)$ is defined as follows:

$$J(x) = \|y - \mathcal{H}(x)\|^2 + \lambda R(x, x_b) \tag{1}$$

where measured data are compared with the solution obtained from a nonlinear model given by $\mathcal{H}(x)$.
In (1), we can recognize a quadratic data-fidelity term, the first term and the general regularization term (or penalty term), the second one. When $\lambda = 1$ and the regularization term can be write as:

$$R(x, x_b) = \|x - x_b\|^2$$

we deal with a three-dimensional variational data assimilation problem (3D-Var DA problem). The purpose is to find an optimal estimate for a vector of states x_t (called the analysis) of a generic system S, at each time $t \in T = \{0, .., n\}$ given:

- a prior estimate vector x_t^b (called the background) achieved by numerical solution of a forecasting model $\mathcal{L}_{t-1,t}(x_{t-1}) = x_t^b$, with error $\delta x_t = x_t^b - x_t$;
- a vector y_t of observations, related to the nonlinear model by δy_t that is an effective measurement error:

$$y_t = H(x_t) + \delta y_t.$$

At each time t, the errors δx_t in the background and the errors δy_t in the observations are assumed to be random with mean zero and covariance matrices $\mathbf{B}$ and $\mathbf{R}$, respectively. More precisely, the covariance $\mathbf{R} =< \delta y_t, \delta y_t^T >$ of observational error is assumed to be diagonal, (observational errors statistically independent). The covariance $\mathbf{B} =< \delta x_t, \delta x_t^T >$ of background error is never assumed to be diagonal as justified in the follow. To minimize, with respect to x_t and for each $t \in T$, the problem becomes:

$$\min_{x_t \in D} J(x_t) = \min_{x_t \in D} \{\frac{1}{2}\|y_t - H(x_t)\|^2_{\mathbf{R}} + \frac{1}{2}\|x_t - x_t^b\|^2_{\mathbf{B}}\} \tag{2}$$

In explicit form, the functional cost of (2) problem can be written as:

$$\begin{aligned} J(x_t) = \tfrac{1}{2}(y_t - H(x_t))^T\mathbf{R}^{-1}(y_t - H(x_t)) + \\ + \tfrac{1}{2}(x_t - x_t^b)^T\mathbf{B}^{-1}(x_t - x_t^b) \end{aligned} \tag{3}$$

It is often numerically convenient to approximate the effects on $H(x_t)$ of small increments of x_t, using the linearization of H. For small increments δx_t, follows [18], it is:

$$H(x_t) \simeq H(x_t^b) + \mathbf{H}\delta x_t$$

where the linear operator $\mathbf{H}$ is the matrix obtained by the first order approximation of the Jacobian of H evaluated at x_t^b.
Now let $d_t = y_t - H(x_t^b)$ be the *misfit*. Then the function J in (3) takes the following form in the increment space:

$$J(\delta x_t) = \tfrac{1}{2}(d_t - \mathbf{H}\delta x_t)^T\mathbf{R}^{-1}(d_t - \mathbf{H}\delta x_t) + \tfrac{1}{2}\delta x_t^T\mathbf{B}^{-1}\delta x_t \tag{4}$$

At this point, at each time t, the minimum of (4) is obtained by requiring $\nabla J = 0$. This gives rise to the linear system:

$$(\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})\delta x_t = \mathbf{H}^T\mathbf{R}^{-1}d_t$$

or equivalently:

$$(I + \mathbf{B}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})\delta x_t = \mathbf{B}\mathbf{H}^T\mathbf{R}^{-1}d_t \tag{5}$$

For each time $t = 0, ..., n$, iterative methods, able to converge toward a practical solution, are needed to solve the linear system (5). However this problem, so as formulated, is generally very ill conditioned. More precisely, by following [15], and assuming that

$$\mathbf{\Psi} = \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} \tag{6}$$

is a diagonal matrix, it can be proved that the conditioning of $I + \mathbf{B}\mathbf{\Psi}$ is strictly related to the conditioning of the matrix $\mathbf{B}$ (the covariance matrix). In general, the matrix $\mathbf{B}$ is a block-diagonal matrix, where each block is related to a single state of vector x_t and it is ill conditioned.
This assertion is exposed in [14] starting from the expression of $\mathbf{B}$ for one-state vectors as:

$$\mathbf{B} = \sigma_b^2\mathbf{C}$$

where σ_b^2 is the background error variance and $\mathbf{C}$ is a matrix that denotes the correlation structure of the background error. Assuming that the correlation structure of matrix $\mathbf{C}$ is homogeneous and depends only on the distance between states and not on positions, an expression of $\mathbf{C}$ as a symmetric matrix with a circulant form is given; i. e. as a Toeplitz matrix. By means of a spectral analysis of its eigenvalues, the ill-conditioning of the matrix $\mathbf{C}$ is checked. As in [7], it follows that $\mathbf{B}$ is ill-conditioned and the matrix $I + \mathbf{B}\mathbf{\Psi}$, of the linear system (5), too. A well-known technique for improving the convergence of iterative methods for solving linear systems is to *preconditioning* the system and thus reduce the condition number of the problem.
In order to precondition the system in (5), it is assumed that $\mathbf{B}$ can be written in the form $\mathbf{B} = \mathbf{V}\mathbf{V}^T$, where $\mathbf{V} = \mathbf{B}^{1/2}$ is the square root of the background error covariance matrix $\mathbf{B}$.

Because $\mathbf{B}$ is symmetric Gaussian, $\mathbf{V}$ is uniquely defined as the symmetric ($\mathbf{V^T} = \mathbf{V}$) Gaussian matrix such that $\mathbf{V^2} = \mathbf{B}$. As explained in [18], the cost function (4) becomes:

$$J(\delta x_t) = \tfrac{1}{2}(d_t - \mathbf{H}\delta x_t)^T \mathbf{R}^{-1}(d_t - \mathbf{H}\delta x_t) + \tfrac{1}{2}\delta x_t^T (\mathbf{V}\mathbf{V^T})^{-1}\delta x_t$$
$$= \tfrac{1}{2}(d_t - \mathbf{H}\delta x_t)^T \mathbf{R}^{-1}(d_t - \mathbf{H}\delta x_t) + \tfrac{1}{2}\delta x_t^T (\mathbf{V^T})^{-1}\mathbf{V}^{-1}\delta x_t$$

Now, by using a new control variable v_t, defined as $v_t = \mathbf{V}^{-1}\delta x_t$, at each time $t \in T$ and observing that $\delta x_t = \mathbf{V}v_t$ we obtain a new cost function:

$$\widetilde{J}(v_t) = \frac{1}{2}(d_t - \mathbf{H}\mathbf{V}v_t)^T \mathbf{R}^{-1}(d_t - \mathbf{H}\mathbf{V}v_t) + \frac{1}{2}v_t^T v_t. \quad (7)$$

Equation (7) is said the *dual problem* of equation (4). Finally, to minimize the cost function $\widetilde{J}(v_t)$ in (7) leads to the new linear system:

$$(I + \mathbf{V}\mathbf{\Psi}\mathbf{V})v_t = \mathbf{V}\mathbf{H}^T\mathbf{R}^{-1}d_t \quad (8)$$

Upper and lower bounds on the condition number of the matrix $I + \mathbf{V}\mathbf{\Psi}\mathbf{V}$ are shown in [14]. In particular it holds that:

$$\mu(I + \mathbf{V}\mathbf{\Psi}\mathbf{V}) << \mu(I + \mathbf{B}\mathbf{\Psi}).$$

Moreover, under some special assumptions, it can be proved that $I + \mathbf{V}\mathbf{\Psi}\mathbf{V}$ is very well-conditioned ($\mu(I + \mathbf{V}\mathbf{\Psi}\mathbf{V}) < 4$).

The OceanVar model

As described in [9], at each time $t \in T$, OceanVar software implements an oceanographic three-dimensional variational DA scheme (3D Var-DA) to produce forecasts of ocean currents for the Mediterranean Sea. The computational kernel is based on the resolution of the linear system defined in (8). To solve it, the Conjugate Gradient (CG) method is used and a basic outline is described in **Algorithm 1**.

Algorithm 1 CG Algorithm

1: $k = 0$; $\mathbf{x}_0$, the initial guess;
2: $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$;
3: $\rho_\mathbf{0} = \mathbf{r_0}$;
4: **while** $\left(\|\mathbf{r}_k\|/\|\mathbf{b}\| > \epsilon \ .and. \ k \leq n\right)$ **do**
5: $\mathbf{q}_k = \mathbf{A}\rho_k$;
6: $\alpha_k = (\mathbf{r}_k, \mathbf{r}_k)/(\rho_k, \mathbf{q}_k)$; $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\rho_\mathbf{k}$;
7: $\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\mathbf{q}_k$; $\beta_k = (\mathbf{r}_{k+1}, \mathbf{r}_{k+1})/(\mathbf{r}_k, \mathbf{r}_k)$;
8: $\rho_{k+1} = \mathbf{r}_{k+1} + \beta_k\rho_k$; $k = k + 1$;
9: **end while**

We focus our attention on step 5.: at each iterative step, a matrix-vector product $\mathbf{A}\,\rho_k$ is required, where

$$\mathbf{A} = \mathbf{I} + \mathbf{V}\mathbf{\Psi}\mathbf{V},$$

$\rho_\mathbf{k}$ is the residual at step k and $\mathbf{\Psi}$ depends on the number of observations and is characterized by a bounded norm (see [15] for details). More precisely, we look to the matrix-vector product

$$\mathbf{q}_k = (I + \mathbf{V}\mathbf{\Psi}\mathbf{V})\rho_\mathbf{k}$$

which can be schematized as shown in **Algorithm 2**.

Algorithm 2 $(I + \mathbf{V}\mathbf{\Psi}\mathbf{V})\rho_k$ Algorithm

1: $z_1 = \mathbf{V}\rho_k$;
2: $z_2 = \mathbf{\Psi}z_1$;
3: $z_3 = \mathbf{V}z_2$;
4: $\mathbf{q}_k = \rho_k + z_3$;

The steps 1. and 3. in **Algorithm 2** consist in a matrix-vector product. These products, as detailed in next section, can be considered discrete Gaussian convolutions and the matrix $\mathbf{V}$, for one-dimensional state vectors, has Gaussian structure. Even for state vectors defined on two (or more) dimensions, the matrix $\mathbf{V}$ can be represented as product of two (or more) Gaussian matrices. Since a single matrix-vector product of this form becomes prohibitively expensive if carried out explicitly, a computational advantage is gained by employing Gaussian RFs to mimic the required Gaussian convolution operators.
In the previous OceanVar scheme, it was implemented a 1st-RF algorithm, as described in [21], [20]. Here, we study the 3rd-RF introduction, based on [27], [23].
The aim of the following sections is to precisely reveal how the n-th order recursive filters are defined and, through the error analysis, to investigate on their effect in terms of error estimate and perfomences.

III. Gaussian recursive filters

In this section we describe Gaussian recursive filters as approximations of the discrete Gaussian convolution used in steps 1. and 3. of **Algorithm 2**. Let denote by

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

the normalized Gaussian function and by $\mathbf{V}$ the square matrix whose entries are given by

$$\mathbf{V}_{i,j} = g(i-j) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(i-j)^2}{2\sigma^2}\right). \quad (9)$$

Now let be $s^0 = (s_1^0, \ldots, s_m^0)^T$ a vector; the discrete Gaussian convolution of s^0 is a new vector $s = (s_1, \ldots, s_m)^T$ defined by means of the matrix-vector product

$$s = \mathbf{V} \otimes s^0 \equiv \mathbf{V}\ s^0. \quad (10)$$

The discrete Gaussian convolution can be considered as a discrete representation of the continuous Gaussian convolution. As is well known, the continuous Gaussian convolution of a function s^0 with the normalized Gaussian function g is a new function s defined as follows:

$$s(x) = [g \otimes s^0](x) = \int_{-\infty}^{+\infty} g(x-\tau)s^0(\tau)d\tau. \quad (11)$$

Discrete and continuous Gaussian convolutions are strictly related. This fact could be seen as follows. Let assume that

$$I = \{x_1 < x_2 < \ldots < x_{m+1}\}$$

is a grid of evaluation points and let set for $i = 1, \ldots, m$

$$s_i \equiv s(x_i), \qquad s_i^0 \equiv s^0(x_i) \quad \text{and} \quad \Delta x_i = x_{i+1} - x_i = 1.$$

By assuming that s^0 is 0 outside of $[x_1, x_{m+1}]$ and by discretizing the integral (11) with a rectangular rule, we obtain

$$s_i = \int_{-\infty}^{+\infty} g(x_i-\tau)s^0(\tau)d\tau = \int_{x_1}^{x_{m+1}} g(x_i-\tau)s^0(\tau)d\tau = \\ = \sum_{j=1}^{m}\int_{x_j}^{x_{j+1}} g(x_i-\tau)s^0(\tau)d\tau \approx \sum_{j=1}^{m}\Delta x_j g(x_i-x_j)s_j^0 = \\ = \sum_{j=1}^{m} g(i-j)s_j^0 = \sum_{j=1}^{m}\mathbf{V}_{i,j}s_j^0 = (\mathbf{V}s^0)_i. \quad (12)$$

An optimal way for approximating the values s_i is given by Gaussian recursive filters. The n-order RF filter computes the vector $s^K = (s_1^K, \ldots, s_m^K)^T$ as follows:

$$\begin{cases} p_i^k = \beta_i s_i^{k-1} + \sum_{j=1}^{n}\alpha_{i,j}p_{i-j}^k \quad i = 1, \ldots, m \\ s_i^k = \beta_i p_i^k + \sum_{j=1}^{n}\alpha_{i,j}s_{i+j}^k \quad i = m, \ldots, 1 \end{cases}. \quad (13)$$

The iteration counter k goes from 1 to K, where K is the total number of filter iterations. Observe that values $p_1^k, \ldots p_n^k$ are computed taking in the sums terms $\alpha_{i,j}p_{i-j}^k$ provided that $i-j \geq 1$. Analogously values $s_m^k, \ldots s_{m-n+1}^k$ are computed taking in the sums terms $\alpha_{i,j}s_{i+j}^k$ provided that $i+j \leq m$. The values $\alpha_{i,j}$ and β_i, at each grid point x_i, are often called *smoothing coefficients* and they obey to the constraint

$$\beta_i = 1 - \sum_{j=1}^{n}\alpha_{i,j}.$$

In this paper we deal with first-order and third-order RFs. The first-order RF expression ($n = 1$) becomes:

$$\begin{cases} p_1^k = \beta_1 s_1^{k-1}, \\ p_i^k = \beta_i s_i^{k-1} + \alpha_i p_{i-1}^k \quad i = 2, \ldots, m \\ s_m^k = \beta_m p_m^k, \\ s_i^k = \beta_i p_i^k + \alpha_i s_{i+1}^k \qquad i = m-1, \ldots, 1. \end{cases} \quad (14)$$

If R_i is the correlation radius at x_i, by setting

$$\sigma_i = \frac{R_i}{\Delta x_i} \quad \text{and} \quad E_i = \frac{K\Delta x_i^2}{R_i^2} = \frac{K}{\sigma_i^2},$$

coefficients α_i e β_i are given by [21]:

$$\alpha_i = 1 + E_i - \sqrt{E_i(E_i+2)}, \qquad \beta_i = \sqrt{E_i(E_i+2)} - E_i. \quad (15)$$

The third-order RF expression ($n = 3$) becomes:

$$\begin{cases} p_i^k = \beta_i s_i^{k-1} + \sum_{j=1}^{3}\alpha_{i,j}p_{i-j}^k \quad i = 1, ..., m \\ s_i^k = \beta_i p_i^k + \sum_{j=1}^{3}\alpha_{i,j}s_{i+j}^k \quad i = m, \ldots, 1. \end{cases} \quad (16)$$

Third-order RF coefficients $\alpha_{i,1}, \alpha_{i,2}, \alpha_{i,3}$ and β_i, for one only filter iteration ($K = 1$), are computed in [11]. If

$$a_i = 3.738128 + 5.788982\sigma_i + 3.382473\sigma_i^2 + \sigma_i^3.$$

the coefficients expressions are:

$$\alpha_{i,1} = (5.788982\sigma_i + 6.764946\sigma_i^2 + 3\sigma_i^3)/a_i$$
$$\alpha_{i,2} = -(3.382473\sigma_i^2 + 3\sigma_i^3)/a_i$$
$$\alpha_{i,3} = \sigma_i^3/a_i$$
$$\beta_i = 1 - (\alpha_{i,1} + \alpha_{i,2} + \alpha_{i,3}) = 3.738128/a_i.$$

In [23] the use of a value $q = q(\sigma_i)$ instead of σ_i is proposed. The q value is:

$$q(\sigma_i) = \begin{cases} 0.98711\sigma_i - 0.96330 & \text{if} \quad \sigma_i > 2.5 \\ 3.97156 - 4.14554\sqrt{1 - 0.26891\sigma_i} & \text{oth.} \end{cases} \quad (17)$$

In order to understand how Gaussian RFs approximate the discrete Gaussian convolution it is useful to represent them in terms of matrix formulation. As explained in [5], the n-order recursive filter computes s^K from s^0 as the solution of the linear system

$$(LU)^K s^K = s^0, \quad (18)$$

where matrices L and U are respectively lower and upper band triangular with nonzero entries

$$U_{i,i} = L_{i,i} = \frac{1}{\beta_i}, \qquad L_{i,i-j} = U_{i,i+j} = -\frac{\alpha_{i,j}}{\beta_i}. \quad (19)$$

By formally inverting the linear system (18) it results

$$s^K = \mathbf{F}_{\mathbf{n}}^{(\mathbf{K})}\, s^0, \quad (20)$$

where $\mathbf{F}_{\mathbf{n}}^{(\mathbf{K})} \equiv (LU)^{-K}$. A direct expression of $\mathbf{F}_{\mathbf{n}}^{(\mathbf{K})}$ and its norm could be obtained, for instance, for the first order recursive filter in the homogenus case ($\sigma_i = \sigma$). However, in the following, it will be shown that $\mathbf{F}_{\mathbf{n}}^{(\mathbf{K})}$ has always bounded norm, i.e.

$$\|\mathbf{F}_{\mathbf{n}}^{(\mathbf{K})}\|_\infty \leq 1. \quad (21)$$

Observe that $\mathbf{F}_{\mathbf{n}}^{(\mathbf{K})}$ is the matrix operator that substitutes the Gaussian operator $\mathbf{V}$ in (10), then a measure of how well s^K approximates s can be derived in terms of the operator distance

$$\|\mathbf{V} - \mathbf{F}_{\mathbf{n}}^{(\mathbf{K})}\|_\infty.$$

Ideally one would expect that $\|\mathbf{V} - \mathbf{F}_{\mathbf{n}}^{(\mathbf{K})}\|$ goes to 0 (and $s^K \to s$) as K approaches to ∞, yet this does not happen due to the presence of edge effects. In the next sections we show the numerical behaviour of the distance $\|\mathbf{V} - \mathbf{F}_{\mathbf{n}}^{(\mathbf{K})}\|$ for some case study and we will show its effects in the CG algorithm.

IV. RF ERROR ANALYSIS

Here we are interested to analyze the error introduced on the matrix-vector operation at step 5. of **Algorithm 1**, when the Gaussian RF is used instead of the discrete Gaussian convolution. As previously explained, in terms of matrices, this is equivalent to change the matrix operator, then **Algorithm 2** can be rewritten as shown in **Algorithm 3**.

Now we are able to give the main result of this paper: indeed the following theorem furnishes an upper bound for the error

Algorithm 3 $(I + \mathbf{F_n^{(K)}}\mathbf{\Psi}\mathbf{F_n^{(K)}})\widetilde{\rho}_k$ Algorithm

1: $\widetilde{z}_1 = \mathbf{F_n^{(K)}}\widetilde{\rho}_k$;
2: $\widetilde{z}_2 = \mathbf{\Psi}\widetilde{z}_1$;
3: $\widetilde{z}_3 = \mathbf{F_n^{(K)}}\widetilde{z}_2$;
4: $\widetilde{\mathbf{q}}_k = \widetilde{\rho}_k + \widetilde{z}_3$;

$\mathbf{q}_k - \widetilde{\mathbf{q}}_k$, made at each single iteration k of the CG (**Algorithm 1**). This bound involves the operator norms

$$\|\mathbf{F_n^{(K)}}\|_\infty, \quad \|\mathbf{\Psi}\|_\infty, \quad \|\mathbf{V}\|_\infty,$$

the distance $\|\mathbf{V} - \mathbf{F_n^{(K)}}\|_\infty$ and the error $\rho_k - \widetilde{\rho}_k$ accumulated on ρ_k at previous iterations.

Theorem 4.1: Let be ρ_k, $\widetilde{\rho}_k$, $\mathbf{q}_k$, $\widetilde{\mathbf{q}}_k$ as in **Algorithm 2** and **Algorithm 3**. Let be $\|\cdot\| = \|\cdot\|_\infty$ and let denote by

$$e_k = \rho_k - \widetilde{\rho}_k$$

the difference between values ρ_k and $\widetilde{\rho}_k$. Then it holds

$$\|\mathbf{q}_k - \widetilde{\mathbf{q}}_k\| \leq (1 + \|\mathbf{V}\|\cdot\|\mathbf{\Psi}\|\cdot\|\mathbf{V}\|)\cdot\|e_k\| + \\ + \|\mathbf{F_n^{(K)}} - \mathbf{V}\|\cdot\|\mathbf{\Psi}\|\cdot\big(\|\mathbf{V}\| + \|\mathbf{F_n^{(K)}}\|\big)\cdot\|\widetilde{\rho}_k\|. \quad (22)$$

Proof: A direct proof follows by using the values z_i and $\widetilde{z}_i$ introduced in Algorithm 2 and in Algorithm 3. It holds:

$$\|z_1 - \widetilde{z}_1\| = \|\mathbf{V}\rho_k - \mathbf{F_n^{(K)}}\widetilde{\rho}_k\| = \|\mathbf{V}\rho_k - \mathbf{V}\widetilde{\rho}_k + \mathbf{V}\widetilde{\rho}_k - \mathbf{F_n^{(K)}}\widetilde{\rho}_k\| \leq \\ \leq \|\mathbf{V}\rho_k - \mathbf{V}\widetilde{\rho}_k\| + \|\mathbf{V}\widetilde{\rho}_k - \mathbf{F_n^{(K)}}\widetilde{\rho}_k\| \leq \\ \leq \|\mathbf{V}\|\cdot\|e_k\| + \|\mathbf{V} - \mathbf{F_n^{(K)}}\|\cdot\|\widetilde{\rho}_k\|.$$

Then, for the difference $z_2 - \widetilde{z}_2$, we get the bound

$$\|z_2 - \widetilde{z}_2\| = \|\mathbf{\Psi}z_1 - \mathbf{\Psi}\widetilde{z}_1\| \leq \|\mathbf{\Psi}\|\cdot\|z_1 - \widetilde{z}_1\| \leq \\ \leq \|\mathbf{\Psi}\|\cdot\|\mathbf{V}\|\cdot\|e_k\| + \|\mathbf{\Psi}\|\cdot\|\mathbf{V} - \mathbf{F_n^{(K)}}\|\cdot\|\widetilde{\rho}_k\|.$$

Hence, for the difference $z_3 - \widetilde{z}_3$, we obtain

$$\|z_3 - \widetilde{z}_3\| = \|\mathbf{V}z_2 - \mathbf{F_n^{(K)}}\widetilde{z}_2\| = \|\mathbf{V}z_2 - \mathbf{V}\widetilde{z}_2 + \mathbf{V}\widetilde{z}_2 - \mathbf{F_n^{(K)}}\widetilde{z}_2\| \leq \\ \leq \|\mathbf{V}\|\cdot\|z_2 - \widetilde{z}_2\| + \|\mathbf{V} - \mathbf{F_n^{(K)}}\|\cdot\|\widetilde{z}_2\| \leq \\ \leq \|\mathbf{V}\|\cdot\|z_2 - \widetilde{z}_2\| + \|\mathbf{V} - \mathbf{F_n^{(K)}}\|\cdot\|\mathbf{\Psi}\|\cdot\|\mathbf{F_n^{(K)}}\|\cdot\|\widetilde{\rho}_k\| \leq \\ \leq \|\mathbf{V}\|\cdot\|\mathbf{\Psi}\|\cdot\|\mathbf{V}\|\cdot\|e_k\| + \|\mathbf{V}\|\cdot\|\mathbf{\Psi}\|\cdot\|\mathbf{V} - \mathbf{F_n^{(K)}}\|\cdot\|\widetilde{\rho}_k\| \\ + \|\mathbf{V} - \mathbf{F_n^{(K)}}\|\cdot\|\mathbf{\Psi}\|\cdot\|\mathbf{F_n^{(K)}}\|\cdot\|\widetilde{\rho}_k\| = \\ \|\mathbf{V}\|\cdot\|\mathbf{\Psi}\|\cdot\|\mathbf{V}\|\cdot\|e_k\| + \|\mathbf{V} - \mathbf{F_n^{(K)}}\|\cdot\|\mathbf{\Psi}\|\big(\|\mathbf{V}\| + \|\mathbf{F_n^{(K)}}\|\big)\cdot\|\widetilde{\rho}_k\|$$

In the second-last inequality we used the fact that

$$\|\widetilde{z}_2\| = \|\mathbf{\Psi}\widetilde{z}_1\| = \|\mathbf{\Psi}\mathbf{F_n^{(K)}}\widetilde{\rho}_k\| \leq \|\mathbf{\Psi}\|\cdot\|\mathbf{F_n^{(K)}}\|\cdot\|\widetilde{\rho}_k\|.$$

Finally, observing that

$$\|\mathbf{q}_k - \widetilde{\mathbf{q}}_k\| = \|\rho_k + z_3 - (\widetilde{\rho}_k + \widetilde{z}_3)\| \leq \\ \leq \|\rho_k - \widetilde{\rho}_k\| + \|z_3 - \widetilde{z}_3\| = \|e_k\| + \|z_3 - \widetilde{z}_3\|,$$

and taking the upper bound of $\|z_3 - \widetilde{z}_3\|$, the thesis is proved. ◇

Previous theorem shows that, at each iteration of the CG algorithm, the error bound on the computed value $\mathbf{q_k}$ at step 5., is characterized by two main terms: the first term can be considered as the contribution of the standard forward error analysis and it is not significant, if $\|e_k\|$ is small; the second term highlights the effect of the introduction of the RF. More in detail, at each iteration step, the computed value $\mathbf{q_k}$ is biased by a quantity proportional to three factors:

- the distance between the original operator (the Gaussian operator $\mathbf{V}$) and its approximation (the operator $\mathbf{F_n^{(K)}}$);
- the norm of $\mathbf{\Psi}$;
- the sum of the operator norms $\|\mathbf{F_n^{(K)}}\|$ and $\|\mathbf{V}\|$.

Table 1: Operator norms

σ	$\|\|\mathbf{F_1^{(1)}}\|\|_\infty$	$\|\|\mathbf{F_3^{(1)}}\|\|_\infty$
5	0.9920	0.9897
20	0.9012	0.8537
50	0.9489	0.8950

As shown in (21) the norm $\mathbf{\Psi}$ is bounded. Besides, the norm of $\mathbf{V}$ is always less or equal to one (because it comes from the discretization of the of the continuous Gaussian convolution). The norm of $\mathbf{F_n^{(K)}}$ is bounded by one too. This fact can be seen by observing the Table 1, where we consider several tests by varying data distributions in the homogeneous case ($\sigma_i = \sigma$), for 1st-RF and 3rd-RF. Starting from these considerations, the error estimate of *Theorem 4.1* can be specialized as:

$$\|\mathbf{q}_k - \widetilde{\mathbf{q}}_k\| \leq (1 + \|\mathbf{\Psi}\|)\,\|e_k\| + 2\|\mathbf{F_n^{(K)}} - \mathbf{V}\|\cdot\|\mathbf{\Psi}\|\|\widetilde{\rho}_k\|. \quad (23)$$

V. Experimental Results

In this section we report some experiments to confirm the discussed theoretical results. In the first part, we deal with the approximations of the discrete operator $\mathbf{V}$ with the first order and of the third order $\mathbf{F_1^{(K)}}$ and $\mathbf{F_3^{(1)}}$ respectively. In the last subsection, we analyze the improving in the performance and in the accuracy terms of the third order RF applied to the case study.

A. 1st-RF and 3rd-RF operators

In the following experiments, we construct the operators $\mathbf{V}$, $\mathbf{F_1^{(1)}}$, $\mathbf{F_1^{(50)}}$ and $\mathbf{F_3^{(1)}}$ in the case of $m = 601$ samples of a random vector $\mathbf{s^0}$. We assume that $\mathbf{s^0}$ comes from a uniform grid with homogeneous condition $\sigma_i = \sigma = 15$. In Figure 1, it is highlighted that the involved discrete operators have different structures. In particular, a first qualitative remark is that the operator $\mathbf{F_1^{(1)}}$ is a poor approximation of $\mathbf{V}$. Conversely, the operator $\mathbf{F_1^{(50)}}$ (Figure 2 on the top) is very close to $\mathbf{V}$ but, as for $\mathbf{F_1^{(1)}}$, there are significant differences with $\mathbf{V}$ in the bottom left and in the top right corners. These dissimilarities in the edges, by a numerical point of view, give some kind of artifacts in the computed convolutions, that determine a vector $\mathbf{s}$ with components, in the initial and final positions, that decay to zero.

Figure 2 bottom shows that the operator $\mathbf{F_3^{(1)}}$ is closer then $\mathbf{F_1^{(1)}}$ and $\mathbf{F_1^{(50)}}$ to the discrete convolution $\mathbf{V}$. In particular, this recursive filter is able to reproduce $\mathbf{V}$ more accurately in the bottom left corner, but unfortunately it does not give good

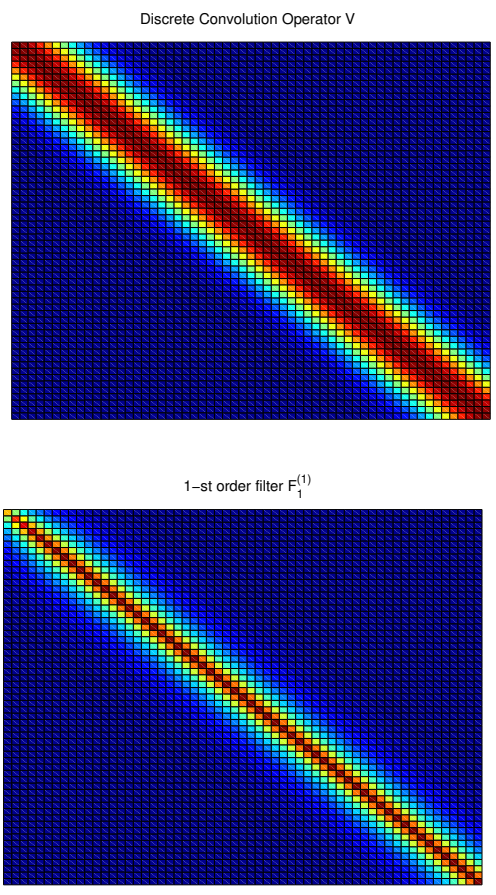


Fig. 1. **Top**. Discrete Gaussian convolution operator $\mathbf{V}$. **Bottom**. 1-st order recursive filter operator F_1

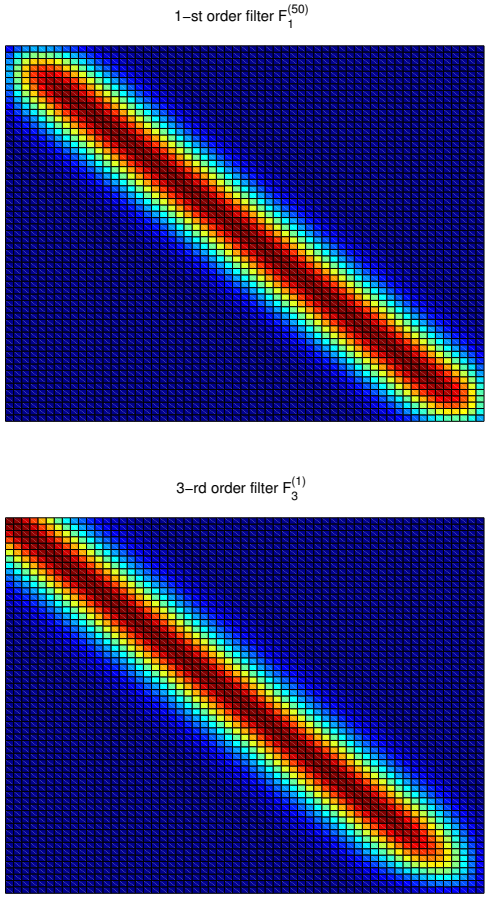


Fig. 2. **Top**. 1-st order recursive filter operator $F_1^{(50)}$ with 50 iterations. **Bottom**. 3-rd order recursive filter operator $F_3^{(1)}$

results on top right corner. In Table 2, for random distributions with homogeneous condition ($\sigma_i = \sigma$), we underline the edge effects by measuring the norms between the discrete convolution $\mathbf{V}$ and the RF filters. Although the $||\mathbf{F_n^{(K)}} - \mathbf{V}||_\infty$ ideally goes to zero as k goes to $+\infty$, this does not happen in practice as observed below.

Table 2: Distance metrics

σ	$\|\|\mathbf{F_1^{(1)}} - \mathbf{V}\|\|_\infty$	$\|\|\mathbf{F_1^{(50)}} - \mathbf{V}\|\|_\infty$	$\|\|\mathbf{F_3^{(1)}} - \mathbf{V}\|\|_\infty$
5	0.2977	0.3800	0.5346
10	0.3895	0.4397	0.5890
25	0.4533	0.4758	0.6221
50	0.4686	0.4809	0.6125

In order to bring out these considerations, we show the application of $\mathbf{V}$, $\mathbf{F_1^{(K)}}$ and $\mathbf{F_3^{(1)}}$ to a periodic signal $\mathbf{s^0}$. We choose $m = 252$ samples of the cos function in $[-2\pi, 2\pi]$ and we perform simulations by using the 1-st RF with $1, 5$ and 50 iterations and 3-rd RF with one iteration. In Figure 3 it is shown the computed Gaussian convolution and the poor approximation of $\mathbf{V}s^0$ on the right side of the test interval, due to the edge effects. A nice result is that our $\mathbf{F_3^{(1)}}$ convolution operator gives better results on the left side of the domain.

Finally, we give some considerations about the accuracy of the studied Gaussian RF schemes, when they are applied to the Dirac rectangular impulse

$$s^0 = (0, \ldots, 0, 1, 0, \ldots).$$

We choose a one-dimensional grid of $m = 301$ points, a constant correlation radius $R = 120, km$, a constant grid space $\Delta x = 6\, km$ and $\sigma = R/\Delta x = 20$. In the numerical experiments to avoid the edge effects, we only consider $\bar{m} = 221$ central values of s^K, i.e.

$$\bar{s}^K = (s^K_{2\sigma}, s^K_{2\sigma+1}, \ldots, s^K_{m-2\sigma-1}, s^K_{m-2\sigma}).$$

Similarly, in Table 3 we measure the operator distances we use $||\bar{\mathbf{F}}_\mathbf{1}^{(1)} - \bar{\mathbf{V}}||_\infty$ and $||\bar{\mathbf{F}}_\mathbf{3}^{(1)} - \bar{\mathbf{V}}||_\infty$, where $\bar{V}$, $\bar{F}_1^{(1)}$ and $\bar{F}_3^{(1)}$ indicate the submatrices obtained, neglecting first and last $2\sigma - 1$ rows and columns.

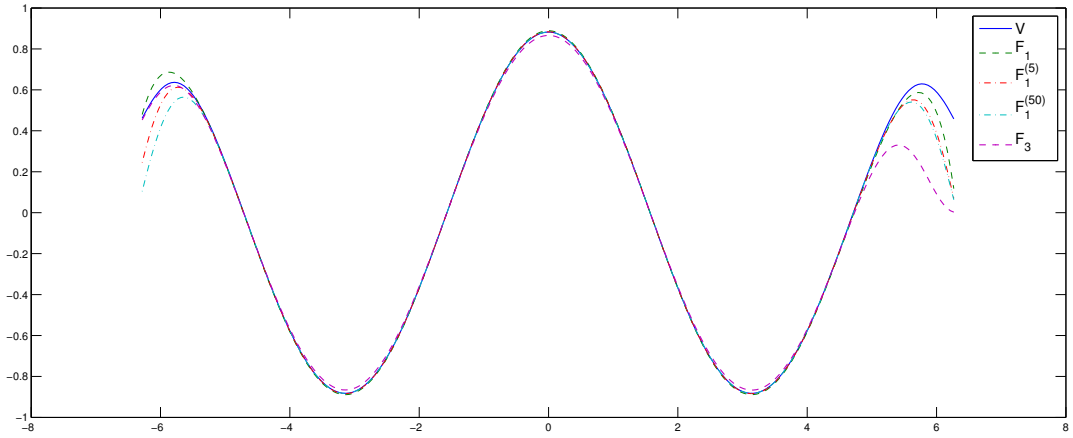


Fig. 3. Discrete convolution $\mathbf{V}$ and Gaussian recursive filtering $\mathbf{F_1^{(K)}}$ with $1, 5, 50$ iterations and $\mathbf{F_3^{(1)}}$ applied to $n = 252$ samples of the periodic function $s^0 = \cos(x)$ in $[-2\pi, 2\pi]$.

Table 3: Convergence history

K	$\|\|\bar{\mathbf{F}}_\mathbf{1}^{(\mathbf{K})} - \bar{\mathbf{V}}\|\|_\infty$	$\|\|\bar{\mathbf{F}}_\mathbf{3}^{(\mathbf{K})} - \bar{\mathbf{V}}\|\|_\infty$
1	0.211	**0.0424**
2	0.13	–
5	0.078	–
50	0.048	–
100	0.0429	–
500	0.0414	–

These case studies show that , neglecting the edge effects, the 3-rd RF filter is more accurate the the 1st-RF order with few iterations. This fact is evident by observing the results in Figure 4 and the operator norms in Table 3. Finally, we remark that the 1-st order RF has to use 100 iteration in order to obtain the same accuracy of the 3-rd order RF. This is a very interesting numerical feature of the third order filter.

B. A case study: Ocean Var

The theoretical considerations of the previous sections are useful to understand the accuracy improvement in the real experiments on Ocean Var. The preconditioned CG is a numerical kernel intensively used in the model minimizations. Implementing a more accurate convolution operators gives benefits on the convergence of GC and on the overall data assimilation scheme [11] . Here we report experimental results of the 3rd-RF in a Global Ocean implementation of OceanVar that follows [22], [12]. These results are extensively discussed in the report [11]. In real scenarios [4], [10], scientific libraries and high performance computing environments are needed. The case study simulations were carried-out on an IBM cluster using 64 processors. The model resolution was about $1/4$ degree and the horizontal grid was tripolar, as described in [19]. This configuration of the model was used at CMCC for global ocean physical reanalyses applications (see [13]). The model has 50 vertical depth levels. The three-dimensional

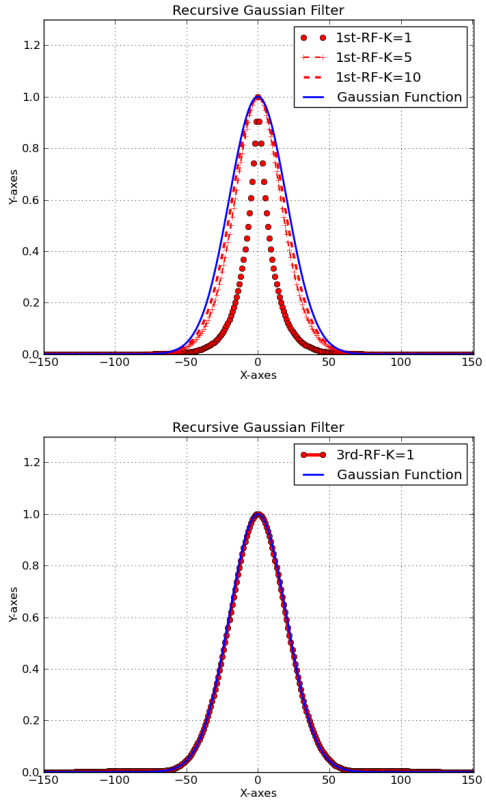


Fig. 4. **Top**. The discrete Gaussian convolution $\mathbf{V}s^0$ (blue) and $\mathbf{F_1^{(K)}}s^0$ for $K = 1, 5, 10$ (red). **Bottom** The discrete Gaussian convolution $\mathbf{V}s^0$ (blue) and $\mathbf{F_3^{(1)}}s^0$ (red).

model grid consists of 736141000 grid-points. The comparison between the 1st-RF and 3rd-RF was carried out for a realistic case study, where all in-situ observations of temperature and

salinity from Expendable bathythermographs (XBTs), Conductivity, Temperature, Depth (CTDs) Sensors, Argo floats and Tropical mooring arrays were assimilated. The observational profiles are collected, quality-checked and distributed by [3]. The global application of the recursive filter accounts for spatially varying and season-dependent correlation length-scales (CLSs). Correlation length-scale were calculated by applying the approximation given in [2] to a dataset of monthly anomalies with respect to the monthly climatology, with inter-annual trends removed.

The obtained performances of a 3Dvar application that uses the 1st-RF with 1, 5 and 10 iterations and the 3rd-RF are shown in Figure 5 with a zoom in the same area of Western Pacific Area as in Figure 5, for the temperature at 100 m of depth. The Figure also displays the differences between the 3rd-RF and the 1st-RF with either 1 or 10 iterations. The patterns of the increments are closely similar, although increments for the case of 1st-RF (K=1) are generally sharper in the case of both short (e.g. off Japan) or long (e.g. off Indonesian region) CLSs. The panels of the differences reveal also that the differences between 3rd-RF and the 1st-RF (K=10) are very small, suggesting once again that the same accuracy of the 3rd-RF can be achieved only with a large number of iterations for the first order recursive filter. Finally, in [12] was also observed that the 3rd-RF compared to the 1st-RF (K=5) and the 1st-RF (K=10) reduces the wall clock time of the software respectively of about 27% and 48%.

VI. Conclusions

Recursive Filters (RFs) are a well known way to approximate the Gaussian convolution and are intensively applied in the meteorology, in the oceanography and in forecast models. In this paper, we deal with the oceanographic 3D-Var scheme OceanVar. The computational kernel of the OceanVar software is a linear system solved by means of the Conjugate Gradient (GC) method. The iteration matrix is related to an error covariance matrix, with a Gaussian correlation structure. In other words, at each iteration, a Gaussian convolution is required. Generally, this convolution is approximated by a first order RF. In this work, we introduced a 3rd-RF filter and we investigated about the main sources of error due to the use of 1st-RF and 3rd-RF operators. Moreover, we studied how these errors influence the CG algorithm and we showed that the third order operator is more accurate than the first order one. Finally, theoretical issues were confirmed by some numerical experiments and by the reported results in the case study of the OceanVar software.

References

[1] M. Abramowitz, I. Stegun - *Handbook of Mathematical Functions*. Dover, New York, 1965.

[2] M. Belo Pereira, L. Berre - *The use of an ensemble approach to study the background-error covariances in a global NWP model*. MOn. Wea. Rev. 134, pp. 2466-2489, 2006.

[3] C. Cabanes, A. Grouazel, K. von Schuckmann, M. Hamon, V. Turpin, C. Coatanoan, F. Paris, S. Guinehut, C. Bppne, N. Ferry, C. de Boyer Montgut, T. Carval, G. Reverding, S. Puoliquen, P.Y. L. Traon - *The CORA dataset: validation and diagnostics of in-situ ocean temperature and salinity measurements*. Ocean Sci 9, pp. 1-18, 2013.

[4] S. Cuomo, A. Galletti, G. Giunta and A. Starace-*Surface reconstruction from scattered point via RBF interpolation on GPU* , Federated Conference on Computer Science and Information Systems (FedCSIS), 2013, pp. 433-440.

[5] G. Dahlquist and A. Bjorck - *Numerical Methods*. Prentice Hall, 573 pp. 1974.

[6] J. Derber, A. Rosati - *A global oceanic data assimilation system*. Journal of Phys. Oceanogr. 19, pp. 1333-1347, 1989.

[7] L. D' Amore, R. Arcucci, L. Marcellino, A. Murli- *HPC computation issues of the incremental 3D variational data assimilation scheme in OceanVarsoftware*. Journal of Numerical Analysis, Industrial and Applied Mathematics, 7(3-4), pp 91-105, 2013.

[8] R. Deriche - *Separable recursive filtering for efficient multi-scale edge detection*. Proc. Int. Workshop Machine Vision Machine Intelligence, Tokyo, Japan, pp 18-23, 1987

[9] S. Dobricic, N. Pinardi - *An oceanographic three-dimensional variational data assimilation scheme*. Ocean Modeling 22, pp 89-105, 2008.

[10] R. Farina, S. Cuomo, P. De Michele, F. Piccialli-*A Smart GPU Implementation of an Elliptic Kernel for an Ocean Global Circulation Model*, APPLIED MATHEMATICAL SCIENCES, 7 (61-64), 2013 pp.3007-3021.

[11] R. Farina, S. Dobricic, S. Cuomo-*Some numerical enhancements in a data assimilation scheme*, AIP Conference Proceedings 1558, 2013, doi: 10.1063/1.4826017.

[12] R. Farina, S. Dobricic, A. Storto, S. Masina, S. Cuomo-*A Revised Scheme to Compute Horizontal Covariances in an Oceanographic 3D-VAR Assimilation System*, CoRR, abs/1404.5756, 2014, http://arxiv.org/abs/1404.5756

[13] N. Ferry, B. Barnier, G. Garric, K. Haines, S. Masina, L. Parent, A. Storto, M. Valdivieso, S. Guinehut, S. Mulet - *NEMO: the modeling engine of global ocean reanalysis*. Mercator Ocean Quaterly Newsletter 46, pp 60-66, 2012.

[14] S. Haben, A. Lawless, N. Nichols - *Conditioning of the 3DVar data assimilation problem*. University of Reading, Dept. of Mathematics, Math Report Series 3, 2009;

[15] S. Haben, A. Lawless, N. Nicholos - *Conditioning and preconditioning of the variational data assimilation problem*. Computers and Fluids 46, pp 252-256, 2011.

[16] L. Haglund - *Adaptive multidimensional filtering*. Linkping University, Sweden, 1992.

[17] A.C. Lorenc - *Iterative analysis using covariance functions and filters*. Quartely Journal of the Royal Meteorological Society 1-118, pp 569-591, 1992.

[18] A.C. Lorenc - *Development of an operational variational assimilation scheme*. Journal of the Meteorological Society of Japan 75, pp 339-346, 1997.

[19] G. Madec, M. Imbard - *A global ocean mesh to overcome the north pole singularity*. Clim. Dynamic 12, pp 381-388, 1996.

[20] R.J. Purser, W.-S. Wu, D.F. Parish, N.M. Roberts - *Numerical aspects of the application of recursive filters to variational statistical analysis. Part II: spatially inhomogeneous and anisotropic covariances*. Monthly Weather Review 131, pp 1524-1535, 2003.

[21] C. Hayden, R. Purser - *Recursive filter objective analysis of meteorological field: applications to NESDIS operational processing*. Journal of Applied Meteorology 34, pp 3-15, 1995.

[22] A. Storto, S. Dobricic, S. Masina, P. D. Pietro - *Assimilating along-track altimetric observations through local hydrostatic adjustments in a global ocean reanalysis system*. Mon. Wea. Rev. 139, pp 738-754, 2011.

[23] L.V. Vliet, I. Young, P. Verbeek - *Recursive Gaussian derivative filters*. International Conference Recognition, pp 509-514, 1998.

[24] L.J. van Vliet, P.W. Verbeek - *Estimators for orientation and anisotropy in digitized images*. Proc. ASCI'95, Heijen (Netherlands), pp 442-450, 1995.

[25] A. T. Weaver, P. Courtier - *Correlation modelling on the sphere using a generalized diffusion equation*. Quarterly Journal of the Royal Meteorological Society 127, pp 1815-1846, 2001.

[26] A. Witkin - *Scale-space filtering*. Proc. Internat. Joint Conf. on Artificial Intelligence, Karlsruhe, germany, pp 1019-1021, 1983.

[27] I.T. Young, L.J. van Vliet - *Recursive implementation of the Gaussian filter*. Signal Processing 44, pp 139-151, 1995.

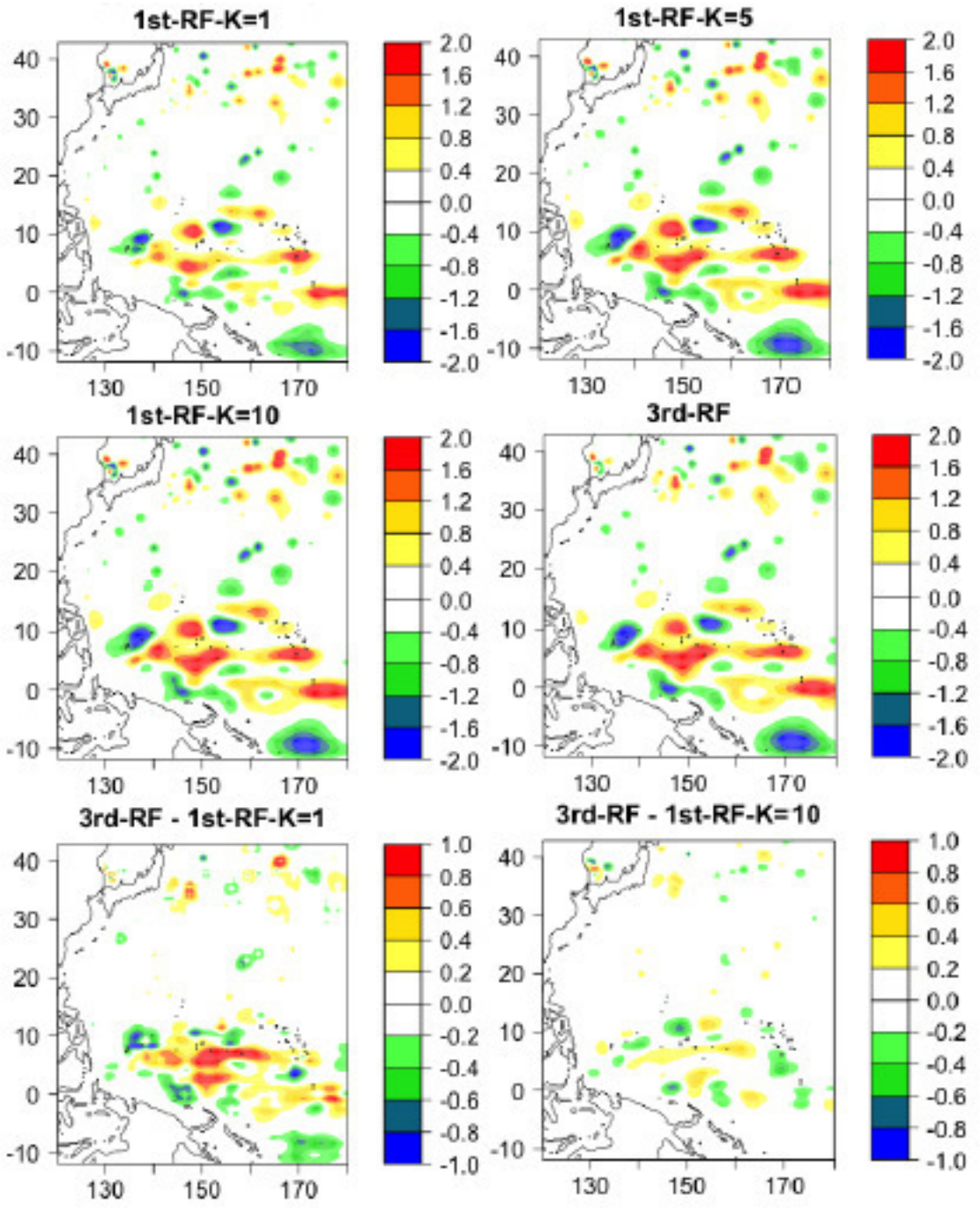


Fig. 5. Analysis increments of temperature at 100 m of depth for the Western Pacific for different configurations of the recursive filter (first two rows of panels). Differences of 100 m temperature analysis increments between 3rd-RF and 1st-RF (K=1) and between 3rd-RF and 1st-RF (K=10) (bottom panels).

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 597–602
DOI: 10.15439/2014F216
ACSIS, Vol. 2

Inexact Newton matrix-free methods for solving complex biotechnological systems

Paweł Drąg
Institute of Computer Engineering, Control and Robotics,
Wrocław University of Technology
Janiszewskiego 11/17, 50-372 Wrocław, Poland
Email: pawel.drag@pwr.edu.pl

Marlena Kwiatkowska
Institute of Environmental Protection Engineering,
Faculty of Environmental Engineering,
Wrocław University of Technology
Pl. Grunwaldzki 9, 50-377 Wrocław, Poland
Email: marlena.kwiatkowska@pwr.edu.pl

***Abstract*—In the article a new approach for solving complex and highly nonlinear differential-algebraic equations (DAEs) was presented. An important kind of applications of DAE systems is modeling of biotechnological processes, which can have a very different course. An efficient solving of equations describing biotechnological industrial inlets results in better optimization of the processes and has a positive impact on the environment. Some of the mentioned processes were characterized by a highly nonlinear dynamics. To obtain the trajectories of the state numerically, the backward differentiation formula was used in the presented method. As a result, a large-scale system of nonlinear algebraic equations was obtained. To solve a such system, the inexact Newton matrix-free approach was proposed. The new algorithm was tested on a mathematical model of a fed-batch fermentor for penicillin production. The numerical simulations were executed in MATLAB using Wroclaw Center for Networking and Supercomputing.**

Keywords*—inexact Newton methods, matrix-free methods, DAE systems, system of nonlinear equations.

I. Introduction

NOWADAYS, biotechnological processes are widely used in many real-life industrial plants. They can be often met in food industry, production of medicines and in many other sectors of industry, especially, when biodegradable components have to be used for the protection of the environment [16], [21].

Very often, the mathematical models of the bioprocesses have highly nonlinear dynamics. Technological and resources constraints on both the state and the control variables are also frequently present. Hence, a commonly used way to describe complex processes are both nonlinear ordinary differential equations and differential-algebraic equations [3], [4].

In recent years, many efforts have been devoted to the model-based optimization of processes in biotechnology and bioengineering. An example of a problem which has received major attention is the dynamic optimization of fed-batch bioreactors [17]. Dynamic optimization allows the computation of the optimal operating policies for these units to ensure the maximization of a predefined performance index. The performance index reflects a productivity or an economical index derived from both the operation profile and the final concentrations [22].

The development of information technology, robust numerical methods and computing capacity, enables to obtain optimal operating policies of the complex biotechnological processes. An efficient solving of the differential-algebraic systems enables the use of optimization strategies, what can improve a process flow significantly [5], [10].

In this work, the general problem of solving dynamical models of bioprocesses described by nonlinear differential-algebraic equations was considered. A solution strategy based on the matrix-free inexact Newton method was presented.

The article consists of 5 sections. The problem of solving complex and highly nonlinear differential-algebraic equations (DAEs) will be introduced in the next section. In the 3*rd* section the matrix-free Newton-Krylov method will be presented. The inexact Newton method will be discussed in the 4*th* section. The inexact Newton matrix-free approach will be tested on the fed-batch fermentor for penicillin production. The numerical results will be presented in the 5*th* section.

II. Statement of the problem

In general, real-life biotechnological systems with dynamics and conservation laws can be described in a *fully-implicit* form

$$\mathcal{B}\Big(\dot{y}(t), y(t), z(t), u(t), p, t\Big) = 0. \tag{1}$$

Here $y(t) \in \mathcal{R}^{n_y}$ represents the differential state trajectory, whereas $z(t) \in \mathcal{R}^{n_z}$ denotes the algebraic state trajectory, $u(t) \in \mathcal{R}^{n_u}$ a vector representing control function and $p \in \mathcal{R}^{n_p}$ indicates a vector of parameters constant in the time. Then, the nonlinear vector-valued function is considered

$$\mathcal{B} : \mathcal{R}^{n_y \times n_z \times n_u \times 1} \rightarrow \mathcal{R}^{n_\mathcal{B}}. \tag{2}$$

On the other hand, when only dynamical features of the systems are pondered, the ordinary differential equations are enough

$$\dot{y}(t) = \mathcal{G}\big(y(t), u(t), p, t\big). \tag{3}$$

Hence, some interesting relations between variables and their physical interpretations can be lost [6].

The first general technique for the numerical solution of the *fully-implicit* DAEs was the backward differential formula. The idea of this technique was that the derivative $\dot{y}(t)$ could be approximated by a linear combination of the solution $y(t)$ at the current mesh point and at several previous mesh points [19].

Previously, the backward differential formula was defined for the differential equations systems coupled to the algebraic

978-83-60810-58-3/$25.00

equations. The application of this method was soon extended to any fully-implicit system of the differential-algebraic equations.

The first order backward differential formula has been considered as the simplest method for solving differential-algebraic systems [14]. It consists of replacing the derivative in eq. (1) by the backward difference quotient

$$F\Big(\frac{y_{n+1}-y_n}{h}, y_{n+1}, z_{n+1}, t_{n+1}\Big) = 0. \tag{4}$$

where $h = t_{n+1} - t_n$.

This procedure results in system of nonlinear equations for y_{n+1} at each step. To obtain the solution from time t_n to time t_{n+1}, the system of equations (4) should be solved.

There are two main assumptions to solve the system (4). The initial value $y(t_0)$ is known and t (time) is the independent variable.

In practical applications, if the time interval, in which the system has to be considered, is known, it can be scaled to the interval $[0, 1]$.

The presented methodology leads to the following equation

$$F(\chi) = 0. \tag{5}$$

This equation is very general and often found in scientific and engineering computing areas. It was assumed, that the function F is considered, where $F : \mathcal{R}^n \to \mathcal{R}^n$ is a nonlinear mapping with the following properties:

(1) There exists a point $\chi^* \in \mathcal{R}^n$ with $F(\chi^*) = 0$.

(2) F is continuously differentiable in a neighborhood of χ^*.

(3) The Jacobian matrix $F'(\chi^*) \equiv \mathcal{J}(\chi^*)$ is nonsingular and for

$$F(\chi) = [F_1, F_2, \cdots, F_n] \tag{6}$$

and

$$\chi \in \mathcal{R}^n, \tag{7}$$

the (i, j)th element (ith row, jth column) of the Jacobian matrix is calculated as

$$\mathcal{J}_{i,j} = \frac{\partial F_i(\chi)}{\partial \chi_j}. \tag{8}$$

There have been a lot of methods for solving the nonlinear equations (5). The most popular and important are both the Newton and different variations of the inexact Newton methods [18].

III. Matrix-free Newton-Krylov method

The matrix-free Newton-Krylov method stands the iterative approach consisting of some nested levels, generally, from two to four. The name of the method come from the primary levels, which are *the Newton correction step* and the loop building up *the Krylov subspace*, out of which each Newton correction is computed [15].

In some applications two additional levels are present. There is *a preconditioner* in the interior to the Krylov loop, and, outside of the Newton loop, *a globalization method* is often required.

A. Newton method

The Newton iteration for $F(\chi) = 0$ derives from a multivariate Taylor expansion about a current point χ_k

$$F(\chi_{k+1}) = F(\chi_k) + F'(\chi_k)(\chi_{k+1} - \chi_k) + \cdots. \tag{9}$$

Neglecting the terms of the higher-order curvature and setting the left-hand side to zero yields a strict Newton method. It is as an iterative process of solving the sequence of the linear systems

$$\mathcal{J}(\chi_k)\delta\chi_k = -F(\chi_k), \tag{10}$$

to obtain $\delta\chi_k$ and to determine

$$\chi_{k+1} = \chi_k + \delta\chi_k, \qquad k = 0, 1, \cdots, \tag{11}$$

where the starting point χ_0 is given, $F'(\chi)$ is a vector-valued function of nonlinear residuals, $\mathcal{J}(\chi)$ is the Jacobian matrix associated with $F'(\chi)$, χ stands the state vector to be found, and k is a nonlinear iteration index.

The Newton iteration is terminated based on a required decrease in the norm of the nonlinear residual

$$\frac{\|F(\chi_k)\|}{\|F(\chi_0)\|} < \Delta_{res}, \tag{12}$$

and a sufficiently small Newton update

$$\frac{\|\delta\chi_k\|}{\|\chi_k\|} < \Delta_{update}. \tag{13}$$

In a scalar example, there is a one-to-one mapping between grid points and rows in the Jacobian. But forming each element of $\mathcal{J}$ requires taking analytic or discrete derivatives of the system of equations with respect to χ. This can be both time consuming and possible source of error for many problems in control and optimization of the biotechnological processes.

B. Krylov method

Krylov subspace methods are approaches for solving large-scale linear systems. They are projection or generalized projection methods for solving

$$A\chi = b, \tag{14}$$

using the Krylov subspace $\mathcal{K}_j$ defined as

$$\mathcal{K}_j = span(r_0, Ar_0, A^2r_0, \cdots, A^{j-1}r_0), \tag{15}$$

where $r_0 = b - A\chi_0$.

These methods require only matrix-vector products, not the individual elements of the matrix A, to perform the iteration. This is the key to their use with the Newton method.

C. Matrix-free Newton-Krylov methods

In the matrix-free Newton–Krylov approach, a Krylov method is used to solve the linear system of equation given by eq. (10). For the Newton step, an initial linear residual r_0 is defined, and an initial guess $\delta\chi_0$ is given

$$r_0 = -F(\chi) - \mathcal{J}(\chi)\delta\chi_0. \tag{16}$$

The nonlinear iteration index k has been omitted, because the Krylov iteration is performed at a fixed k. Let j be the Krylov iteration index. Since the Krylov solution is a Newton correction, and a locally optimal move was just made in the direction of the previous Newton correction, the initial iterate for the Krylov iteration for $\delta\chi_0$ is typically zero. This is asymptotically a reasonable guess in the context of the Newton step, as the converged value for $\delta\chi_0$ should approach zero in late Newton iterations.

When the Generalized Minimal RESidual method (GMRES) is used, in the jth iteration $\|\mathcal{J}\delta\chi_j + F(\chi)\|_2$ is minimized within a subspace of small dimension, relative to the number of unknowns, in a least-square sense [20]. $\delta\chi_j$ is drawn from the subspace spanned by the Krylov vectors, $\{r_0, \mathcal{J}r_0, \mathcal{J}^2 r_0, \cdots, \mathcal{J}^{j-1}r_0\}$, and can be written as

$$\delta\chi_j = \sum_{i=0}^{j-1} \beta_i \mathcal{J}^i r_0, \tag{17}$$

where the scalars β_i minimize the residual.

Upon examining eq. (17) one can see, that GMRES requires the Jacobian only in the form of the matrix-vector products, which may be approximated by

$$\mathcal{J}v \approx [F(\chi + \varepsilon v) - F(\chi)]/\varepsilon, \tag{18}$$

where ε is a small perturbation.

Equation (18) is a first order Taylor series expansion approximation to the product of the Jacobian $\mathcal{J}$ and a vector v.

In a simple case, when the two coupled nonlinear equations are considered $F_1 = (\chi_1, \chi_2) = 0, F_2 = (\chi_1, \chi_2) = 0$, the Jacobian matrix takes a form

$$\mathcal{J} = \begin{bmatrix} \frac{\partial F_1}{\partial \chi_1} & \frac{\partial F_1}{\partial \chi_2} \\ \\ \frac{\partial F_2}{\partial \chi_1} & \frac{\partial F_2}{\partial \chi_2} \end{bmatrix}. \tag{19}$$

The matrix-free Newton-Krylov method does not require the formation of this matrix. Instead, a result vector, that approximates this matrix multiplied by a vector, was formed.

$$\frac{F(\chi + \varepsilon v) - F(\chi)}{\varepsilon} = \begin{bmatrix} \frac{F_1(\chi_1+\varepsilon v_1, \chi_2+\varepsilon v_2) - F_1(\chi_1,\chi_2)}{\varepsilon} \\ \\ \frac{F_2(\chi_1+\varepsilon v_1, \chi_2+\varepsilon v_2) - F_2(\chi_1,\chi_2)}{\varepsilon} \end{bmatrix}. \tag{20}$$

Approximation of $F(\chi+\varepsilon v)$ with a first order Taylor series expansion about χ takes a form

$$F'(\chi_1, \chi_2) \approx \begin{bmatrix} \frac{F_1(\chi_1,\chi_2) + \varepsilon v_1 \frac{\partial F_1}{\partial \chi_1} + \varepsilon v_2 \frac{\partial F_1}{\partial \chi_2} - F_1(\chi_1,\chi_2)}{\varepsilon} \\ \\ \frac{F_2(\chi_1,\chi_2) + \varepsilon v_1 \frac{\partial F_2}{\partial \chi_1} + \varepsilon v_2 \frac{\partial F_2}{\partial \chi_2} - F_2(\chi_1,\chi_2)}{\varepsilon} \end{bmatrix}, \tag{21}$$

which simplifies

$$\mathcal{J}v = \begin{bmatrix} v_1 \frac{\partial F_1}{\partial \chi_1} + v_2 \frac{\partial F_1}{\partial \chi_2} \\ \\ v_1 \frac{\partial F_2}{\partial \chi_1} + v_2 \frac{\partial F_2}{\partial \chi_2} \end{bmatrix}. \tag{22}$$

The error in this approximation is proportional to ε.

The most attractive advantages of the matrix-free approach is a Newton-like nonlinear convergence without costs of forming and storing the true Jacobian. In practice, one forms a matrix for preconditioning purposes. However, the matrices employed in preconditioning can be simpler than true Jacobian of the problem, so the algorithm is properly said to be *Jacobian-free* [15].

Since the use of an iterative technique to solve eq. (10) does not require the exact solution of the linear system, the resulting algorithm is categorized as *the inexact Newton method*.

IV. Inexact Newton method

The Newton method is attractive because its quadratically rate of convergence from any sufficiently good initial point. But the computational cost can be expensive, especially, when the size of the problem is very large. In each iteration step the Newton equation

$$F(\chi_k) + \mathcal{J}(\chi_k)\delta\chi_k = 0 \tag{23}$$

should to be solved. Here χ_k denotes the current iterate, and $\mathcal{J}(\chi_k)$ is the Jacobian matrix of $F(x)$ at point χ_k. The solution $\delta\chi_k^N$ of the Newton equation is known as the Newton correction or the Newton step. Once the Newton step is obtained, the next iterate is given by

$$\chi_{k+1} = \chi_k + \delta\chi_k^N. \tag{24}$$

The inexact Newton method is a generalization of the Newton method [8], [12]. It is any method, which for given an initial guess χ_0, generates a sequence χ_k of approximations to χ^* as in Algorithm 1.

ALGORITHM 1. The inexact Newton method
1. Given $\chi_0 \in \mathcal{R}^n$
2. For $k = 0, 1, 2, \cdots$ until χ_k converges
   2.1 Choose some $\eta_k \in [0, 1)$
   2.2 Inexactly solve the Newton equation (10) and obtain a step $\delta\chi_k$, such that
   $\|F(\chi_k) + \mathcal{J}(\chi_k)\delta\chi_k\| \le \eta_k \|F(\chi_k)\|$. $(\star)$
   2.3 Let $\chi_{k+1} = \chi_k + \delta\chi_k$.

In the Algorithm 1, η_k is the forcing term in the kth iteration, $\delta\chi_k$ is the inexact Newton step and $(\star)$ is the inexact Newton condition.

In each iteration step of the inexact Newton method, a real number $\eta_k \in [0, 1)$ should be chosen. Then the inexact

Newton step $\delta\chi_k$ was obtained by solving the Newton equation approximately.

Since $F(\chi_k) + \mathcal{J}(\chi_k)\delta\chi_k$ is both residual of the Newton equations and the local linear model of $F(\chi)$ at χ_k, the inexact Newton condition ($\star$) reflects both the reduction in the norm of the local linear model and certain accuracy in solving the Newton equations. In this way, the role of forcing terms is to control the accuracy degree of solving the Newton equations. In particular, if $\eta_k = 0$ for all k, then the inexact Newton method is reduced into the Newton method.

The inexact Newton method, like the Newton method, is locally convergent.

Theorem 1 ([8]): Assume that $F : \mathcal{R}^n \to \mathcal{R}^n$ is continuously differentiable, $\chi^* \in \mathcal{R}^n$ such that $\mathcal{J}(\chi^*)$ is nonsingular. Let $0 < \eta_{max} < \beta < 1$ be the given constants. If the forcing terms η_k in the inexact Newton method satisfy $\eta_k \leq \eta_{max} < \beta < 1$ for all k, then there exists $\varepsilon > 0$, such that for any $\chi_0 \in N_\varepsilon(\chi^*) \equiv \{\chi : \|\chi - \chi^*\| < \varepsilon\}$, the sequence $\{\chi_k\}$ generated by the inexact Newton method converges to χ^*, and

$$\|\chi_{k+1} - \chi^*\|_* \leq \beta\|\chi_k - \chi^*\|_*, \tag{25}$$

where $\|v\|_* = \|\mathcal{J}(\chi^*)v\|$.

If the forcing terms $\{\eta_k\}$ in the inexact Newton method are uniformly strict less than 1, then by Theorem 1, the method is locally convergent. The following result states the convergence rate of the inexact Newton method.

Theorem 2 ([8]): Assume that $F : \mathcal{R}^n \to \mathcal{R}^n$ is continuously differentiable, $\chi^* \in \mathcal{R}^n$ such that $\mathcal{J}(\chi^*)$ is nonsingular. If the sequence $\{\chi_k\}$ generated by the inexact Newton method converges to χ^*, then

(1) χ_k converges to χ^* superlinearly when $\eta_k \to 0$;

(2) χ_k converges to χ^* quadratically if $\eta_k = \mathcal{O}(\|F(\chi_k)\|)$ and $\mathcal{J}(\chi)$ is Lipschitz continuous at χ^*.

Theorem 2 indicates, that the convergence rate of the inexact Newton method is determined by the choice of the forcing terms.

Various ways for selection the forcing terms have been widely discussed and tested in [1] and [13].

V. Case study

As the case study a fed-batch reactor for the production of penicillin [2] was considered. The objective was to maximize the amount of penicillin using the feed rate as the control variable. The duration of the process was specified at 120 hours.

The mathematical statement of the dynamical optimization problem is as follows.

Find $u(t)$ and t_f over $t \in [t_0, t_f]$ to maximize

$$J = x_2(t_f) \cdot x_4(t_f) \tag{26}$$

subject to differential-algebraic system

$$\frac{dx_1}{dt} = h_1 x_1 - u\left(\frac{x_1}{500x_4}\right), \tag{27}$$

$$\frac{dx_2}{dt} = h_2 x_1 - 0.01x_2 - u\left(\frac{x_2}{500x_4}\right), \tag{28}$$

$$\frac{dx_3}{dt} = -h_1\frac{x_1}{0.47} - h_2\frac{x_1}{1.2} - x_1\frac{0.029x_3}{0.0001 + x_3} + \frac{u}{x_4}\left(1 - \frac{x_3}{500}\right), \tag{29}$$

$$\frac{dx_4}{dt} = \frac{u}{500}, \tag{30}$$

$$h_1 = 0.11\left(\frac{x_3}{0.006x_1 + x_3}\right), \tag{31}$$

$$h_2 = 0.0055\left(\frac{x_3}{0.0001 + x_3(1 + 10x_3)}\right), \tag{32}$$

where x_1, x_2 and x_3 are the biomass, penicillin and substrate concentration (g/L), and x_4 is the volume (L). The initial conditions are

$$x(t_0) = [1.5 \quad 0 \quad 0 \quad 7]^T. \tag{33}$$

There are several path constraints for state variables

$$0 \leq x_1 \leq 40, \tag{34}$$

$$0 \leq x_2 \leq 25, \tag{35}$$

$$0 \leq x_3 \leq 10. \tag{36}$$

The upper and lower bounds on the control variable (feed rate of substrate) are

$$0 \leq u \leq 50. \tag{37}$$

The control problem of the fed-batch fermentor for penicillin production was solved with the matrix-free inexact Newton method, presented in the article.

At first, the overall time domain was divided into 1200 equidistant intervals. The resulting model consisted of 7200 nonlinear algebraic equations and the same number of variables and it was of the form

$$x_{1,n+1} - x_{1,n} - \Delta t\Big(h_{1,n+1}x_{1,n+1} - u\frac{x_{1,n+1}}{500x_{4,n+1}}\Big) = 0, \tag{38}$$

$$\vdots$$

$$x_{4,n+1} - x_{4,n} - \Delta t\frac{u}{500} = 0, \tag{39}$$

$$h_{1,n+1} - \Delta t\Big(0.11 \times \frac{x_{3,n+1}}{0.006x_{1,n+1} + x_{3,n+1}}\Big) = 0, \tag{40}$$

$$h_{2,n+1} - \Delta t\Big(0.0055 \times \frac{x_{3,n+1}}{0.0001 + x_{3,n+1}(1 + 10x_{3,n+1})}\Big) = 0, \tag{41}$$

for $\quad n = 0, 1, \cdots, 1200.$

The initial conditions were known only for the first stage $n = 0$. In this way, there are 7200 decision variables connected with initial values for both differential and algebraic state variables. There is one variable, which is the assumed value of the feed rate and represents the control variable.

The initial values for the decision variables were as follows

$$\chi_{1,x_{1,1}}, \cdots, \chi_{1200,x_{1,1200}} = 1.5, \tag{42}$$

$$\chi_{1201,x_{2,1}}, \cdots, \chi_{2400,x_{2,1200}} = 0.0, \tag{43}$$

$$\chi_{2401,x_{3,1}}, \cdots, \chi_{3600,x_{3,1200}} = 0.0, \tag{44}$$

$$\chi_{3601,x_{4,1}}, \cdots, \chi_{4800,x_{4,1200}} = 7.0, \tag{45}$$

$$\chi_{4801,h_{1,1}}, \cdots, \chi_{6000,h_{1,1200}} = 10.0, \tag{46}$$

$$\chi_{6001,h_{2,1}}, \cdots, \chi_{7200,h_{2,1200}} = 10.0, \tag{47}$$

In the simulations the following rule choice of the forcing terms was used

$$\eta = \min\left\{\frac{1}{k_{iter}+2}, \|F(\chi_{iter})\|\right\}, \tag{48}$$

where $iter$ denotes the number of the previously iterate [9].

For the constant control function u, the final value of the objective function was $81.1943g$. The obtained value of the control function was $u_{const} = 12.5000$. The assumed duration of the whole process was adjusted to 120 hours. Simulations were performed with the accuracy $\Delta_{res} = \Delta_{update} = 10^{-6}$. There are the optimal trajectories of both the biomass and penicillin concentrations in the Fig. 1.

In the simulation for solving the Newton equation (10), the Generalized Minimal RESidual method (GMRES) was used. In GMRES, the Arnoldi basis vector form the trial subspace out of which the solution was constructed. One matrix-vector product was required per iteration to create each new trial vector, and the iterations are terminated based on a by-product estimate of the residual that does not require explicit construction of intermediate residual vector of solutions. It was a major beneficial feature of the algorithm.

In the case study, the Jacobian matrix in the Newton equation consisted on more than $50 \cdot 10^6$ cells. It means, that the matrix-vector product would be impossible to obtain by ordinary methods.

The first proposition was to use the sparsity of the matrix, especially for the storage and speed-up of the computations. In the Jacobian matrix only 0.048% elements has another value than zero. The second proposition is the Jacobian-free approach.

These two remarks, enables us to solve the fed batch fermentor for penicillin production described by the nonlinear differential-algebraic equations.

The numerical simulations were executed in MATLAB using Wroclaw Center for Networking and Supercomputing

VI. Conclusion

In this paper the new approach for solving the nonlinear differential-algebraic equations in the *fully-implicit* form was presented. The method consists of two main remarks. The first, that the Newton equation can be solved inexactly. The appropriate choice of the forcing terms to obtain the well behaved inexact Newton method preserve locally the superlinearly convergence rate. The second remark is that, the matrix-free approach enables us to consider a large-scale systems

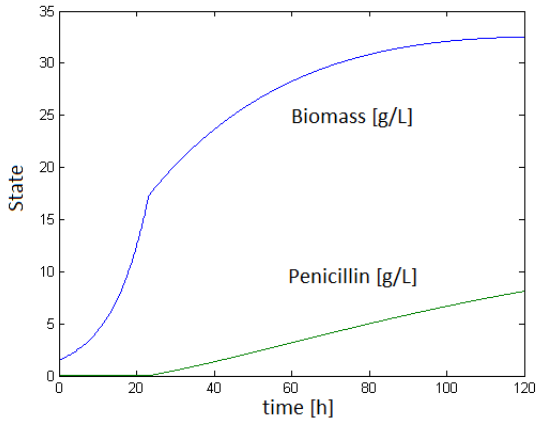


Fig. 1. The optimal trajectories of both the biomass and penicillin concentrations.

with thousands of variables. The sparse representation of the Jacobian matrix and a function, which calculate the matrix-vector product effectively, makes large-scale computations possible.

The algorithm was tested on the nonlinear DAE system, which described the fed batch fermentor for penicillin production. The discretized large-scale model consisted on 7200 nonlinear algebraic equations and the same number of variables.

The presented approach can be applied in real-life industrial plants, to optimize and control the biotechnological processes [7], [23]. The high degree of utilization of resources ensures a high profit and negligible waste.

At the next step, the new preconditioned Jacobian-free algorithms, which could solve large-scale optimization problems efficiently, will be studied and adjusted for new challenges in solving the dynamical optimization problems [11].

Acknowledgment

The project was supported by the grant Młoda Kadra B30036 at Wrocław University of Technology.

References

[1] H.-B. An, Z.-Y. Mo, X.-P. Liu. 2007. A choice of forcing terms in inexact Newton method. Journal of Computational and Applied Mathematics. 200:47-60, http://dx.doi.org/10.1016/j.cam.2005.12.030.

[2] J.R. Banga, E. Balsa-Canto, C.G. Moles, A.A. Alonso. 2005. Dynamic optimization of bioprocesses: Efficient and robust numerical strategies. Journal of Biotechnology. 117:407-419, http://dx.doi.org/10.1016/j.jbiotec.2005.02.013.

[3] J.T. Betts. 2010. Practical Methods for Optimal Control and Estimation Using Nonlinear Programming, Second Edition. SIAM, Philadelphia, http://dx.doi.org/10.1137/1.9780898718577.

[4] L.T. Biegler. 2010. Nonlinear Programming. Concepts, Algorithms and Applications to Chemical Processes. SIAM, Philadelphia, http://dx.doi.org/10.1137/1.9780898719383.

[5] L.T. Biegler, S. Campbell, V. Mehrmann. 2012. DAEs, Control, and Optimization. Control and Optimization with Differential-Algebraic Constraints. SIAM, Philadelphia, http://dx.doi.org/10.1137/9781611972252

[6] K.E. Brenan, S.L. Campbell, L.R. Petzold. 1996. Numerical Solution of Initial- Value Problems in Differential-Algebraic Equations. SIAM, Philadelphia, http://dx.doi.org/10.1137/1.9781611971224

[7] R. Brunet, G. Guillen-Gosalbez, L. Jimenez. 2010. Cleaner design of single-product biotechnological facilities through the integration of process simulation, multiobjective optimization, life cycle assessment, and principal compopnent analysis. Ind. Eng. Chem. Res. 51:410-424, http://dx.doi.org/10.1021/ie2011577.

[8] R.S. Dembo, S.C. Eisenstat, T. Steihaug. 1982. Inexact Newton Methods. SIAM Journal on Numerical Analysis. 19:400-408, http://dx.doi.org/10.1137/0719025.

[9] R.S. Dembo, T. Steihaug. 1983. Truncated-Newton algorithm for large-scale unconstrained optimization. Mathematical Programming. 26:190-212, http://dx.doi.org/10.1007/BF02592055.

[10] M. Diehl, H.G. Bock, J.P. Schlöder, R. Findeisen, Z. Nagy, F. Allgower. 2002. Real-time optimization and nonlinear model predictive control of processes governed by differential-algebraic equations. Journal of Process Control. 12:577-585, http://dx.doi.org/10.1016/S0959-1524(01)00023-3.

[11] P. Drąg, K. Styczeń. 2012. A Two-Step Approach for Optimal Control of Kinetic Batch Reactor with electroneutrality condition. Przegląd Elektrotechniczny. 6:176-180.

[12] S.C. Eisenstat, H.F. Walker. 1994. Globally convergent inexact Newton methods. SIAM Journal on Optimization. 4:393-422, http://dx.doi.org/10.1137/0804022.

[13] S.C. Eisenstat, H.F. Walker. 1996. Choosing the forcing terms in an inexact Newton method. SIAM Journal on Scientific Computing. 17:16-32, http://dx.doi.org/10.1137/0917003.

[14] C.W. Gear. 1971. The simultaneous numerical solution of differential-algebraic equations. IEEE Transactions on Circuit Theory. 18:89-95, http://dx.doi.org/10.1109/TCT.1971.1083221.

[15] D.A. Knoll, D.E. Keyes. 2004. Jacobian-free Newton-Krylov methods: a survey of approaches and applications. Journal of Computational Physics. 193:357-397, http://dx.doi.org/10.1016/j.jcp.2003.08.010.

[16] M. Kwiatkowska. 2012. Antimicrobial PVC composites. Processing technologies and functional properties of polymer nanomaterials for food packaging : International COST Workshop, Wroclaw, Poland, September 11-12, pp. 40-41.

[17] D. Niu, M. Jia, F. Wang, D. He. 2013. Optimization of nosiheptide fed-batch fermentation process based on hybrid model. Ind. Eng. Chem. Res. 52:3373-3380, http://dx.doi.org/10.1021/ie3022169.

[18] J. Nocedal, S.J. Wright. 2006. Numerical Optimization. Second Edition. Springer, New York, http://dx.doi.org/10.1007/978-0-387-40065-5

[19] L. Petzold. 1982. Differential/Algebraic Equations are not ODEs. SIAM Journal on Scientific Computing. 3:367-384, http://dx.doi.org/10.1137/0903023.

[20] Y. Saad, M. H. Schultz. 1986. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Stat. Comput. 7:856-869, http://dx.doi.org/10.1137/0907058.

[21] D.L. Stoner, A.P. Poloski, J.A. Johnson, C.R. Tolle. 2001. Optimization and Control of Dynamic Bioprocesses. Organic Process Research and Development. 5:299-307, http://dx.doi.org/10.1021/op0100091.

[22] V.S Vassiliadis, R.W.H. Sargent, C.C. Pantelides. 1994. Solution of a Class of Multistage Dynamic Optimization Problems. 1. Problems without Path Constraints. Ind. Eng. Chem. Res. 33:2111-2122, http://dx.doi.org/10.1021/ie00033a014.

[23] S.R.R. Vetukuri, L.T. Biegler, A. Walther. 2010. An inexact trust-region algorithm for the optimization of periodic adsorption processes. Ind. Eng. Chem. Res. 49:12004-12013, http://dx.doi.org/10.1021/ie100706c.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 603–612
DOI: 10.15439/2014F222
ACSIS, Vol. 2

Finite Element Numerical Integration on Xeon Phi coprocessor

Filip Krużel
Cracow University of Technology
ul. Warszawska 24, 31-155 Kraków, Poland
Email: fkruzel@pk.edu.pl

Krzysztof Banaś
AGH University of Science and Technology
al. Mickiewicza 30, 30-059 Kraków, Poland
Email: kbanas@pk.edu.pl

***Abstract*—In the present article we describe the implementation of the finite element numerical integration algorithm for the Xeon Phi coprocessor. The coprocessor is an extension of the idea of the many-core specialized unit for calculations and, by assumption, its performance has to be competitive with the current families of GPUs. Its main advantage is the built-in set of 512-bit vector registers and the ease of transferring existing codes from normal x86 architectures. However, the differences between standard x86 architectures and Xeon Phi do not guarantee performance portability. We choose an alternative approach and, instead of porting standard multithreaded code, we adapt to Xeon Phi previously developed OpenCL algorithms for finite element numerical integration. The algorithm is tested for standard FEM approximations of selected problems. The obtained timing results allow to compare the performance of the OpenCL kernels executed on the Xeon Phi and the contemporary GPUs.**

I. Motivation

IN RECENT years there has been a noticeable increase of popularity of programming with the use of graphic cards. Their computing power allows for significant acceleration of calculations for properly implemented programs. However, there is a price to be paid, in the form of complex programming model with a complicated memory organization [1], [2]. Huge performance of GPUs can be seriously limited due to data transfers between different memory levels. Therefore, an important step is to design an algorithm that takes into account characteristics of memory access mechanisms for a particular architecture.

The development of multi-core architectures has resulted in many interesting ideas for further evolution of hardware for scientific and technical calculations. GPUs are an example of massively multi-core microprocessors with the large number of relatively simple cores equipped with small amount of memory. Another development trend in microprocessor architecture is to increase the amount and width of vector execution units within a single processor, clearly visible in recent general purpose cores [3]. The other idea was to combine the architecture of a general purpose processor with SIMD units encountered in graphics cards. The first example which achieved a fairly considerable popularity is the CellBE architecture used in Sony Playstation 3 consoles and successfully adapted to the scientific purposes as PowerXCell 8i processor [4]. In our previous studies, we focused on the development of an algorithm for the finite element numerical integration on the aforementioned processors which resulted in the development of highly efficient implementations for higher order elements of the Discontinous Galerkin approximation [5]. At the same time authors developed a version for the graphic card which was then successfully redesigned to use a standard approximation of the finite element method [6]. The resulting version of the algorithm has been tested on different types of graphic cards and the results of these tests will soon be presented in [7]. Both mentioned architectures can be considered as predecessors of the new Intel Xeon Phi architecture. This architecture combines large number of cores with wide vector units in each core. Opposite to standard GPUs, coprocessor cores are less numerous and their complexity lies in between standard, general purpose cores and simple GPU cores. As in GPUs, Xeon Phi shares the same way of memory organization and therefore all codes developed for graphic cards should be easily adapted to coprocessor architecture, but as in all types of such architectures, data movement between different levels of memory may become an issue of primary importance [8]. With the introduction of Intel Xeon Phi numerical coprocessors, there is a need to test the previously developed algorithms on the new architecture and verify whether the widely advertised adaptability of existing codes also applies to the transition from GPU to coprocessor.

II. Numerical Integration Algorithm

Numerical integration algorithm is one of the most important parts of the finite element method codes. FEM assumes the division of the whole computational domain into elements for which the integrals, corresponding to pairs of element basis functions, are calculated and the results are collected in local, element stiffness matrices. Local load vectors are also obtained through integration of corresponding terms. Final structure of the formula to calculate the example entry to the element stiffness matrix depends on the form of the weak statement for the considered problem and can look as (1).

This work is developed within the project DEC-2011/01/B/ST6/00674 financed by Polish National Science Centre

$$A^e_{i_s j_s} = \int_{\Omega_e} \sum_{i_D, j_D} C^{i_D j_D} \frac{\partial \phi_{i_s}}{x_{i_D}} \frac{\partial \phi_{j_s}}{x_{j_D}} d\Omega \quad (1)$$

In the formula above, $C^{i_D j_D}$ are coefficients that depend on the problem with $i_D, j_D = 0,1,2,3$ and ϕ_{i_s}, ϕ_{j_s} are global basis functions.

In order to calculate the integrals we need to perform the change of variables, which means that the integration is made for a particular type of reference element. The transformation from the reference element to the real element is denoted by $x(\xi)$. For a reference element we use shape functions instead of global basis functions and apply one of the forms of quadrature. In our case, we used one of the most popular Gaussian quadrature. This quadrature allows for the transformation of the integral to the sum over integration points within the reference domain. Number of integration points is dependent on the required accuracy of the calculations and the type of the reference element. For N_Q integration points with coordinates ξ^Q and weights w^Q we can transform the integral (1) to the sum (2).

$$A^e_{i_S j_S} = \sum_{i_Q}^{N_Q} \sum_{i_D, j_D} C^{i_D j_D} \frac{\partial \hat{\phi}_{i_S}}{\partial \xi^Q} \frac{\partial \xi^Q}{x_{i_D}} \frac{\partial \hat{\phi}_{j_S}}{\partial \xi^Q} \frac{\partial \xi^Q}{x_{j_D}} det\, \boldsymbol{J}_{T_e} w^Q \quad (2)$$

Where $\hat{\phi}_{i_s}$ and $\hat{\phi}_{j_s}$ are shape functions and $\boldsymbol{J}_{T_e}$ is the Jacobian matrix of transformation $x(\xi)$.

Performance of numerical integration algorithm depends greatly on the problem being solved (weak formulation) and the approximation method employed. With the use of standard linear approximation the time of the creation of element stiffness matrix is relatively small. From the computational point of view, numerical integration algorithm consists of multiple independent calculations for each element. For this reason, the computational cost increases with growing number of elements. Calculated integrals correspond to the different terms in the weak formulation of the problem for which there is a need to define the matrix of coefficients for integration. Therefore, for the various problems we obtain different combinations of integration components for partial integrals of the test functions.

The problem dependent contribution mainly consists of the set of coefficient for numerical integration. Besides standard i_D and j_D indices that corresponds to the different spatial derivatives for test and trial functions, there can be also second pair of indices i_E, j_E. This indices are introduced, because for vector problems, the same approximation can be used for different unknowns in the solved system of partial differential equations (PDEs). Hence, for each combination of i_D and j_D there may be a small matrix of coefficients with the N_E dimension equal to the number of equations in solved system of PDEs. Moreover, in the most general cases there may be different values of coefficients at each integration point. Hence for the generic numerical integration algorithm array of coefficient should be considered in a form $C^{i_Q i_D j_D i_E j_E}$. The problem dependent indices indicate that element stiffness matrix entry is also dependent on the problem solved. Hence, we can define full equation for our computations, with the definition of $\frac{\partial \hat{\phi}_i}{\partial \xi}$ as $\psi^{i_Q i_D i_S}$ and $det\, \boldsymbol{J}_{T_e} w^Q$ as vol^{i_Q} (3)

$$A^e_{i_E j_E i_S j_S} = \sum_{i_Q}^{N_Q} \sum_{i_S, j_S}^{N_S} \sum_{i_E, j_E}^{N_E} \sum_{i_D, j_D}^{N_D} C^{i_Q i_D j_D i_E j_E} \psi^{i_Q i_D i_S} \psi^{j_Q j_D j_S} vol^{i_Q} \quad (3)$$

The corresponding right hand side vector is calculated using the formula (4)

$$b^e_{i_E i_S} = \sum_{i_Q}^{N_Q} \sum_{i_S}^{N_S} \sum_{i_E}^{N_E} \sum_{i_D}^{N_D} D^{i_Q i_D i_E} \psi^{i_Q i_D i_S} vol^{i_Q} \quad (4)$$

As the conclusion of the numerical integration problem definition we provide the algorithm for computing stiffness matrices and load vectors for a set of elements of the same type and the order of approximation :

1: read quadrature data ξ_Q and weights w_Q for the reference element of particular type.
2: **for** $e=1$ **to** N_e **do**
3: read problem dependent coefficients common for all integration points (e.g. material data, previous iterations (or time steps) degrees of freedom etc.)
4: read element geometry data for $x(\xi)$ transformation
5: initialize element stiffness matrix $A^e_{i_E j_E i_S j_S}$ and element load vector $b^e_{i_E i_S}$
6: **for** $i_Q=1$ **to** N_Q **do**
7: read or calculate(on a basis of the coordination of the integration points) values of shape functions and their derivatives with respect to their local coordinates for a given integration point.
8: read or calculate jacobian matrix, its determinant and inverse.
9: calculate vol^{i_Q}
10: using the jacobian matrix calculate derivatives of shape functions $\hat{\phi}_{i_Q}$ with respect to the global coordinates for a given integration point
11: basing on the values of unknowns obtained through the use of $\hat{\phi}_{i_Q}$ calculate the C^{i_Q} and D^{i_Q} coefficients for a given quadrature point
12: **for** $i_S=1$ **to** N_S **do**
13: **for** $j_S=1$ **to** N_S **do**
14: **for** $i_E=1$ **to** N_E **do**
15: **for** $j_E=1$ **to** N_E **do**
16: **for** $i_D=0$ **to** N_D **do**

```
17:            for jD=0 to ND do
18:              Ae[iS][jS][iE][jE]+=
                                  C[iQ][iE][jE][iD][jD]×
                                  ψ[iQ][iS][iD]×
                                  ψ[jQ][jS][jD]×vol[iQ]
19:            end for (jD)
20:          end for (iD)
21:          if iS=jS && iE=jE then
22:            for iD=0 to ND do
23:              be[iS][iE]+=D[iQ][iE][iD]×ψ[iQ][iS][iD]
24:          end for (jE)
25:        end for (iE)
26:      end for (jS)
27:    end for (iS)
28:  end for (iQ)
29: end for (e)
```

As we see from algorithm above, we can either read or compute most of the necessary components for numerical integration. This leads us to the conclusion that we can steer the amount of data sent from the memory and the amount of computations, depending on the available hardware resources and the problem solved.

In our case, we focused on the problem of convection-diffusion for N_E = 0 in two cases - one with simple Laplace equation, where the coefficient matrix C is sparse and coefficients appear only on the main diagonal in the case of $i_D = j_D$ (3 coefficients for stiffness matrices, one for the right hand side (RHS) vector) and a second with enhanced convection-diffusion problem for the full sixteen coefficients for stiffness matrix and four for RHS vector. Furthermore, for solving the Laplace task all coefficient were the same for all Gaussian integration points for stiffness matrix and different for the RHS vector. In the second, convection-diffusion task, all coefficients were constant for all Gauss points. For our reference elements we use prisms with 6 degrees of freedom. Our assumptions are illustrated by the data in Table I.

TABLE I.
NUMBER OF PARAMETERS FOR NUMERICAL INTEGRATION OF PRISMATIC ELEMENT

N_Q		6
N_S		6
N_geo_dofs		6
Nr_coeff_SM	**Laplace**	3
Nr_coeff_LV		6
Nr_coeff_SM	**Convection-diffusion**	16
Nr_coeff_LV		4

For optimization of the data transfer we need to decide which coefficients should be computed on the host system side and which on the accelerator side. This depends on the available resources and the type of the solved problem. Amount of data to send/store for one element can be observed in Table II.

TABLE II.
NUMBER OF DATA ELEMENTS FOR ARRAYS USED IN NUMERICAL INTEGRATION FOR PRISMATIC ELEMENTS

Gauss data		24
Shape functions at point		24
Shape functions total		144
Geometric data		18
Jacobian terms at point		10
Jacobian terms total		60
Coefficients at point	**Laplace**	4
Coefficients total		9
Coefficients at point	**Convection-diffusion**	20
Coefficients total		20

For the GPU implementation the most important part is a proper way of data transfer organization and utilization of a limited resources. In order to port the code to the Xeon Phi coprocessor we need to reorganize the code, based on the experience gained when implementing the numerical integration for the PowerXCell 8i architecture.

III. INTEL XEON PHI

With the development of multi-core architectures and a simultaneous trend of using the graphics cards for the calculation, an idea came up to combine several different architectures in a single hardware unit whose individual elements would be responsible for processing different type of code fragments. The first device of this type – mentioned earlier PowerXCell 8i processor was unveiled by IBM and was equipped with two core with IBM Power architecture (Power Processing Element) and a few specialized SIMD cores (Synergistic Processing Elements). Its hybrid design allowed for sending to SPE a pieces of code for which you can apply the SIMD paradigm in order to speed up calculations [9]. Truncated version of this processor has been successfully applied for commercial purposes in Sony Playstation 3 consoles and its scientific version was part of the Roadrunner computer which in 2008 exceeded the petaflops performance barrier [10]. PowerXCell 8i processor was a very big step in the development of architecture and despite the discontinuation of its production it has become a base used by other manufacturers for a hardware development for high-performance computing. At the same time Intel was working on its

line of graphics cards codenamed Larabee trying to eliminate the main disadvantage in programming CellBE or GPU architectures, which is complicated programming model. The main features of this architecture was the use of a very wide vector units (512bit), texture units taken from the GPU, the coherence memory hierarchy and compatibility with x86 architecture [11]. On the basis of this project Intel Many Integrated Core (MIC) architecture was developed, which was successfully applied in Intel Xeon Phi coprocessors [12]. These coprocessors are sold as a PCI-express cards (Fig 1.) and are equipped with its own operating system based on Linux, and depending on the version 57-61 cores with hard-

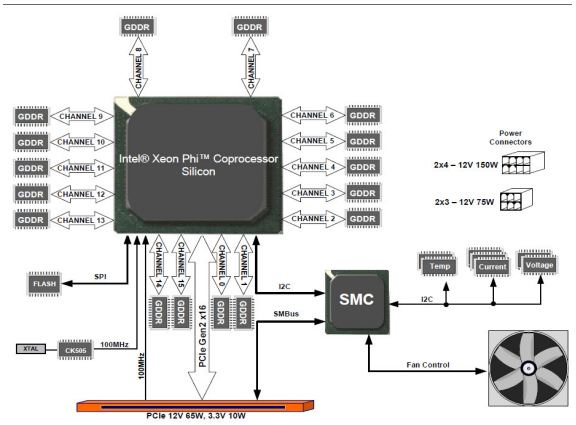


Fig 1. Xeon Phi coprocessor board schematic [13]

ware multithreading support (4 threads per core).

For testing purposes, we used 5110P coprocessor equipped with 8GB of RAM and 60 cores with a speed of 1GHz. However, in order to function properly, a single coprocessor core and 2GB of memory are reserved for the internal operating system which results in a 236 available threads and 6GB of memory for performing the calculations [14]. MIC Architecture cores design is based mainly on the Pentium architecture but it is enhanced with 512-bit vector units. The x86 compatible architecture theoretically allows for easy transfer of existing code to be used on the coprocessor with a significant increase in performance. Fig 2. shows the internal structure of the single coprocessor core.

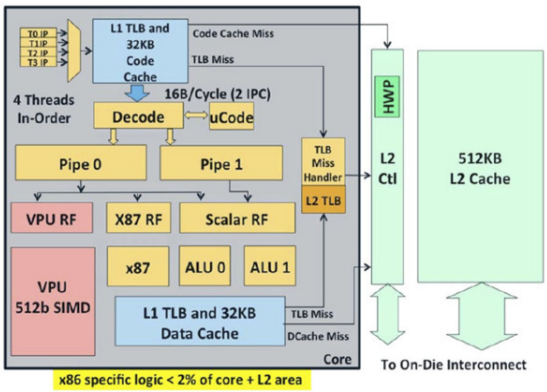


Fig 2. Single coprocessor core [15]

Every core is connected to the ultra fast interface, and thanks to a coherent cache memory, the data between cores

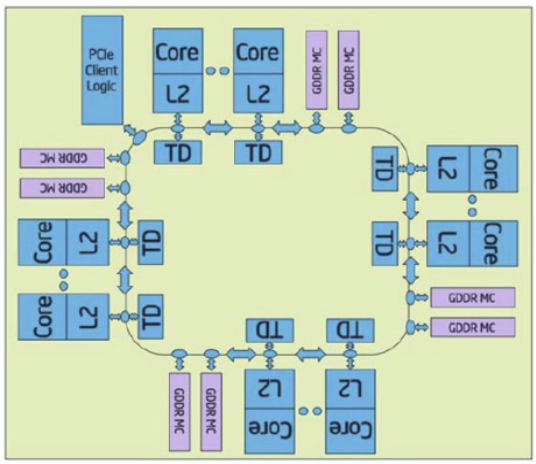


Fig 3. Xeon Phi microarchitecture [15]

are exchanged almost immediately (Fig. 3).

IV. OpenCL Programming Model

OpenCL is a software development platform that supports many kinds of available hardware, from standard CPUs, through hybrid architectures to the GPUs [16]. In recent years this platform has gained popularity due to its portability and similarity to the previously used CUDA programming model developed by Nvidia [17]. OpenCL code is compiled and run for a given platform, representing the environment for code execution. Each platform is equipped with sets of devices of three types: CPU, GPU or Accelerator. For one host system there could be many platforms installed, varying on the vendor and supported devices. Host system runs standard code and manages the execution of an OpenCL code on device. OpenCL code is called a kernel and is written in a slightly modified C language, with the special extensions to manage different types of devices. Each device in the platform is composed of compute units, that are further divided into processing elements. Individual threads are running on processing elements with capabilities depending on the architecture of device. In OpenCL nomenclature all threads are called work items and they are grouped into work groups. This allows for direct hardware mapping for different architectures. OpenCL programming model is shown on Fig. 4. Threads within a single work group execute concurrently and can be synchronized using fast system calls. Moreover, they can share some of the data in their fast shared memory, called local memory in OpenCL nomenclature. Different work-groups are scheduled independently and have

their own resources. OpenCL execution model specifies a

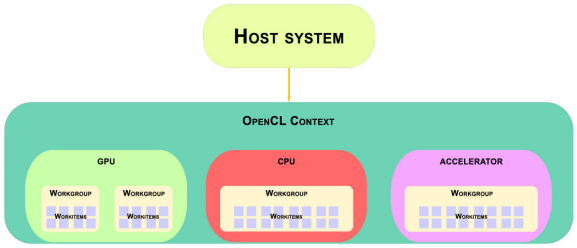


Fig 4. OpenCL programming model

set of events that has to occur in order to run a kernel.

The first phase includes initialization of the OpenCL platform, data structures and checking the available devices. Then the kernel code has to be prepared (read or compile) for running on the devices. Because of the ability of passing arguments to kernels, the space for them has to be allocated on device, before executing a kernel. Host is also responsible for preparing and allocating the space needed for variables and arrays in different types of memory that are explicitly available to programmers. All memory transactions are performed by sending a request to the OpenCL management layer. Then the requests are realized asynchronously to the host code. The same strategy is used to request kernel execution or transfer of data, back from device memory to the host memory [18].

In OpenCL the programmers have several types of memory regions explicitly available to use. Each of memory objects can be created in OpenCL memory model with different mappings to hardware resources. Individual variables defined inside kernel belong to private memory. Each thread has its own copy of each variable, and they can be stored in scalar or vector registers. Other memory regions can be assigned through specific qualifiers. Typical memory regions are divided into three types – global, constant and local. Global memory stores variables that are visible to all threads executing the kernel. Constant memory is also available for all threads but it is only accessible for reading. Variables stored in fast local memory are shared by threads in a single work-group. Because of the portability of created code OpenCL contains procedures that allows for adapting to different platforms and devices [19]. The code can query the environment to get information about many available resources. For our case we compare the available resources of all three types of devices – CPU, GPU and Accelerator. The results are presented in Table III. As we notice CPU and a Xeon Phi cards share the same amount of local and constant memory which indicates the same origins of this architectures. The Tesla K20m card used for testing our GPU implementations of numerical integration has bigger local memory size but less compute units. Hence, one can conclude that it should be slower than the other devices, but OpenCL hardware layer does not provide information on deeper division of compute units into processing elements. OpenCL in both, CPU and Intel MIC architectures treat their cores as a single compute unit but it ignores all CUDA or STREAM cores in GPUs. Our reference Tesla K20m card is equipped with 13 compute units with the Kepler architecture [20] that indicates that we have a massive amount of 192 processing elements per one compute unit, giving total 2496 cores available [21]. Despite of that all three architectures are treated as a direct opponents in the domain of high performance computing. This happens because each of these architectures has its own unique characteristics that allow for direct

TABLE III.
COMPARISON OF DIFFERENT TYPES OF DEVICES AVAILABLE IN OPENCL

OpenCL properties	CPU	GPU	Accelerator
CL_DEVICE_NAME	Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz	Tesla K20m	Intel(R) Many Integrated Core Acceleration Card
CL_DEVICE_VENDOR	Intel(R) Corporation	NVIDIA Corporation	Intel(R) Corporation
CL_DEVICE_VERSION	OpenCL 1.2 (Build 67279)	OpenCL 1.1 CUDA	OpenCL 1.2 (Build 67279)
global memory size (MB)	32083.020	4799.563	5773.180
global max alloc size (MB)	8020.755	1199.891	1924.391
local memory size (kB)	32	48	32
constant memory size (kB)	128	64	128
cache memory size (kB)	256	208	0
cache line size (B)	64	128	0
number of compute units	24	13	236

comparison between them (e.g. architecture of cores, clock speed, vector registers etc.). This determines also, that in order to fully exploit possibilities of the hardware, all existing algorithms should be adapted separately for each of the architectures. The task of numerical integration becomes non-trivial and therefore very interesting from the performance point of view.

V. Numerical Integration on Intel Xeon Phi

For our tests we use ModFEM code - a computational framework developed for solving various scientific and engineering problems by the adaptive finite element method [22]. Due to its modular structure it allows to test different levels of the FEM. Therefore, we can easily separate the numerical integration algorithm and make a parallel versions for different architectures.

For numerical integration algorithm we have several levels of parallelism available. The chosen way of parallelisation will depend on the size of data and the number of calculations in the solved problem. Therefore, we can choose which loop from algorithm 1 should be divided. On first level we can parallelize the outermost loop over elements. Then, we can divide the loop over integration points and subsequently two inner loops over shape functions. In our previous works [5],[6] we have tested several strategies for higher order finite elements. Because of the quite big sizes of computed matrices in the problems described above, we have tested division of the inner loops over shape functions, and also a loop over Gaussian integration points. In the current case we decided to test standard approximation module. Therefore, our stiffness matrices and load vectors are quite small as we can see in Tables I and II. Hence, as the method of parallelization the most natural way of parallelizing the loop over elements is selected.

As it was mentioned above we decided to test two cases for our implementation – small Laplace and big Convection-Diffusion problem. Moreover, we have tested this problems with the use of double precision and single precision variables to check the differences between DP and SP hardware units. For our tests on graphic cards we have tried two versions of kernels – one with the stiffness matrix and load vector stored in registers and the second with the matrices stored in shared memory. In this article we will reference to them by using acronyms REG_ONE_EL for register and SHM_ONE_EL for shared memory version. Both versions has their own advantages and disadvantages. The first one allows for using very fast registers, and it saves local (shared) memory for other data, but with the limited number of registers available it can easily cause register spilling and therefore lose the efficiency of the algorithm. The second version allows for saving fast registers, but it uses a slower shared memory. Our ONE_EL versions assume that the whole element is computed by the one work-group, although one work-group can (and should) of course compute more than one element. At a first stage, host code has to compute all necessary sizes of data and thus, all needed divisions of the loops. For our reference platform we use a system equipped with NVIDIA Tesla K20m GPU, whose parameters are presented in Table III. The main difference that we must assume during the transformation of the GPU algorithm for the Xeon Phi implementation, is the size of warp/wavefront. This size (equals 32 for NVIDIA or 64 for AMD) indicates the minimal size of work-group that should be used on a given device. Due to the hardware division of every compute unit of Tesla GPU, we must also provide proper (high enough) ratio of compute unit occupancy. According to [23], Intel Xeon Phi fully utilizes its vector registers when the work-group size is set to 16. This allows for the most optimal automatic vectorization that can fully use the advantages of a very wide vector registers to store variables and use vector computations on the hardware units. Other difference lay in the use of the shared memory, because all OpenCL memory levels are mapped into Xeon Phi global memory. Hence, the use of shared and constant memory should be minimized and all possible data should be declared locally to allow proper vectorization. Of course, in the case of such a complicated algorithm there is no possibility to fit all data in registers, so we must find a proper way of preparing and storing the data. For these reasons, in opposite to GPUs SHM and REG versions that assumes only stiffness matrix allocation, we have considered more complex options for Xeon Phi.

For our tests we use a computational domain with 782336 prismatic elements. Because of the minimal work-group size that should be used for a certain architecture this indicates that we have to compute data of 785408 elements on Xeon Phi and 798720 on GPU, which in this second case is 16384 elements more than our computational domain size. While this amount seems to be very large, in fact it is only 2% more calculations and it is absolutely necessary for the proper mapping to the hardware. Due to the fact that one work-group has to compute 64 elements at once, we must divide the number of elements per compute unit by this size, so we will receive 832 work-groups that will work on 960 elements. For our Xeon Phi accelerator we have accordingly 236 work-groups with 208 elements to compute. Therefore, for GPU we have a total number of 53248 threads, while for Xeon Phi there are only 3776 threads. All precomputed values needed for calculations are shown in Table IV.

After calculations of all necessary divisions, the space needed for calculation is computed, and the data preparation phase begins. At this stage all needed buffers on the kernel side are prepared and the necessary data are computed. For our algorithm we need the following data:

- execution parameters – all values earlier computed on the host side that may be necessary for our computations – e.g. the number of elements per kernel and per work-group. This data can be stored in constant memory because we do

not need to change it. For Xeon Phi case where constatnt memory is a part of global we can assume direct read from the global memory.

TABLE IV.
PARAMETERS FOR NUMERICAL INTEGRATION OF PRISMATIC ELEMENT

	Xeon Phi	Tesla K20m
number of elements to compute	782336	782336
number of elements for kernel	785408	798720
compute units	236	13
number of elements per CU	3328	61440
number of elements per wg	208	960
wg size	16	64
number of work groups	236	832

- Gauss points data – all necessary Gaussian integration points data – their coordinates and associated weights can also be stored in constant memory or read from global in Xeon Phi case.
- Values of the shape functions and their derivatives on a reference element – needed for all Jacobian calculations and obtained in the same way as previous data.
- Geometric data (coordinates) for all elements – stored in global memory of the device. Here we can assume several different cases – we can copy it to local memory for each element separately (main method for REG version), copy it in coalesced way for all elements in work group (main method for SHM version) or use it directly form global memory (Xeon Phi).
- Problem dependent coefficients – send to global memory for all elements. Here we can repeat the methods from the geometric data above but for Xeon Phi we also decided to copy it directly to the registers to speed up the calculations.

After preparing the data above we can start our computations. Firstly if we are using shared and constant memory we must read all necessary execution parameters, Gauss data and values of the reference shape functions. At this stage for SHM version we have to declare local arrays for stiffness matrix and load vector. After preparation we are entering the outer loop over elements processed by a thread. According to the Table IV for Xeon Phi it is 208 elements per work-group of size 16 which indicates that each thread has to compute 13 elements, while for Tesla it will be 960 elements per work-group of size 64, that results in 15 elements per single iteration. Inside this loop we are reading all geometrical and coefficient data for one element. As it was mentioned above for SHM version we can organize this data for so-called coalescent access which theoretically enable higher performance of data transfer allowing for simultaneously read all data by all threads within one work-group. For Xeon Phi we can use global memory directly. Because of the use of the local memory on GPU after reading this data we need to establish a synchronization point with the use of a barrier, which can slow the flow of calculations a little bit in opposite to Xeon Phi and its direct global memory access. The next step includes defining (for REG and PHI versions) and zero the local stiffness matrix and load vector. Afterwards, we are entering the loop over Gauss points where we have to compute the Jacobian transformation matrix and its inverse on the basis of the previously obtained Gauss and geometrical data. After this calculations we are entering the innermost loops over the shape functions. After computing the values of shape functions and their derivatives for a real element based on their values for the reference element and earlier computed Jacobian matrix, we can compute a final entry to the stiffness matrix and load vector according to the algorithm 1. For SHM version we need to compute the right offset for storing the computed matrix in local memory. After computations for each Gauss points we can send the data to the device global memory. After all computations, the data stored are read back to the host system memory where they can be checked and used for further FEM computations. The amount of data send to and received from device global memory is shown in Table V.

TABLE V.
AMOUNT OF DATA SEND FOR NUMERICAL INTEGRATION

Device	Problem	Variable types	In data size [MB]	Out data size [MB]
Xeon Phi	**Laplace**	**double**	169,65	263,89
		float	84,82	131,95
	Conv-diff	**double**	238,76	263,89
		float	119,38	131,95
Tesla K20m	**Laplace**	**double**	172,52	268,37
		float	86,26	134,19
	Conv-diff	**double**	242,81	268,37
		float	121,41	134,19

VI. TESTS RESULTS

For the best comparison we use the same SHM and REG algorithms for our tests on Xeon Phi. Moreover, basing on our experiments and the [23] we have prepared the more optimal version with the direct global memory use and maximization of the register usage which we refer as PHI. The performance results obtained are presented in tables VI and VII. For simplifying the comparison between our

Xeon Phi card and a reference Nvidia Tesla K20m we provide corresponding figures.

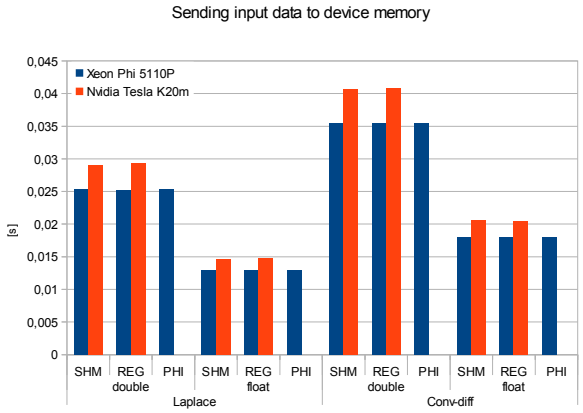


Fig 5. Sending input data to device memory

As we see from Fig. 5 the time for sending the data of comparable sizes are almost the same for Xeon Phi and Nvidia Tesla, but in all cases Xeon seems to be slightly better than Tesla.

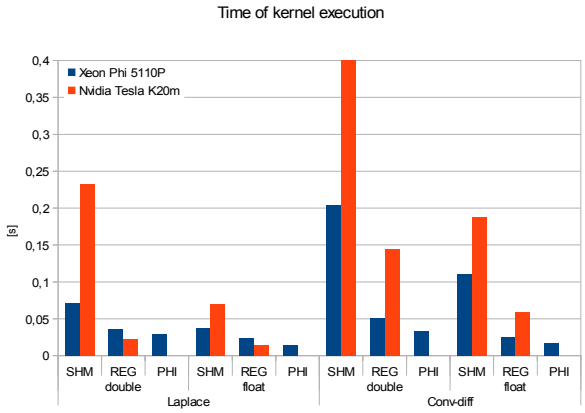


Fig 6. Time of kernel execution

Time of kernel execution (Fig. 6) shows more differences depending on the tested problem and used algorithm. A version with the use of shared memory turns out to be non optimal in our cases, but Xeon Phi is a much more faster than the Tesla card. What is more interesting we can see that the REG version with stiffness matrix stored in registers is slightly more faster than the PHI version for a small Laplace problem. All this advantage is lost when we use more complicated Convection-diffusion problem. This may indicate that Xeon Phi need quite big amount of data to fully utilize its vector registers and take advantage of it.

TABLE VI.
TEST RESULTS FOR INTEL XEON PHI

Problem	Variable types	Kernel version	Sending Input Data to device memory		Executing kernel [s]	Copying Output Data from device memory	
			[s]	[GB/s]		[s]	[GB/s]
Laplace	double	SHM	0,02531	6,70280	0,07172	0,85134	0,30998
		REG	0,02526	6,71710	0,03618	0,84913	0,31079
		PHI	0,02540	6,67933	0,02967	0,84930	0,31073
	float	SHM	0,01291	6,56829	0,03724	0,42433	0,31096
		REG	0,01293	6,56175	0,02363	0,42609	0,30967
		PHI	0,01286	6,59605	0,01492	0,42536	0,31021
Conv-diff	double	SHM	0,03538	6,74895	0,20397	0,85181	0,30981
		REG	0,03545	6,73447	0,05066	0,85110	0,31007
		PHI	0,03544	6,73691	0,03256	0,85159	0,30989
	float	SHM	0,01800	6,63413	0,11021	0,42630	0,30952
		REG	0,01795	6,64973	0,02525	0,43486	0,30343
		PHI	0,01791	6,66460	0,01755	0,42597	0,30976

TABLE VII.
TEST RESULTS FOR TESLA K20M

Problem	Variable types	Kernel version	Sending Input Data to device memory		Executing kernel [s]	Copying Output Data from device memory	
			[s]	[GB/s]		[s]	[GB/s]
Laplace	double	SHM	0,029054	5,938045	0,232612	0,094975	2,82569
		REG	0,029334	5,881336	0,022464	0,164163	1,634778
	float	SHM	0,014627	5,897442	0,069616	0,142744	0,94004
		REG	0,014691	5,871696	0,013967	0,047422	2,829597
Conv-diff	double	SHM	0,040705	5,965142	0,80109	0,509872	0,526347
		REG	0,040874	5,940472	0,144406	0,094887	2,82831
	float	SHM	0,020599	5,893718	0,188041	0,046696	2,873589
		REG	0,020489	5,925403	0,058719	0,04689	2,861695

Unfortunately, all this gained performance is lost during the copying the output data back from the accelerator to the host memory (Fig. 7). As we see on Xeon Phi the organization of the global memory has no impact on the obtained results, in opposite to the Tesla card.

Table VIII shows the obtained results in Gflops – basing on that we can see that our algorithm reaches almost 15% of theoretical peak for both double and single precision according to [24]. This can lead us to the conclusion that there is a certain margin of performance that can be used for further optimization.

TABLE VIII.
PERFORMANCE ON XEON PHI

Problem	Variable types	Kernel version	Performance [GFLOPS]
Laplace	double	SHM	31,92
		REG	63,28
		PHI	91,09
	float	SHM	61,48
		REG	96,89
		PHI	155,47
Conv-diff	double	SHM	18,51
		REG	74,53
		PHI	149,75
	float	SHM	34,26
		REG	149,52
		PHI	257,06

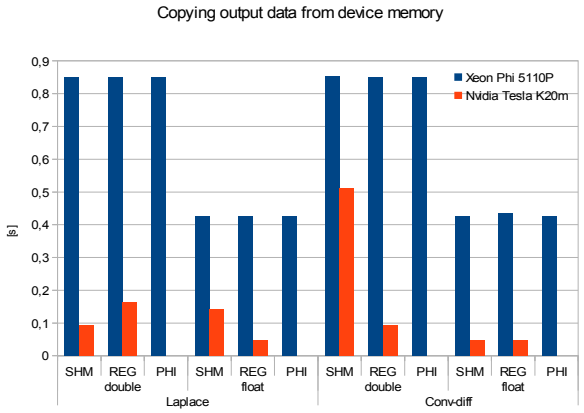


Fig 7. Copying output data from device memory

VII. CONCLUSIONS AND FUTURE WORK

As we have shown in this article Intel Xeon Phi could be an efficient and easy to use hardware for the finite element calculations. However, it needs quite big changes in actually developed GPU codes. In our further work we will try to manually vectorize the calculations and change the data retrieving algorithm to be more efficient. The first method will allow for comparing the automatic vectorization option of the compiler and check if it fully utilizes very wide 512-bit vector registers. Second method will allow to catch up with the Tesla GPU speed of data transfer and will let to make a full comparison of the competitive architectures.

REFERENCES

[1] NVIDIA, *CUDA C Programming Guide*, version 6.0, 2014.

[2] AMD, *AMD Accelerated Parallel Processing. OpenCL Programming Guide*, revision 2.7, 2013.

[3] Intel, *Intel 64 and IA-32 Architectures Optimization Reference Manual*, April 2012.

[4] IBM, *Cell Broadband Engine Programming Handbook Including the PowerXCell 8i Processor*, version 1.11, May 2008.

[5] F. Krużel, and K. Banaś, "Vectorized OpenCL implementation of numerical integration for higher order finite elements," Computers & Mathematics with Applications, vol. 66 (10), pp. 2030-2044, 2013, http://dx.doi.org/10.1016/j.camwa.2013.08.026

[6] K. Banaś, P. Płaszewski, and P. Macioł, "Numerical integration on GPUs for higher order finite elements", Computers & Mathematics with Applications, vol. 67 (6), pp. 1319-1344, 2014, http://dx.doi.org/10.1016/j.camwa.2014.01.021

[7] K. Banaś, and F. Krużel, "Large scale numerical integration on GPU", submitted for publication.

[8] N. K. Govindaraju, S. Larsen, J. Gray, and D. Manocha, "A memory model for scientific algorithms on graphics processors," SC 2006 Conference, Proceedings of the ACM/IEEE, Nov. 2006, http://dx.doi.org/10.1109/SC.2006.2

[9] K. Rojek, and L. Szustak, "Adaptation of double-precision matrix multiplication to the Cell Broadband Engine architecture," in: PPAM'09: Proceedings of the 8th international conference on Parallel processing and applied mathematics, Springer-Verlag, Berlin, Heidelberg, pp. 535-546, 2010.

[10] K. J. Barker, K. Davis, A. Hoisie, D. K. Kerbyson, M. Lang, S. Pakin, and J. C. Sancho, "Entering the petaflop era: The architecture and performance of Roadrunner," High Performance Computing, Networking, Storage and Analysis, pp. 1-11, Nov. 2008, http://dx.doi.org/10.1109/SC.2008.5217926

[11] L. Seiler, D. Carmean, E. Sprangle, T. Forsyth, M. Abrash, P. Dubey, et al., "Larrabee: a many-core x86 architecture for visual computing", in SIGGRAPH '08: ACM SIGGRAPH 2008 papers, pp. 1-15, 2008, http://dx.doi.org/10.1145/1399504.1360617

[12] R. Goodwins, "Intel unveils many-core Knights platform for HPC", www.zdnet.co.uk, 2010.

[13] Intel, *Intel Xeon Phi Coprocessor Datasheet*, June 2013.

[14] F. Roth, *System Administration for the Intel Xeon Phi Coprocessor*, 2013.

[15] T. P. Morgan, Intel teaches Xeon Phi x86 coprocessor snappy new tricks, www.theregister.co.uk, 2012.

[16] Khronos OpenCLWorking Group, The OpenCL Specification, Ed. A. Munshi, version 1.2, revision 19, 2012.

[17] N. Wilt, *The CUDA Handbook: A Comprehensive Guide to GPU Programming*, Addison-Wesley Professional, 2013.

[18] B. Gaster, D. Kaeli, L. Howes, P. Mistry, and D. Schaa, *Heterogeneous Computing With OpenCL*, Elsevier Science & Technology, 2011.

[19] S. Rul, H. Vandierendonck, J. D' Haene, and K. De Bosschere, "An experimental study on performance portability of OpenCL kernels", in: Application Accelerators in High Performance Computing, 2010 Symposium, Knoxville, TN, USA, p. 3, 2010.

[20] NVIDIA, "NVIDIA'a Next Generation CUDA Compute Architecture: Kepler GK110. The Fastest, Most Efficient HPC Architecture Ever Built", Whitepaper, ver. 1.0, 2012.

[21] NVIDIA, "Tesla K-Series Datasheet", Oct. 2013.

[22] K. Michalik, K. Banaś, P. Płaszewski, and P. Cybułka, "ModFem : a computational framework for parallel adaptive finite element simulations", Computer Methods in Materials Science, vol 13 (1), pp 3-8, 2013.

[23] Intel, *Intel SDK for OpenCL Applications XE 2013 R2 Optimization Guide,* 2013.

[24] Intel, *Intel Xeon Phi Product Family Performance,* revision 1.4, 12th December 2013.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 613–622
DOI: 10.15439/2014F374
ACSIS, Vol. 2

Performance analysis of a scalable algorithm for 3D linear transforms

Ivan Lirkov
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
Acad. G. Bonchev, bl. 25A
1113 Sofia, Bulgaria
ivan@parallel.bas.bg
http://parallel.bas.bg/~ivan/

Marcin Paprzycki
Maria Ganzha
Systems Research Institute, Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
paprzyck@ibspan.waw.pl, maria.ganzha@ibspan.waw.pl
http://www.ibspan.waw.pl/~paprzyck/
http://inf.ug.edu.pl/~mganzha/

Stanislav Sedukhin
Graduate School of Computer Science and Engineering
The University of Aizu
Tsuruga, Ikki-Machi, Aizu-Wakamatsu City
Fukushima, 965-8580 Japan
sedukhin@u-aizu.ac.jp

Paweł Gepner
Intel Corporation
Pipers Way
Swindon Wiltshire SN3 1RJ
United Kingdom
pawel.gepner@intel.com

***Abstract*—Practical realizations of 3D forward/inverse separable discrete transforms, such as Fourier transform, cosine/sine transform, etc. are frequently the principal limiters that prevent many practical applications from scaling to a large number of processors. Specifically, existing approaches, which are based primarily on 1D or 2D data decompositions, prevent the 3D transforms from effectively scaling to the maximum (possible / available) number of computer nodes. Recently, a novel, highly scalable, approach to realize forward/inverse 3D transforms has been proposed. It is based on a 3D decomposition of data and geared towards a torus network of computer nodes. The proposed algorithms requires compute-and-roll time-steps, where each step consists of an execution of multiple GEMM operations and concurrent movement of cubical data blocks between nearest-neighbor nodes (directly using the logical arrangements of the nodes within the torus). The proposed 3D orbital algorithms gracefully avoids the, required, 3D data transposition. The aim of this paper is to present a preliminary experimental performance study of the proposed implementation on two different high-performance computer architectures.**

I. Introduction

THREE-DIMENSIONAL (3D) discrete transforms (DT) such as Fourier transform, cosine/sine transform, Hartley transform, Walsh-Hadamard transform, etc., are known to play a fundamental role in many application areas, such as spectral analysis, digital filtering, signal and image processing, data compression, medical diagnostics, etc. Continuously increasing demands for high speed computing, in a constantly increasing number of many real-world applications, have stimulated the development of a number of "fast algorithms," such as the Fast Fourier Transform (FFT), characterized by dramatic reduction of arithmetic complexity. However, further reduction of execution (wall-clock) time is possible only by overlapping these arithmetic operations, i.e. using parallel implementation.

There exists three different approaches to parallel implementation of the 3D forward/inverse discrete transforms. Two of them are particularly well suited for the Fourier transform.

The first one is the 1D or "slab" decomposition of the initial 3D data. In this approach, $N \times N \times N$ data is divided into 2D slabs of size $N \times N \times b$, where $b = N/P$ and P is the number of computer nodes. The scalability of the slab-based approach, or the maximum number of nodes that can be used concurrently, is limited by the number of data elements along a single dimension of the 3D transform.

The second approach is the 2D or "pencil" decomposition, of a 3D $N \times N \times N$ initial data, among a 2D array of $P \times P$ computer nodes. Here, the initial cube is divided into a 1D "pencil" of size $N \times b \times b$, and is assigned to each node (as above, $b = N/P$). This approach increases the maximum number of nodes than can be effectively used in computations, from N to N^2. Parallel 3D FFT implementation with a 2D data decomposition has been discussed, among others, in [1], [3], [4].

In both of these, so-called, "transposed" approaches, the computational part and the inter-node communication part are separated. Moreover, a computational part inside each node is implemented by using either 2D or 1D fast (recursive) algorithm for a "slab"-based or a "pencil"-based decomposition, respectively, *without* any inter-node communication. However, upon completion of each computational part, in order to support contiguity of memory accesses, a transposition of the 3D data array is required, to put data of appropriate dimension(s) into each node. Here, at least one or two transpositions would be needed for the 1D or 2D data decomposition-based approaches, respectively. Each of such transpositions of

978-83-60810-58-3/$25.00

3D data is typically implemented by a global "all-to-all" inter-node, message-passing communication.

The last approach is the 3D or "cube" decomposition, which was recently proposed in [5]. The 3D or "cubic" decomposition of an $N \times N \times N$ initial data among $P \times P \times P$ computer nodes, allows a 3D data "cube" of size $b \times b \times b$ to be assigned to each computer node. It is easy to realize that, here, the theoretical scalability is further improved from N^2 to N^3. In this approach, blocked GEMM-based algorithms are used to compute the basic one-dimensional N-size transform, not on a single but on the $P = N/b$ cyclically interconnected (for data reuse) nodes of a 3D torus network. In this way, the proposed algorithm *integrates* local intra-node computation with a nearest-neighbour inter-node communication, at each step of the three-dimensional processing. It is important to observe that the proposed algorithm, with its 3D data decomposition, and the torus-oriented communication scheme, *completely eliminates global communication*. In addition, computation and local communication can be overlapped. Finally, note that in the considered approach, the 3D transform is represented as three chained sets of cubical tensor-by-matrix or matrix-by-tensor multiplications, which are executed in a 3D torus network of computer nodes by the fastest and extremely scalable orbital algorithms.

Te main contribution of this paper is to experimentally evaluate the performance of the latter algorithm. To do this, we have implemented overlapping of computation and communication for the 3D data decomposition and used GEMM kernels available on selected computers. The experimental performance of the 3D Discrete Cosine Transform (DCT) and Discrete Fourier Transform (DFT), with the 3D data decomposition, has been evaluated on a Linux cluster and on the Blue Gene/P supercomputer.

II. 3D Separable Transform

Let us start by introducing basic definitions concerning 3D separable transforms. Let $X = [x(n_1, n_2, n_3)]$, $0 \le n_1, n_2, n_3 < N$, be an $N \times N \times N$ cubical grid of input data, or a three-way data tensor. A separable forward 3D transform of X is another cubical grid of an $N \times N \times N$ data or a three-way tensor $\dddot{X} = [\dddot{x}(k_1, k_2, k_3)]$, where for all $0 \le k_1, k_2, k_3 < N$:

$$\dddot{x}(k_1, k_2, k_3) = \sum_{n_3=0}^{N-1} \sum_{n_2=0}^{N-1} \sum_{n_1=0}^{N-1} x(n_1, n_2, n_3) \cdot c(n_1, k_1) \cdot c(n_2, k_2) \cdot c(n_3, k_3) \quad (1)$$

A separable inverse, or backward, 3D transform of a three-way tensor $\dddot{X} = [\dddot{x}(k_1, k_2, k_3)]$ is expressed as:

$$x(n_1, n_2, n_3) = \sum_{k_3=0}^{N-1} \sum_{k_2=0}^{N-1} \sum_{k_1=0}^{N-1} \dddot{x}(k_1, k_2, k_3) \cdot c(n_1, k_1) \cdot c(n_2, k_2) \cdot c(n_3, k_3) \quad (2)$$

where $0 \le n_1, n_2, n_3 < N$ and $X = [x(n_1, n_2, n_3)]$ is an output $N \times N \times N$ cubical tensor.

We will use the notations from [5] to describe the proposed parallel algorithm. First, we divide the input data $X = [x(n_1, n_2, n_3)]$ into $P_1 \times P_2 \times P_3$ data rectangular cuboid, where each cuboid $X(N_1, N_2, N_3), 0 \le N_i < P_i$, has the size of $b_1 \times b_2 \times b_3$, i.e. $b_i = N/P_i$. Then, the forward 3D transform can be expressed as a block version of the multi-linear matrix multiplication:

$$\dddot{X}(K_1, K_2, K_3) = \sum_{N_3=0}^{P_3-1} \sum_{N_2=0}^{P_2-1} \sum_{N_1=0}^{P_1-1} X(N_1, N_2, N_3) \times C(N_1, K_1) \times C(N_2, K_2) \times C(N_3, K_3), \quad (3)$$

where $0 \le K_i < P_i$ and $C(N_s, K_s), s = 1, 2, 3$, is the (N_s, K_s)-th block of the transform matrix C.

Due to the separability of the linear transforms, the 3D transform can be split into three data dependent sets of 1D transforms. At the *first stage*, the 1D transform of $X(N_1, N_2, :)$ is performed for all (N_1, N_2) pairs, as a block tensor-by-matrix multiplication:

$$\dot{X}(N_1, N_2, K_3) = \sum_{N_3=0}^{P_3-1} X(N_1, N_2, N_3) \times C(N_3, K_3).$$

At the *second stage*, the 1D transform of $\dot{X}(:, N_2, K_3)$ is implemented for all (N_2, K_3) pairs, as the second block tensor-by-matrix multiplication:

$$\ddot{X}(K_1, N_2, K_3) = \sum_{N_1=0}^{P_1-1} \dot{X}(N_1, N_2, K_3) \times C(N_1, K_1).$$

At the *third stage*, the 1D transform of $\ddot{X}(K_1, :, K_3)$ is implemented for all (K_1, K_3) pairs, as the third block tensor-by-matrix multiplication:

$$\dddot{X}(K_1, K_2, K_3) = \sum_{N_2=0}^{P_2-1} \ddot{X}(K_1, N_2, K_3) \times C(N_2, K_2).$$

By slicing the cubical data, i.e. representing the three-way tensors as the set of matrices, it is possible to formulate the 3D transform as a conventional block *matrix-by-matrix multiplication* with its transpose/nontranspose versions. In this case, the initial data grid $X(N_1, N_2, N_3)$, is divided into 1D "slices" along one axis. Then, the 3D transform can also be computed in three data-dependent stages as chaining sets of block matrix-by-matrix products.

III. Algorithm Description

A. Multi-node Implementation

In the proposed approach it is assumed that each computer node CN(Q,R,S) has six bi-directional links labeled as $\pm Q$, $\pm R$ and $\pm S$. These nodes are toroidally interconnected. During processing, some blocks of tensor data are rolled, i.e. cyclically shifted, along (+) or opposite (-) axis (orbit). The first two stages implement the set of space-independent 2D forward transforms, in parallel, along the R-axis (orbit) slabs.

Note that, each stage of both forward and inverse transforms, with 3D data decomposition, has a common structure, i.e. steps of "compute-and-roll".

A three-stage orbital implementation of the 3D forward transform in a 3-dimensional network of toroidally interconnected nodes CN(Q,R,S) proceeds as follows.

Stage I.
$\dot{X}(N_1, N_2, K_3) = \sum_{0 \le N_3 < P_3} X(N_1, N_2, N_3) \times C(N_3, K_3)$:

- **for all** CN(Q, R, S) **do** P_3 times:
  1) **compute**: $\dot{X} \leftarrow X \times C + \dot{X}$
  2) **data roll**: $\xleftarrow{+S}$ X $\xleftarrow{-S}$

Stage II. $\ddot{X}(K_1, N_2, K_3) =$
$\sum_{0 \le N_1 < P_1} C(N_1, K_1)^T \times \dot{X}(N_1, N_2, K_3)$:

- **for all** CN(Q, R, S) **do** P_1 times:
  1) **compute**: $\ddot{X} \leftarrow C^T \times \dot{X} + \ddot{X}$
  2) **data roll**: $\xleftarrow{+Q}$ $\dot{X}$ $\xleftarrow{-Q}$

Stage III.
$\dddot{X}(K_1, K_2, K_3) = \sum_{0 \le N_2 < P_2} \ddot{X}(K_1, N_2, K_3) \times C(N_2, K_2)$:

- **for all** CN(Q, R, S) **do** P_2 times:
  1) **compute**: $\dddot{X} \leftarrow \ddot{X} \times C + \dddot{X}$
  2) **data roll**: $\xleftarrow{+R}$ $\ddot{X}$ $\xleftarrow{-R}$

For more details, see [5].

It should be noted that the implementation described here is a modification of the parallel algorithms proposed in [5]. The main differences between our implementation and the original algorithm are:

1) The implemented parallel algorithm works only for the 3D DCT and the 3D DFT;
2) The proposed implementation uses additional arrays to store elements of the coefficient matrix C. In the case of the DCT, we use one array with $4N$ elements; while for the DFT two arrays with N elements each. In this way, we avoid rolling the coefficient matrix. In other words, we simplify the communication, while paying the price of somewhat increasing (by $O(N)$ elements) the total memory utilization.

Since the tensor-by-matrix, or the matrix-by-tensor, multiplications can be expressed as the set of matrix-by-matrix multiplications, we can use an existing GEMM subroutines, from the BLAS library [2], to compute the 3D transform.

B. Multi-thread Implementation

There exists two possible ways to compute the tensor-by-matrix multiplication on computers with multi-core processors. The first one is to use the multi-threaded library, such as the Engineering and Scientific Subroutine Library (ESSL, see http://www-03.ibm.com/systems/software/essl/index.html) or the Intel Math Kernel Library (MKL, see http://software.intel.com/en-us/articles/intel-mkl/). Here, each slice of the tensor is computed by multiple threads. The other possible approach is to use OpenMP. In the current implementation, we have linked our code to the multi-threaded library for the parallelization on a single (multi-core) node of the computer system.

IV. Experimental Results

A portable parallel code was designed and implemented in C. The parallelization was based on the MPI standard [6], [7]. In the code, we used the BLAS subroutines SGEMM, DGEMM, CGEMM, and ZGEMM to perform matrix-by-matrix multiplication. In order to obtain a better mapping of the processors to the physical interconnect topology of computers actually used in experiments, functions MPI_Dims_create and MPI_Cart_create were used to create a logical 3D Cartesian grid of processors. Let us also note that we used one MPI process per computer node.

The parallel code has been tested on the following systems: (1) a cluster computer *Galera*, located in the Polish Informatics Center TASK, and (2) two IBM Blue Gene/P machines, one at the Bulgarian Supercomputing Center, and one at the HPC Center of the West University of Timisoara (UVT).

In our experiments, times have been collected using the MPI provided timer, and we report the best results from multiple runs. In the following tables, we report the elapsed (wall-clock) time T_p, in seconds, using p MPI processes, and the parallel speed-up $S_p = T_1/T_p$.

Tables I and II show the results collected on the Galera. It is a Linux cluster with 336 nodes, and two Intel Xeon quad core processors per node. Each processor runs at 2.33 GHz. Processors within each node share 8, 16, or 32 GB of memory. Nodes are interconnected with a high-speed InfiniBand network (see also http://www.task.gda.pl/kdm/sprzet/Galera). When running our code on Galera, we used the Intel C compiler, and compiled the code with the options "-O3 -openmp". To use the BLAS subroutines, we linked our code to the optimized multi-threaded Intel MKL library.

The symbol * in the tables denotes that, in the given case, the memory of p nodes was not large enough to compute the 3D transform for data of size $N \times N \times N$.

The reported execution time for $N = 100$ shows that the problem is "small" and can be executed on one node of the cluster (no need for parallelization). Here, there is no significant improvement from using two or more nodes. However, already for the problems of size $N = 600$ a significant performance gain can be observed (see, also, below). Considering the fact that some of the applications that need 3D transforms involve "real-time processing of data," it is worthy noting that, using the proposed method, similar time is required to find the solution on a single node for the problem of size $N = 600$ as finding solution using 256 nodes for the problem of size $2000 < N < 2400$.

Table III contains the speed-up obtained on the Galera. For the largest problem, which can be executed on a single node, the parallel efficiency is above 50% for the number of nodes up to 16 for the DCT and up to 32 for the DFT. We note that the main advantage of the parallel algorithm is that the code allows performing the 3D transform for very large data. Taking into account the largest cases reported in Tables I and II, we

TABLE I
EXECUTION TIME FOR THE 3D DISCRETE COSINE TRANSFORM ON GALERA.

N	nodes								
	1	2	4	8	16	32	64	128	256
	single precision								
	forward transform								
100	0.08	0.06	0.06	0.18	0.07	0.14	0.20	0.19	0.20
200	0.23	0.18	0.15	0.10	0.09	0.10	0.11	0.14	0.18
300	0.86	0.54	0.31	0.21	0.14	0.15	0.17	0.12	0.21
400	2.18	1.26	0.71	0.41	0.32	0.25	0.23	0.16	0.22
600	9.59	5.58	3.06	1.80	0.98	0.68	0.45	0.36	0.35
800	*	14.51	7.70	4.29	2.51	1.43	0.85	0.72	0.52
1000	*	*	17.91	10.12	5.47	3.07	1.77	1.28	0.99
1200	*	*	*	18.88	10.71	5.80	3.31	2.48	1.56
1400	*	*	*	34.11	19.02	10.23	5.67	3.97	2.76
1600	*	*	*	*	27.45	14.83	8.11	5.11	3.74
2000	*	*	*	*	*	35.00	18.50	11.16	7.48
2400	*	*	*	*	*	75.50	35.10	22.12	13.56
2800	*	*	*	*	*	*	63.67	36.11	23.61
3200	*	*	*	*	*	*	*	53.51	34.92
	backward transform								
100	0.03	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02
200	0.20	0.13	0.08	0.05	0.03	0.02	0.01	0.01	0.03
300	0.83	0.49	0.29	0.17	0.11	0.08	0.05	0.06	0.04
400	2.08	1.16	0.65	0.38	0.25	0.18	0.09	0.09	0.07
600	9.38	5.44	2.96	1.61	0.98	0.56	0.35	0.29	0.20
800	*	14.00	7.36	4.13	2.36	1.24	0.72	0.58	0.43
1000	*	*	17.65	9.56	5.14	3.00	1.75	1.33	0.88
1200	*	*	*	18.04	10.73	5.82	3.19	2.10	1.58
1400	*	*	*	33.53	18.49	9.85	5.48	3.45	2.46
1600	*	*	*	*	26.68	14.36	7.68	5.57	3.56
2000	*	*	*	*	*	34.52	18.50	12.05	7.77
2400	*	*	*	*	*	70.40	37.74	21.43	13.99
2800	*	*	*	*	*	*	65.15	38.27	23.74
3200	*	*	*	*	*	*	*	58.55	33.11
	double precision								
	forward transform								
100	0.07	0.05	0.07	0.14	0.08	0.15	0.19	0.20	0.48
200	0.39	0.27	0.20	0.16	0.15	0.12	0.16	0.16	0.20
300	1.51	0.89	0.51	0.37	0.31	0.21	0.19	0.22	0.20
400	4.09	2.26	1.27	0.78	0.49	0.34	0.25	0.27	0.24
600	18.49	9.63	5.29	2.90	1.77	1.04	0.62	0.60	0.44
800	*	*	14.72	8.00	4.36	2.56	1.41	1.15	0.92
1000	*	*	*	18.11	9.69	5.40	3.12	2.36	1.72
1200	*	*	*	36.27	18.10	10.00	5.68	4.11	2.95
1400	*	*	*	*	32.85	18.05	9.78	6.59	4.55
1600	*	*	*	*	*	30.76	15.28	9.17	6.89
2000	*	*	*	*	*	*	*	20.80	14.69
2400	*	*	*	*	*	*	*	39.98	26.38
2800	*	*	*	*	*	*	*	68.79	47.75
	backward transform								
100	0.04	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.02
200	0.36	0.21	0.13	0.07	0.05	0.04	0.04	0.02	0.03
300	1.43	0.82	0.48	0.32	0.17	0.13	0.08	0.06	0.07
400	3.94	2.09	1.24	0.75	0.41	0.28	0.17	0.14	0.13
600	18.10	9.06	4.88	2.74	1.60	1.02	0.57	0.45	0.28
800	*	*	13.87	7.39	4.13	2.30	1.37	1.14	0.71
1000	*	*	*	16.92	9.17	5.02	2.90	2.54	1.68
1200	*	*	*	34.43	17.00	9.43	4.78	3.88	2.76
1400	*	*	*	*	31.85	16.97	8.90	7.09	4.76
1600	*	*	*	*	*	27.07	14.07	10.75	7.14
2000	*	*	*	*	*	*	*	22.22	15.28
2400	*	*	*	*	*	*	*	42.08	25.26
2800	*	*	*	*	*	*	*	71.30	44.40

TABLE II
EXECUTION TIME FOR THE 3D DISCRETE FOURIER TRANSFORM ON GALERA.

N	nodes								
	1	2	4	8	16	32	64	128	256
	single precision								
	forward transform								
100	0.13	0.07	0.15	0.14	0.12	0.14	0.21	0.19	0.23
200	0.61	0.39	0.26	0.24	0.12	0.19	0.20	0.21	0.22
300	2.52	1.56	0.93	0.55	0.42	0.37	0.32	0.27	0.27
400	7.08	3.93	2.16	1.24	0.75	0.54	0.36	0.36	0.36
600	32.02	17.42	9.34	5.15	3.20	1.72	1.01	0.77	0.57
800	*	*	26.21	14.13	7.99	4.40	2.46	1.49	1.13
1000	*	*	*	31.98	17.89	9.75	5.33	3.12	2.05
1200	*	*	*	63.48	34.49	18.41	9.79	6.19	3.78
1400	*	*	*	*	61.93	32.31	17.73	10.15	6.09
1600	*	*	*	*	*	51.22	27.99	15.03	8.85
2000	*	*	*	*	*	*	61.28	34.67	18.51
2400	*	*	*	*	*	*	153.02	66.21	37.28
2800	*	*	*	*	*	*	*	124.36	67.65
	backward transform								
100	0.06	0.04	0.03	0.02	0.01	0.01	0.01	0.01	0.01
200	0.58	0.33	0.20	0.11	0.08	0.09	0.06	0.04	0.03
300	2.44	1.43	0.83	0.47	0.31	0.20	0.11	0.08	0.06
400	6.93	3.80	2.04	1.13	0.64	0.44	0.26	0.22	0.17
600	31.49	16.84	8.99	4.92	2.90	1.77	0.94	0.68	0.43
800	*	*	25.66	13.67	7.44	4.39	2.32	1.67	1.00
1000	*	*	*	31.46	17.56	9.25	5.03	3.20	2.31
1200	*	*	*	64.16	33.55	17.38	10.00	5.88	3.92
1400	*	*	*	*	60.57	32.48	17.19	10.54	6.77
1600	*	*	*	*	*	51.21	26.56	15.28	9.66
2000	*	*	*	*	*	*	62.10	36.40	22.04
2400	*	*	*	*	*	*	173.82	66.95	37.35
2800	*	*	*	*	*	*	*	136.07	66.60
	double precision								
	forward transform								
100	0.17	0.14	0.09	0.17	0.06	0.20	0.19	0.20	0.29
200	1.23	0.73	0.42	0.37	0.19	0.17	0.22	0.24	0.22
300	5.24	2.96	1.57	1.02	0.66	0.47	0.34	0.32	0.33
400	14.63	8.21	4.30	2.43	1.36	0.88	0.60	0.47	0.43
600	74.84	36.03	18.82	10.47	5.95	3.24	1.88	1.30	0.98
800	*	*	*	28.70	15.78	8.34	5.05	2.96	1.92
1000	*	*	*	*	36.49	18.98	10.29	6.13	3.89
1200	*	*	*	*	71.56	36.18	19.53	11.37	6.88
1400	*	*	*	*	*	64.20	34.80	19.71	11.65
1600	*	*	*	*	*	*	58.76	29.84	18.58
2000	*	*	*	*	*	*	*	70.18	37.90
2400	*	*	*	*	*	*	*	204.81	69.83
	backward transform								
100	0.11	0.07	0.04	0.03	0.02	0.04	0.01	0.01	0.01
200	1.20	0.67	0.39	0.22	0.16	0.10	0.07	0.06	0.11
300	5.09	2.81	1.51	0.86	0.57	0.31	0.21	0.12	0.11
400	14.32	7.81	4.16	2.37	1.30	0.78	0.45	0.35	0.33
600	85.48	35.01	18.05	9.79	5.25	2.99	1.81	1.07	0.78
800	*	*	*	28.14	14.88	8.13	4.33	3.01	2.18
1000	*	*	*	*	35.58	17.76	9.97	6.14	4.40
1200	*	*	*	*	72.70	36.49	18.90	11.06	7.65
1400	*	*	*	*	*	64.13	33.69	19.83	11.94
1600	*	*	*	*	*	*	54.91	30.35	18.84
2000	*	*	*	*	*	*	*	71.07	39.38
2400	*	*	*	*	*	*	*	136.01	74.78

TABLE III
SPEED-UP ON GALERA.

N	nodes							
	2	4	8	16	32	64	128	256
	single precision DCT							
	forward transform							
100	1.26	1.28	0.46	1.20	0.58	0.40	0.43	0.41
200	1.32	1.57	2.31	2.52	2.38	2.17	1.65	1.29
300	1.61	2.74	4.16	6.03	5.75	5.17	7.13	4.08
400	1.73	3.06	5.25	6.74	8.55	9.46	14.01	9.77
600	1.72	3.14	5.34	9.82	14.10	21.26	26.72	27.47
	backward transform							
100	1.75	2.31	2.83	5.07	3.37	3.56	9.85	1.21
200	1.54	2.59	4.29	6.83	9.73	15.57	15.37	7.95
300	1.70	2.91	4.91	7.62	10.44	17.70	13.50	20.04
400	1.79	3.19	5.43	8.19	11.66	22.21	22.79	31.49
600	1.72	3.17	5.83	9.58	16.85	26.89	32.52	47.97
	double precision DCT							
	forward transform							
100	1.36	1.02	0.50	0.86	0.45	0.36	0.34	0.15
200	1.45	1.92	2.38	2.63	3.36	2.49	2.46	1.94
300	1.69	2.93	4.08	4.93	7.21	7.77	6.95	7.60
400	1.81	3.23	5.24	8.41	11.98	16.25	15.15	17.22
600	1.92	3.50	6.38	10.48	17.83	29.61	32.63	41.63
	backward transform							
100	1.60	2.30	3.76	5.36	6.55	4.05	4.37	1.66
200	1.73	2.79	4.79	7.18	9.10	8.76	19.52	13.77
300	1.74	3.00	4.46	8.24	10.64	18.67	23.22	20.88
400	1.89	3.18	5.24	9.60	14.23	22.51	27.85	30.30
600	2.00	3.71	6.60	11.31	17.73	31.83	39.98	64.69
	single precision DFT							
	forward transform							
100	1.85	0.85	0.94	1.09	0.94	0.61	0.66	0.55
200	1.56	2.34	2.57	4.91	3.20	3.08	2.89	2.73
300	1.62	2.72	4.42	5.98	6.90	7.88	9.21	9.25
400	1.80	3.28	5.72	9.41	13.06	19.86	19.51	19.50
600	1.84	3.43	6.21	10.02	18.65	31.72	41.76	56.01
	backward transform							
100	1.66	2.31	3.62	5.10	5.26	10.08	9.24	6.03
200	1.76	2.95	5.08	7.56	6.79	10.56	15.38	22.71
300	1.71	2.94	5.22	7.95	12.19	22.02	30.82	38.79
400	1.82	3.39	6.13	10.88	15.83	26.44	31.29	41.30
600	1.87	3.50	6.41	10.85	17.79	33.33	46.24	73.48
	double precision DFT							
	forward transform							
100	1.16	1.94	0.98	2.79	0.84	0.87	0.84	0.57
200	1.68	2.95	3.36	6.56	7.36	5.68	5.17	5.50
300	1.77	3.34	5.12	7.93	11.23	15.32	16.52	15.72
400	1.78	3.40	6.01	10.77	16.60	24.59	31.07	33.90
600	2.08	3.98	7.15	12.57	23.09	39.85	57.58	76.71
	backward transform							
100	1.63	2.64	3.50	5.65	2.82	9.63	9.58	13.64
200	1.79	3.06	5.32	7.59	12.19	17.53	20.07	10.50
300	1.81	3.36	5.95	8.99	16.44	24.28	42.88	44.84
400	1.83	3.44	6.03	11.02	18.38	31.52	40.92	43.24
600	2.44	4.74	8.73	16.29	28.59	47.11	79.60	110.08

can see that increasing the number of nodes from 128 to 256 results in efficiency of 60-69% for the DCT, and 40-60% for the DFT (depending if the transform forward or backward and if it runs in single or double precision).

Tables IV and V present times collected on the IBM Blue Gene/P supercomputers. For our experiments we used the BG/P machine located at the Bulgarian Supercomputing Center and a slightly different one located at the HPC Center of the West University of Timisoara (UVT). The supercomputer in Bulgaria has two BG/P racks, while the supercomputer in Romania has one BG/P rack. One BG/P rack consists of 1024 compute nodes with quad core PowerPC 450 processors (running at 850 MHz). Each node of the Bulgarian rack has 2 GB of RAM, while each node of the Romanian rack has 4 GB of RAM. For the point-to-point communications a 3.4 Gb 3D mesh network is used (for more details, see http://www.scc.acad.bg/ and http://hpc.uvt.ro/infrastructure/bluegenep/). In our experiments, to compile the code we have used the IBM XL C compiler and compiled the code with the following options: "-O5 -qstrict -qarch=450d -qtune=450 -qsmp=omp". To use the BLAS subroutines, we linked our code to the multi-threaded ESSL library.

TABLE IV
EXECUTION TIME FOR 3D DISCRETE COSINE TRANSFORM ON IBM BLUE GENE/P.

N	nodes										
	1	2	4	8	16	32	64	128	256	512	1024
	single precision										
	forward transform										
100	0.09	0.06	0.05	0.04	0.03	0.02	0.02	0.02	0.01	0.01	0.01
200	1.01	0.62	0.38	0.24	0.17	0.10	0.06	0.05	0.04	0.03	0.03
300	4.45	2.43	1.73	1.05	0.59	0.32	0.18	0.12	0.08	0.06	0.05
400	14.40	7.69	4.77	2.64	1.47	0.77	0.42	0.27	0.17	0.10	0.08
600	70.59	36.30	19.09	10.87	6.29	3.33	1.79	1.13	0.54	0.31	0.19
800	*	117.36	57.54	33.19	18.21	9.39	4.91	2.83	1.35	0.75	0.43
1000	*	*	140.94	75.20	44.04	22.93	11.94	7.15	3.46	1.88	1.09
1200	*	*	*	139.73	76.14	40.04	20.30	11.95	5.93	3.20	1.95
1400	*	*	*	*	135.50	70.74	38.77	22.47	11.29	6.03	3.23
1600	*	*	*	*	229.10	120.66	63.27	35.68	17.12	9.10	5.04
	backward transform										
100	0.09	0.05	0.04	0.03	0.02	0.01	0.01	0.01	0.01	0.01	0.01
200	1.01	0.63	0.38	0.24	0.17	0.09	0.05	0.04	0.03	0.02	0.02
300	4.48	2.48	1.83	1.06	0.60	0.32	0.17	0.11	0.07	0.05	0.03
400	14.49	7.92	4.87	2.72	1.51	0.78	0.41	0.26	0.16	0.09	0.07
600	70.91	36.59	20.04	11.26	6.44	3.34	1.77	1.14	0.55	0.31	0.18
800	*	118.42	59.93	34.29	18.77	9.44	5.00	2.89	1.35	0.76	0.44
1000	*	*	146.71	77.66	44.82	22.49	11.88	7.27	3.48	1.93	1.11
1200	*	*	*	143.08	78.76	40.51	20.29	12.18	5.94	3.22	1.98
1400	*	*	*	*	140.26	72.08	38.59	22.61	11.37	6.04	3.28
1600	*	*	*	*	236.68	120.59	63.96	36.31	17.45	9.18	5.20
	double precision										
	forward transform										
100	0.10	0.07	0.05	0.05	0.04	0.03	0.02	0.02	0.01	0.01	0.01
200	1.15	0.72	0.42	0.26	0.18	0.10	0.07	0.06	0.05	0.03	0.03
300	4.97	2.80	1.84	1.07	0.60	0.35	0.21	0.13	0.11	0.08	0.06
400	16.27	8.89	5.02	2.77	1.50	0.84	0.47	0.28	0.20	0.13	0.11
600	*	39.37	20.28	11.34	6.53	3.48	1.91	1.17	0.66	0.39	0.25
800	*	*	66.34	35.33	18.89	10.00	5.34	2.94	1.60	0.91	0.56
1000	*	*	*	85.23	47.80	25.34	13.45	7.65	4.05	2.28	1.30
1200	*	*	*	*	78.03	41.22	22.15	12.31	6.94	3.80	2.32
1400	*	*	*	*	*	85.32	44.44	25.87	13.90	7.54	3.82
1600	*	*	*	*	*	147.61	70.00	36.61	19.86	10.74	5.88
	backward transform										
100	0.10	0.06	0.05	0.03	0.02	0.01	0.01	0.01	0.01	0.01	0.01
200	1.16	0.73	0.43	0.26	0.16	0.09	0.05	0.05	0.03	0.02	0.02
300	5.01	2.83	1.87	1.09	0.61	0.33	0.19	0.12	0.10	0.06	0.04
400	16.33	9.01	5.12	2.87	1.52	0.84	0.46	0.28	0.17	0.10	0.09
600	*	39.78	20.65	11.70	6.65	3.53	1.93	1.18	0.66	0.39	0.25
800	*	*	68.21	36.05	19.00	10.20	5.50	3.05	1.60	0.93	0.57
1000	*	*	*	87.68	48.31	24.94	13.41	7.71	4.11	2.27	1.30
1200	*	*	*	*	80.46	42.39	22.17	12.32	7.12	3.86	2.35
1400	*	*	*	*	*	86.05	45.25	25.79	14.12	7.74	3.84
1600	*	*	*	*	*	135.85	70.88	37.47	20.28	10.86	6.04

Here, again the execution time for $N = 100$ shows that the code can be executed on one node and it is not necessary to use the parallel algorithm. Note that the memory of a single node of the IBM supercomputer is substantially smaller than that on the Galera cluster and is not sufficient for solving large problems. While both BG/P machines have the same processors, the one located in Romania has larger memory (with 4 GB memory per node). This is thus the machine used to run experiments with larger data sets. Due to the lack of space, and relative similarity of results, we do not report results obtained on both machines separately (when running problems of the same size). Note that individual processors on supercomputer are slower than these on the Galera cluster. For the double precision DFT the Blue Gene is approximately three times slower than the Galera.

Let us also observe that almost the same time was spent solving the problem of size $N = 600$ on a single node as it was spent when solving problem of size $N = 1600$ on 64 nodes. This indicates that the BG/P is more efficient in supporting parallel computing than the Galera cluster.

Table VI shows the speed-up obtained on the Blue Gene. Because of smaller memory per node we calculated the actual speed-up only for $N = 100, 200, 300, 400$. Furthermore, only for the single precision DCT the speed-up for $N = 600$ is reported. For $N = 400$ the parallel efficiency is more than 50% on up to 64 nodes for the DCT and on up to 512 nodes for the DFT.

An interesting observation comes from comparing results reported in Tables VI and VI, as well as those found in Tables II and V. For instance, in the most complex problem (where such

TABLE V
EXECUTION TIME FOR 3D DISCRETE FOURIER TRANSFORM ON IBM BLUE GENE/P.

N	nodes										
	1	2	4	8	16	32	64	128	256	512	
	single precision										
	forward transform										
100	0.25	0.15	0.09	0.06	0.05	0.03	0.02	0.02	0.02	0.01	0.01
200	3.65	1.94	1.05	0.58	0.33	0.18	0.11	0.07	0.06	0.04	0.03
300	17.95	9.47	5.28	2.83	1.52	0.80	0.45	0.25	0.17	0.10	0.08
400	55.63	28.70	14.60	7.69	4.06	2.13	1.13	0.62	0.34	0.20	0.14
600	*	140.97	72.54	37.82	19.51	10.20	5.27	2.97	1.58	0.89	0.49
800	*	*	223.40	113.48	57.70	29.60	15.04	7.99	4.18	2.22	1.22
1000	*	*	*	276.03	268.76	72.45	36.94	20.16	11.15	6.01	3.26
1200	*	*	*	*	287.60	146.05	74.36	38.99	20.12	10.52	5.90
1400	*	*	*	*	*	270.33	137.73	74.55	39.40	21.26	10.33
1600	*	*	*	*	*	446.59	226.00	115.32	58.73	30.08	15.87
	backward transform										
100	0.25	0.15	0.08	0.05	0.04	0.02	0.02	0.01	0.01	0.01	0.01
200	3.66	1.95	1.09	0.60	0.32	0.17	0.10	0.06	0.05	0.03	0.03
300	18.03	9.54	5.33	2.88	1.54	0.81	0.44	0.25	0.16	0.09	0.07
400	55.79	28.83	14.76	7.79	4.14	2.13	1.13	0.63	0.34	0.19	0.13
600	*	141.86	73.07	38.26	19.81	10.23	5.32	2.96	1.60	0.88	0.49
800	*	*	224.63	114.12	58.22	29.86	15.28	8.10	4.28	2.25	1.25
1000	*	*	*	279.21	234.82	72.90	37.02	20.33	11.23	6.02	3.32
1200	*	*	*	*	289.29	146.65	75.11	39.30	20.49	10.59	5.98
1400	*	*	*	*	*	271.09	139.31	74.81	39.44	21.30	10.61
1600	*	*	*	*	*	446.08	228.34	116.13	58.99	30.17	16.20
	double precision										
	forward transform										
100	0.29	0.18	0.11	0.07	0.05	0.04	0.03	0.02	0.02	0.02	0.02
200	3.98	2.19	1.17	0.67	0.37	0.21	0.13	0.08	0.06	0.04	0.04
300	19.06	10.39	5.62	3.08	1.68	0.91	0.52	0.30	0.18	0.11	0.09
400	61.59	32.04	16.59	8.77	4.53	2.43	1.35	0.73	0.40	0.24	0.15
600	*	*	79.35	41.38	21.69	11.25	6.09	3.22	1.78	1.01	0.59
800	*	*	*	125.18	63.35	33.33	17.23	9.00	4.80	2.63	1.40
1000	*	*	*	*	152.19	78.01	41.13	21.59	11.52	6.21	3.48
1200	*	*	*	*	*	157.86	81.49	42.20	22.05	11.80	6.46
1400	*	*	*	*	*	*	152.59	78.99	41.33	22.14	10.87
1600	*	*	*	*	*	*	253.92	128.75	65.41	33.98	17.91
	backward transform										
100	0.28	0.17	0.11	0.06	0.04	0.03	0.02	0.01	0.02	0.01	0.02
200	3.99	2.21	1.19	0.67	0.37	0.21	0.12	0.07	0.05	0.03	0.03
300	19.13	10.44	5.69	3.13	1.66	0.91	0.51	0.30	0.17	0.10	0.07
400	61.82	32.24	17.12	8.92	4.63	2.46	1.34	0.72	0.39	0.23	0.16
600	*	*	80.27	41.79	21.96	11.39	6.01	3.34	1.80	1.01	0.58
800	*	*	*	127.55	64.07	33.35	17.19	9.12	4.77	2.66	1.43
1000	*	*	*	*	153.40	78.62	41.46	21.56	11.57	6.19	3.47
1200	*	*	*	*	*	156.72	81.52	43.23	22.79	12.18	6.55
1400	*	*	*	*	*	*	152.42	79.23	41.22	22.18	11.00
1600	*	*	*	*	*	*	257.88	131.82	66.12	34.10	18.32

comparison was possible), for the backward double precision DFT, for $N = 400$ and 256 nodes, speedup obtained on Galera is 43, while on the BG/P it reaches 157. Furthermore, for the same problem (backward double precision DFT) the execution time on Galera on 256 nodes is 18 seconds, which is almost exactly the time needed to compute the same problem on 1024 nodes of the BG/P. Overall, this indicates that, in the case of the BG/P, somewhat slower nodes have been combined with superior network infrastructure, which is exactly the opposite than in the case of the Galera cluster (where more powerful processors are connected through a slower network).

Finally, in Figure 1, we represent execution time of the code, which performs one forward and one backward DFT. Results are presented for single and double precision, for problems of size $N = 400$ and $N = 600$. Here, it becomes even clearer that for both problems, using more than 64 nodes on the Galera cluster results in, so called, Amdahl's effect (where adding more resources does not result in a commensurate time reduction). This is not the case for the BG/P machines. Nevertheless, for up to 256 nodes, for $N = 600$, the cluster is faster in completing the task.

V. CONCLUDING REMARKS

The aim of this paper was to describe our attempt at implementing a slightly simplified version of a novel algorithm for 3D forward/inverse discrete transforms, and to report its performance on two different parallel computers. Obtained results show that the proposed approach allows solution of large 3D problems on a supercomputer as well as on a cluster. Furthermore, the initial estimates indicate quite good scalability of the proposed implementation. It should be noted

TABLE VI
SPEED-UP ON IBM BLUE GENE/P.

N	nodes									
	2	4	8	16	32	64	128	256	512	1024
	single precision DCT									
	forward transform									
100	1.57	1.97	2.51	2.76	3.95	4.56	5.13	6.84	16.40	15.43
200	1.62	2.67	4.29	5.89	10.26	16.76	20.42	24.84	29.33	35.17
300	1.83	2.57	4.25	7.53	13.82	24.36	37.75	52.39	78.08	88.86
400	1.87	3.02	5.45	9.78	18.58	34.42	53.36	87.08	140.70	180.09
600	1.94	3.70	6.50	11.22	21.20	39.47	62.53	130.89	228.25	372.52
	backward transform									
100	1.58	2.07	2.97	3.78	6.80	9.29	10.69	13.50	15.14	14.33
200	1.61	2.64	4.18	5.91	11.03	19.81	27.45	35.69	52.30	56.93
300	1.80	2.45	4.24	7.52	13.96	25.90	41.00	66.22	95.26	133.22
400	1.83	2.98	5.33	9.58	18.60	35.57	54.84	89.59	156.33	218.11
600	1.94	3.54	6.30	11.01	21.22	40.11	62.13	129.60	230.57	388.31
	double precision DCT									
	forward transform									
100	1.44	1.99	2.17	2.84	3.75	4.58	5.45	8.54	14.48	15.43
200	1.60	2.72	4.46	6.52	11.00	17.46	20.23	23.82	32.91	39.41
300	1.78	2.70	4.64	8.22	14.38	24.21	37.95	45.96	65.02	86.02
400	1.83	3.24	5.88	10.88	19.29	34.97	58.15	82.13	128.29	154.47
	backward transform									
100	1.52	2.15	2.77	3.94	6.73	10.06	10.26	7.04	13.75	14.75
200	1.60	2.68	4.49	7.38	13.33	22.49	24.98	38.40	53.37	51.18
300	1.77	2.67	4.60	8.22	15.01	26.68	41.23	47.99	80.81	114.23
400	1.81	3.19	5.70	10.74	19.38	35.45	58.13	94.07	160.43	172.13
	single precision DFT									
	forward transform									
100	1.66	2.91	4.06	5.11	7.90	11.60	11.96	12.87	21.17	22.00
200	1.89	3.47	6.25	11.16	20.41	34.17	49.06	61.54	92.30	110.87
300	1.89	3.40	6.34	11.78	22.36	39.70	71.21	104.45	173.67	222.25
400	1.94	3.81	7.23	13.69	26.16	49.44	90.28	162.72	281.33	404.33
	backward transform									
100	1.70	3.17	4.58	6.08	10.38	16.00	22.77	18.99	20.82	20.72
200	1.88	3.37	6.14	11.54	21.50	38.36	56.38	74.31	110.65	141.26
300	1.89	3.38	6.27	11.67	22.20	40.56	73.45	111.63	202.70	263.55
400	1.93	3.78	7.16	13.47	26.13	49.31	88.81	165.38	293.04	427.32
	double precision DFT									
	forward transform									
100	1.62	2.56	4.05	5.75	7.86	11.45	13.54	13.19	11.92	15.16
200	1.82	3.38	5.96	10.77	18.85	31.19	49.06	61.22	89.00	99.92
300	1.83	3.39	6.19	11.34	20.90	36.34	64.60	103.16	168.98	220.23
400	1.92	3.71	7.02	13.61	25.32	45.66	84.68	153.37	260.39	407.95
	backward transform									
100	1.66	2.67	4.37	6.65	11.66	17.58	20.57	18.39	20.99	15.04
200	1.81	3.37	5.93	10.90	19.39	33.92	54.65	77.85	117.11	140.65
300	1.83	3.37	6.11	11.49	20.95	37.25	64.82	111.71	186.97	277.44
400	1.92	3.61	6.93	13.35	25.16	46.24	85.44	157.80	270.35	397.71

that the code was tested on the machines in case of which we deal with a discrepancy between the physical layout of the computing nodes and the layout assumed by the method. Nevertheless, we believe that the initial results are encouraging enough to continue work. Here, the first step will be to perform more involved testing of the performance to establish performance profile (especially for the largest problems). We also plan to investigate the performance on the cluster utilizing Intel Phi coprocessors.

ACKNOWLEDGMENTS

Computer time grants from the TASK computing center in Gdansk, Poland, the Bulgarian Supercomputing Center, and the HPC Center from West University of Timisoara are kindly acknowledged. This research was partially supported by grants DCVP 02/1 and I01/5 from the Bulgarian NSF. Work presented here is a part of the Poland-Bulgaria collaborative grant "Parallel and distributed computing practices." Work of Marcin Paprzycki was completed in part, while he was visiting University of Aizu.

REFERENCES

[1] O. Ayala and L.P. Wang. Parallel implementation and scalability analysis of 3D fast Fourier transform using 2D domain decomposition. *Parallel Computing*, 2012. DOI: 10.1016/j.parco.2012.12.002

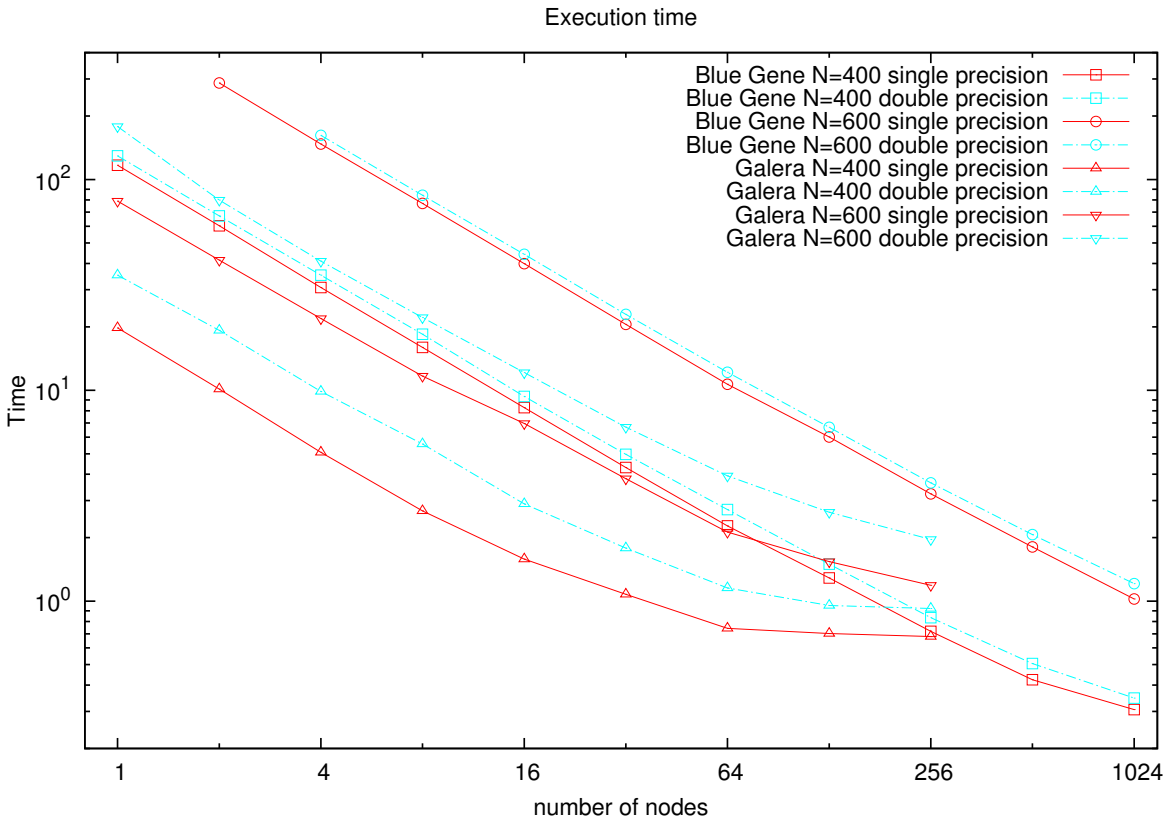


Fig. 1. Execution time of code which performs forward and backward DFT for $N = 400, 600$.

[2] Jack J. Dongarra, Jeremy Du Croz, Sven Hammarling, and Iain S. Duff. A set of level 3 basic linear algebra subprograms. *ACM Transactions on Mathematical Software (TOMS)*, 16(1):1–17, 1990. DOI: 10.1145/77626.79170

[3] Maria Eleftheriou, José E. Moreira, Blake G. Fitch, and Robert S. Germain. A volumetric FFT for Blue Gene/L. In Timothy Mark Pinkston and Viktor K. Prasanna, editors, *High Performance Computing - HiPC 2003*, volume 2913 of *Lecture Notes in Computer Science*, pages 194–203. Springer Berlin Heidelberg, 2003. DOI: 10.1007/978-3-540-24596-4_21

[4] Ning Li and Sylvain Laizet. 2DECOMP&FFT - A Highly Scalable 2D Decomposition: Library and FFT Interface. In *Cray User Group 2010 conference*, pages 1–13, 2010.

[5] Stanislav G. Sedukhin, Co-design of Extremely Scalable Algorithms/Architecture for 3-Dimensional Linear Transforms, Technical Report TR2012-001, The University of Aizu, July 2012.

[6] Snir, M., Otto, S., Huss-Lederman, S., Walker, D., Dongarra, J.: MPI: The Complete Reference, second edition, Volume 1, The MPI Core. Scientific and engineering computation series, The MIT Press, Cambridge, Massachusetts (1998), ISBN: 9780262692151

[7] Walker, D., Dongarra, J.: MPI: a standard Message Passing Interface. Supercomputer, 12 (1), 56–68 (1996), ISSN 0168-7875

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 623–629
DOI: 10.15439/2014F91
ACSIS, Vol. 2

MuPAD codes which implement limit-computable functions that cannot be bounded by any computable function

Apoloniusz Tyszka
University of Agriculture
Faculty of Production and Power Engineering
Balicka 116B, 30-149 Kraków, Poland
Email: rttyszka@cyf-kr.edu.pl

***Abstract*—Let $E_n = \{x_k = 1,\ x_i + x_j = x_k,\ x_i \cdot x_j = x_k :\ i,j,k \in \{1,\ldots,n\}\}$. For a positive integer n, let $f(n)$ denote the smallest non-negative integer b such that for each system $S \subseteq E_n$ with a solution in non-negative integers $x_1,\ldots,x_n$ there exists a solution of S in non-negative integers not greater than b. We prove that if a function $\Gamma : \mathbb{N} \setminus \{0\} \to \mathbb{N}$ is computable, then f dominates Γ i.e. there exists a positive integer m such that $\Gamma(n) < f(n)$ for any $n \geq m$. For positive integers n, m, let $g(n,m)$ denote the smallest non-negative integer b such that for each system $S \subseteq E_n$ with a solution in $\{0,\ldots,m-1\}^n$ there exists a solution of S in $\{0,\ldots,b\}^n$. Then,**

$$g(n,m) \leq m-1, \tag{1}$$

$$0 = g(n,1) < 1 = g(n,2) \leq g(n,3) \leq g(n,4) \leq \ldots \tag{2}$$

and

$$g(n,f(n)) < f(n) = g(n,f(n)+1) =$$
$$g(n,f(n)+2) = g(n,f(n)+3) = \ldots \tag{3}$$

We present an infinite loop in *MuPAD* which takes as input a positive integer n and returns $g(n,m)$ on the m-th iteration.

***Index Terms*—Hilbert's Tenth Problem, infinite loop, limit-computable function, *MuPAD*, trial-and-error computable function.**

LIMIT-computable functions, also known as trial-and-error computable functions, have been thoroughly studied, see [6, pp. 233–235] for the main results. Our first goal is to present an infinite loop in *MuPAD* which finds the values of a limit-computable function $f : \mathbb{N} \setminus \{0\} \to \mathbb{N} \setminus \{0\}$ by an infinite computation, where f dominates all computable functions. There are many limit-computable functions $f : \mathbb{N} \setminus \{0\} \to \mathbb{N} \setminus \{0\}$ which cannot be bounded by any computable function. For example, this follows from [2, p. 38, item 4], see also [5, p. 268] where Janiczak's result is mentioned. Unfortunately, for all known such functions f, it is difficult to write a suitable computer program. The sophisticated choice of a function f will allow us to do so.

Let

$$E_n = \{x_k = 1,\ x_i + x_j = x_k,\ x_i \cdot x_j = x_k :\ i,j,k \in \{1,\ldots,n\}\}.$$

For a positive integer n, let $f(n)$ denote the smallest non-negative integer b such that for each system $S \subseteq E_n$ with a solution in non-negative integers $x_1,\ldots,x_n$ there exists a solution of S in non-negative integers not greater than b. This definition is correct because there are only finitely many subsets of E_n. For positive integers n, m, let $g(n,m)$ denote the smallest non-negative integer b such that for each system $S \subseteq E_n$ with a solution in $\{0,\ldots,m-1\}^n$ there exists a solution of S in $\{0,\ldots,b\}^n$. Then, conditions (1)-(3) stated in the abstract hold.

Obviously, $f(1) = 1$. The system

$$\begin{cases} x_1 = 1 \\ x_1 + x_1 = x_2 \\ x_2 \cdot x_2 = x_3 \\ x_3 \cdot x_3 = x_4 \\ \ldots \\ x_{n-1} \cdot x_{n-1} = x_n \end{cases}$$

has a unique integer solution, namely $\left(1,2,4,16,\ldots,2^{2^{n-3}},2^{2^{n-2}}\right)$. Therefore, $f(n) \geq 2^{2^{n-2}}$ for any $n \geq 2$.

The Davis-Putnam-Robinson-Matiyasevich theorem states that every recursively enumerable set $\mathcal{M} \subseteq \mathbb{N}^n$ has a Diophantine representation, that is

$$(a_1,\ldots,a_n) \in \mathcal{M} \iff$$
$$\exists x_1,\ldots,x_m \in \mathbb{N}\quad W(a_1,\ldots,a_n,x_1,\ldots,x_m) = 0 \qquad (\mathrm{R})$$

for some polynomial W with integer coefficients, see [3]. The polynomial W can be computed, if we know the Turing machine M such that, for all $(a_1,\ldots,a_n) \in \mathbb{N}^n$, M halts on $(a_1,\ldots,a_n)$ if and only if $(a_1,\ldots,a_n) \in \mathcal{M}$, see [3]. The representation (R) is said to be single-fold, if for any $a_1,\ldots,a_n \in \mathbb{N}$ the equation $W(a_1,\ldots,a_n,x_1,\ldots,x_m) = 0$ has at most one solution $(x_1,\ldots,x_m) \in \mathbb{N}^m$. Yu. Matiyasevich conjectures that each recursively enumerable set $\mathcal{M} \subseteq \mathbb{N}^n$ has a single-fold Diophantine representation, see [4].

Let $\mathcal{R}ng$ denote the class of all rings $\boldsymbol{K}$ that extend $\mathbb{Z}$.

Lemma ([8, p. 720]). *Let $D(x_1,\ldots,x_p) \in \mathbb{Z}[x_1,\ldots,x_p]$. Assume that $\deg(D,x_i) \geq 1$ for each $i \in \{1,\ldots,p\}$. We can compute a positive integer $n > p$ and a system $T \subseteq E_n$ which satisfies the following two conditions:*

978-83-60810-58-3/$25.00

`Condition 1.` *If* $\boldsymbol{K} \in \mathcal{R}ng \cup \{\mathbb{N},\ \mathbb{N} \setminus \{0\}\}$*, then*

$$\forall \tilde{x}_1, \ldots, \tilde{x}_p \in \boldsymbol{K} \left(D(\tilde{x}_1, \ldots, \tilde{x}_p) = 0 \Longleftrightarrow \right.$$

$$\left. \exists \tilde{x}_{p+1}, \ldots, \tilde{x}_n \in \boldsymbol{K}\ (\tilde{x}_1, \ldots, \tilde{x}_p, \tilde{x}_{p+1}, \ldots, \tilde{x}_n)\ \textit{solves}\ T \right)$$

`Condition 2.` *If* $\boldsymbol{K} \in \mathcal{R}ng \cup \{\mathbb{N},\ \mathbb{N} \setminus \{0\}\}$*, then for each* $\tilde{x}_1, \ldots, \tilde{x}_p \in \boldsymbol{K}$ *with* $D(\tilde{x}_1, \ldots, \tilde{x}_p) = 0$*, there exists a unique tuple* $(\tilde{x}_{p+1}, \ldots, \tilde{x}_n) \in \boldsymbol{K}^{n-p}$ *such that the tuple* $(\tilde{x}_1, \ldots, \tilde{x}_p, \tilde{x}_{p+1}, \ldots, \tilde{x}_n)$ *solves* T*.*

Conditions 1 and 2 imply that for each $\boldsymbol{K} \in \mathcal{R}ng \cup \{\mathbb{N},\ \mathbb{N} \setminus \{0\}\}$*, the equation* $D(x_1, \ldots, x_p) = 0$ *and the system* T *have the same number of solutions in* $\boldsymbol{K}$*.*

Theorem 1. *If a function* $\Gamma : \mathbb{N} \setminus \{0\} \to \mathbb{N}$ *is computable, then there exists a positive integer* m *such that* $\Gamma(n) < f(n)$ *for any* $n \geq m$*.*

Proof. The Davis-Putnam-Robinson-Matiyasevich theorem and the Lemma for $\boldsymbol{K} = \mathbb{N}$ imply that there exists an integer $s \geq 3$ such that for any non-negative integers x_1, x_2,

$$(x_1, x_2) \in \Gamma \Longleftrightarrow \exists x_3, \ldots, x_s \in \mathbb{N}\ \ \Phi(x_1, x_2, x_3, \ldots, x_s), \quad \texttt{(E)}$$

where the formula $\Phi(x_1, x_2, x_3, \ldots, x_s)$ is a conjunction of formulae of the forms $x_k = 1$, $x_i + x_j = x_k$, $x_i \cdot x_j = x_k$ $(i, j, k \in \{1, \ldots, s\})$. Let $[\cdot]$ denote the integer part function. For each integer $n \geq 6 + 2s$,

$$n - \left[\frac{n}{2}\right] - 3 - s \geq 6 + 2s - \left[\frac{6+2s}{2}\right] - 3 - s \geq 6 + 2s - \frac{6+2s}{2} - 3 - s = 0$$

For an integer $n \geq 6 + 2s$, let S_n denote the following system

$$\begin{cases} \text{all equations occurring in} \\ \qquad \Phi(x_1, x_2, x_3, \ldots, x_s) \\ n - \left[\frac{n}{2}\right] - 3 - s \text{ equations} \\ \qquad \text{of the form } z_i = 1 \\ \qquad\qquad t_1 = 1 \\ \qquad\qquad t_1 + t_1 = t_2 \\ \qquad\qquad t_2 + t_1 = t_3 \\ \qquad\qquad \ldots \\ \qquad\qquad t_{[\frac{n}{2}]-1} + t_1 = t_{[\frac{n}{2}]} \\ \qquad\qquad t_{[\frac{n}{2}]} + t_{[\frac{n}{2}]} = w \\ \qquad\qquad w + y = x_1 \\ \qquad\qquad y + y = y \text{ (if } n \text{ is even)} \\ \qquad\qquad y = 1 \text{ (if } n \text{ is odd)} \\ \qquad\qquad x_2 + t_1 = u \end{cases}$$

with n variables. By the equivalence `(E)`, S_n is satisfiable over $\mathbb{N}$. If a n-tuple $(x_1, x_2, x_3, \ldots, x_s, \ldots, w, y, u)$ of non-negative integers solves S_n, then by the equivalence `(E)`,

$$x_2 = \Gamma(x_1) = \Gamma(w + y) = \Gamma\left(2 \cdot \left[\frac{n}{2}\right] + y\right) = \Gamma(n)$$

Therefore, $u = x_2 + t_1 = \Gamma(n) + 1 > \Gamma(n)$. This shows that $\Gamma(n) < f(n)$ for any $n \geq 6 + 2s$. □

Theorem 2. *There exists a computable function* $\varphi : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ *which satisfies the following conditions:*

`1)` *For each non-negative integers* n *and* l*,*

$$\varphi(n, l) \leq l$$

`2)` *For each non-negative integer* n*,*

$$0 = \varphi(n, 0) < 1 = \varphi(n, 1) \leq \varphi(n, 2) \leq \varphi(n, 3) \leq \ldots$$

`3)` *For each non-negative integer* n*, the sequence* $\{\varphi(n, l)\}_{l \in \mathbb{N}}$ *is bounded from above.*

`4)` *The function*

$$\mathbb{N} \ni n \overset{\theta}{\longrightarrow} \theta(n) = \lim_{l \to \infty} \varphi(n, l) \in \mathbb{N} \setminus \{0\}$$

dominates all computable functions.

`5)` *For each non-negative integer* n*,*

$$\varphi(n, \theta(n) - 1) < \theta(n) = \varphi(n, \theta(n)) =$$

$$\varphi(n, \theta(n) + 1) = \varphi(n, \theta(n) + 2) = \ldots$$

Proof. Let us say that a tuple $y = (y_1, \ldots, y_n) \in \mathbb{N}^n$ is a *duplicate* of a tuple $x = (x_1, \ldots, x_n) \in \mathbb{N}^n$, if

$$(\forall k \in \{1, \ldots, n\}\ (x_k = 1 \Longrightarrow y_k = 1)) \wedge$$
$$(\forall i, j, k \in \{1, \ldots, n\}\ (x_i + x_j = x_k \Longrightarrow y_i + y_j = y_k)) \wedge$$
$$(\forall i, j, k \in \{1, \ldots, n\}\ (x_i \cdot x_j = x_k \Longrightarrow y_i \cdot y_j = y_k))$$

For non-negative integers n and l, we define $\varphi(n, l)$ as the smallest non-negative integer b such that for each $x \in \{0, \ldots, l\}^{n+1}$ there exists a duplicate of x in $\{0, \ldots, b\}^{n+1}$. Theorem 1 implies the claim of item `4)` whereas the following *MuPAD* code performs a Turing computation of $\varphi(n, l)$.

```
input("input the value of n",n):
input("input the value of l",l):
n:=n+1:
X:=[i $ i=0..l]:
Y:=combinat::cartesianProduct(X $i=1..n):
W:=combinat::cartesianProduct(X $i=1..n):
for s from 1 to nops(Y) do
for t from 1 to nops(Y) do
m:=0:
for i from 1 to n do
if Y[s][i]=1 and Y[t][i]<>1
then m:=1 end_if:
for j from i to n do
for k from 1 to n do
if Y[s][i]+Y[s][j]=Y[s][k] and
Y[t][i]+Y[t][j]<>Y[t][k]
then m:=1 end_if:
if Y[s][i]*Y[s][j]=Y[s][k] and
Y[t][i]*Y[t][j]<>Y[t][k]
then m:=1 end_if:
end_for:
end_for:
end_for:
if m=0 and
max(Y[t][i] $i=1..n)<max(Y[s][i] $i=1..n)
then W:=listlib::setDifference(W,[Y[s]])
end_if:
```

```
end_for:
end_for:
print(max(max(W[z][u] $u=1..n) $z=1..nops(W))):
```

Code 1
A Turing computation of $\varphi(n,l)$

□

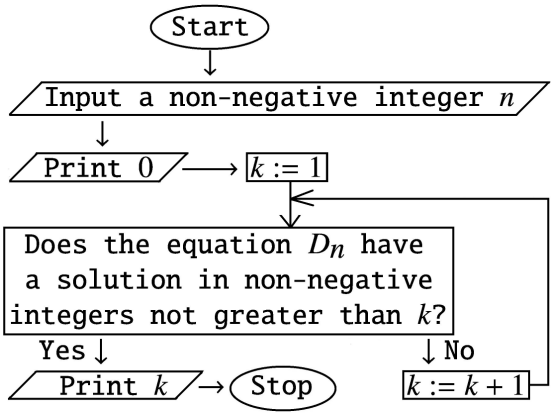
Code 1 is also stored in [10]. The following algorithm performs an infinite computation of $f(n)$, because it returns $g(n,m)$ on the m-th iteration, where m stands for any positive integer.

```
input("input the value of n",n):
i:=0:
while TRUE do
print(φ(n-1,i)):
i:=i+1:
end_while:
```

Algorithm 1
An infinite computation of $f(n)$

A slightly changed *MuPAD* code that implements Algorithm 1 is stored in [10, Code 4].

Let us fix a computable enumeration $D_0, D_1, D_2, \ldots$ of all Diophantine equations. The following flowchart illustrates an infinite computation of a limit-computable function that cannot be bounded by any computable function.



Algorithm 2
A loop whose execution does not always terminate, and that defines a partially computable function that cannot be bounded by any computable function from $\mathbb{N}$ to $\mathbb{N}$

For each non-negative integer n, the function has a non-zero value at n if and only if the equation D_n has a solution in non-negative integers. Unfortunately, the function does not have any easy implementation.

The following *MuPAD* code is stored in [10].

```
input("input the value of n",n):
print(0):
A:=op(ifactor(210*(n+1))):
B:=[A[2*i+1] $i=1..(nops(A)-1)/2]:
S:={}:
for i from 1 to floor(nops(B)/4) do
if B[4*i]=1 then
S:=S union {B[4*i-3]} end_if:
if B[4*i]=2 then S:=S union
{[B[4*i-3],B[4*i-2],B[4*i-1],"+"]}
end_if:
if B[4*i]>2 then S:=S union
{[B[4*i-3],B[4*i-2],B[4*i-1],"*"]}
end_if:
end_for:
m:=2:
repeat
C:=op(ifactor(m)):
W:=[C[2*i+1]-1 $i=1..(nops(C)-1)/2]:
T:={}:
for i from 1 to nops(W) do
for j from 1 to nops(W) do
for k from 1 to nops(W) do
if W[i]=1 then T:=T union {i} end_if:
if W[i]+W[j]=W[k] then
T:=T union {[i,j,k,"+"]} end_if:
if W[i]*W[j]=W[k] then
T:=T union {[i,j,k,"*"]} end_if:
end_for:
end_for:
end_for:
m:=m+1:
until S minus T={} end_repeat:
print(max(W[i] $i=1..nops(W))):
```

Code 2
A loop whose execution does not always terminate, and that defines a partially computable function that cannot be bounded by any computable function from $\mathbb{N}$ to $\mathbb{N}$

Theorem 3. *The above code implements a limit-computable function $\xi : \mathbb{N} \to \mathbb{N}$ that cannot be bounded by any computable function. The code takes as input a non-negative integer n, returns 0, and computes a system S of polynomial equations. If the loop terminates for S, then the next instruction returns $\xi(n)$. If the loop does not terminate, then $\xi(n) = 0$. The loop defines a partially computable function that cannot be bounded by any computable function from $\mathbb{N}$ to $\mathbb{N}$.*

Proof. Let $n \in \mathbb{N}$, and let ${p_1}^{t(1)} \cdot \ldots \cdot {p_s}^{t(s)}$ be a prime factorization of $210 \cdot (n+1)$, where $t(1), \ldots, t(s)$ denote positive integers. Obviously, $p_1 = 2$, $p_2 = 3$, $p_3 = 5$, and $p_4 = 7$.

For each positive integer i that satisfies $4i \leq s$ and $t(4i) = 1$, the code constructs the equation $x_{t(4i-3)} = 1$.

For each positive integer i that satisfies $4i \leq s$ and $t(4i) = 2$, the code constructs the equation $x_{t(4i-3)} + x_{t(4i-2)} = x_{t(4i-1)}$.

For each positive integer i that satisfies $4i \leq s$ and $t(4i) > 2$, the code constructs the equation $x_{t(4i-3)} \cdot x_{t(4i-2)} = x_{t(4i-1)}$.

The last three facts imply that the code assigns to n a finite and non-empty system S which consists of equations of the

forms: $x_k = 1$, $x_i + x_j = x_k$, and $x_i \cdot x_j = x_k$. Conversely, each such system S is assigned to some non-negative integer n.

Starting with the instruction $m := 2$, the code tries to find a solution of S in non-negative integers by performing a brute-force search. If a solution exists, then the search terminates and the code returns a non-negative integer $\xi(n)$ such that the system S has a solution in non-negative integers not greater than $\xi(n)$. In the opposite case, the execution of the code never terminates.

A negative solution to Hilbert's Tenth Problem ([3]) and the Lemma for $\boldsymbol{K} = \mathbb{N}$ imply that the code implements a limit-computable function $\xi : \mathbb{N} \to \mathbb{N}$ that cannot be bounded by any computable function. □

The execution of the last code does not terminate for $n = 7 \cdot 11 \cdot 13 \cdot 17 \cdot 19 - 1 = 323322$, when the code tries to find a solution of the system $\{x_1 + x_1 = x_1,\ x_1 = 1\}$. Execution terminates for any $n < 323322$, when the code returns 0 and next 1 or 0. The last claim holds only theoretically. In fact, for $n = 2^{18} - 1 = 262143$, the algorithm of the code returns 1 solving the equation $x_{19} = 1$ on the $\left(2 \cdot 3 \cdot 5 \cdot 7 \cdot 11 \cdot 13 \cdot 17 \cdot 19 \cdot 23 \cdot 29 \cdot 31 \cdot 37 \cdot 41 \cdot 43 \cdot 47 \cdot 53 \cdot 59 \cdot 61 \cdot 67^2 - 1\right)$-th iteration.

Let $\mathcal{P}$ denote a predicate calculus with equality and one binary relation symbol, and let Λ be a computable function that maps $\mathbb{N}$ onto the set of sentences of $\mathcal{P}$. The following pseudocode in *MuPAD* implements a limit-computable function $\sigma : \mathbb{N} \to \mathbb{N}$ that cannot be bounded by any computable function.

```
input("input the value of n",n):
print(0):
k:=1:
while Λ(n) holds in all models of size k do
k:=k+1:
end_while:
print(k):
```

Algorithm 3

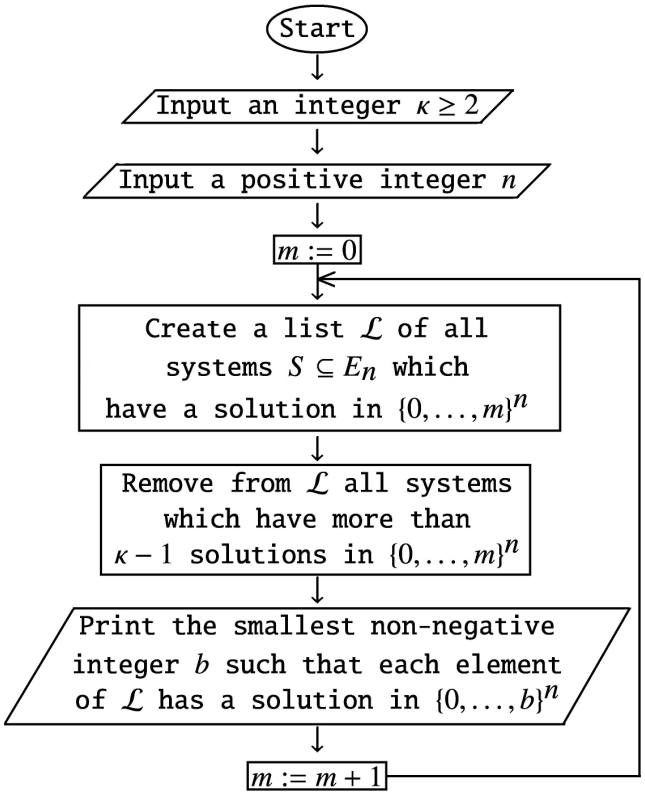
A loop whose execution does not always terminate, and that defines a partially computable function that cannot be bounded by any computable function from $\mathbb{N}$ to $\mathbb{N}$

The proof follows from the fact that the set of sentences of $\mathcal{P}$ that are true in all finite and non-empty models is not recursively enumerable, see [1, p. 129], where it is concluded from Trakhtenbrot's theorem. The author has no idea how to transform the pseudocode into a correct computer program.

The commercial version of *MuPAD* is no longer available as a stand-alone product, but only as the *Symbolic Math Toolbox* of *MATLAB*. Fortunately, the presented codes can be executed by *MuPAD Light*, which was and is free, see [11]. Similar codes in *MuPAD Light* are presented and discussed at http://arxiv.org/abs/1310.5363.

Limit-computable functions are related to the question of the decidability of Diophantine equations with a finite number of solutions in non-negative integers. Let $\kappa \in \{2, 3, 4, \ldots, \omega, \omega_1\}$. For a positive integer n, let $f_K(n)$ denote the smallest non-negative integer b such that for each system $S \subseteq E_n$ which has a solution in non-negative integers $x_1, \ldots, x_n$ and which has less than κ solutions in non-negative integers $x_1, \ldots, x_n$, there exists a solution of S in non-negative integers not greater than b. Since $f_{\omega_1} = f$, f_{ω_1} is limit-computable by Algorithm 1.

Obviously, $f_2(n)$ is the smallest non-negative integer b such that for each system $S \subseteq E_n$ with a unique solution in non-negative integers $x_1, \ldots, x_n$ this solution belongs to $[0, b]^n$. If $\kappa < \omega$, then the function f_K is limit-computable as the flowchart below describes an infinite computation of $f_K(n)$.



Algorithm 4

An infinite computation of $f_K(n)$

The following *MuPAD* code is stored in [10, Code 3] and performs an infinite computation of $f_2(n)$.

```
input("input the value of n",n):
X:=[0]:
while TRUE do
Y:=combinat::cartesianProduct(X $i=1..n):
W:=combinat::cartesianProduct(X $i=1..n):
for s from 1 to nops(Y) do
for t from 1 to nops(Y) do
m:=0:
for i from 1 to n do
if Y[s][i]=1 and Y[t][i]<>1 then m:=1 end_if:
for j from i to n do
for k from 1 to n do
```

```
if Y[s][i]+Y[s][j]=Y[s][k] and
Y[t][i]+Y[t][j]<>Y[t][k] then m:=1 end_if:
if Y[s][i]*Y[s][j]=Y[s][k] and
Y[t][i]*Y[t][j]<>Y[t][k] then m:=1 end_if:
end_for:
end_for:
end_for:
if m=0 and s<>t then
W:=listlib::setDifference(W,[Y[s]]) end_if:
end_for:
end_for:
print(max(max(W[z][u] $u=1..n) $z=1..nops(W))):
X:=append(X,nops(X)):
end_while:
```

Code 3
An infinite computation of $f_2(n)$

Theorem 5 implies that f_2 dominates any function $h : \mathbb{N} \setminus \{0\} \to \mathbb{N}$ with a single-fold Diophantine representation. Therefore, Matiyasevich's conjecture on single-fold Diophantine representations implies that f_2 dominates all computable functions from $\mathbb{N} \setminus \{0\}$ to $\mathbb{N}$.

Obviously, $f_K(1) = 1$ and $f_K(n) \geq 2^{2^{n-2}}$ for any $n \geq 2$. Theorem 1 implies that the equality

$$f_K = \{(1,1)\} \cup \left\{\left(n, 2^{2^{n-2}}\right) : n \in \{2,3,4,\ldots\}\right\}$$

is false for $\kappa = \omega_1$. The above equality is also false for any $\kappa \in \{2,3,4,\ldots,\omega\}$. The conjecture in [8] is false. The conjecture in [9] is false. The last three results were recently communicated to the author.

The representation (R) is said (here and further) to be κ-fold, if for any $a_1,\ldots,a_n \in \mathbb{N}$ the equation $W(a_1,\ldots,a_n,x_1,\ldots,x_m) = 0$ has less than κ solutions $(x_1,\ldots,x_m) \in \mathbb{N}^m$

Theorem 4. *([7, Theorem 2]) Let us consider the following three statements:*

(*a*) *There exists an algorithm $\mathcal{A}$ whose execution always terminates and which takes as input a Diophantine equation D and returns the answer* YES *or* NO *which indicates whether or not the equation D has a solution in non-negative integers, if the solution set $Sol(D)$ satisfies* $\mathrm{card}(Sol(D)) < \kappa$.

(*b*) *The function f_K is majorized by a computable function.*

(*c*) *If a set $\mathcal{M} \subseteq \mathbb{N}^n$ has a κ-fold Diophantine representation, then $\mathcal{M}$ is computable.*

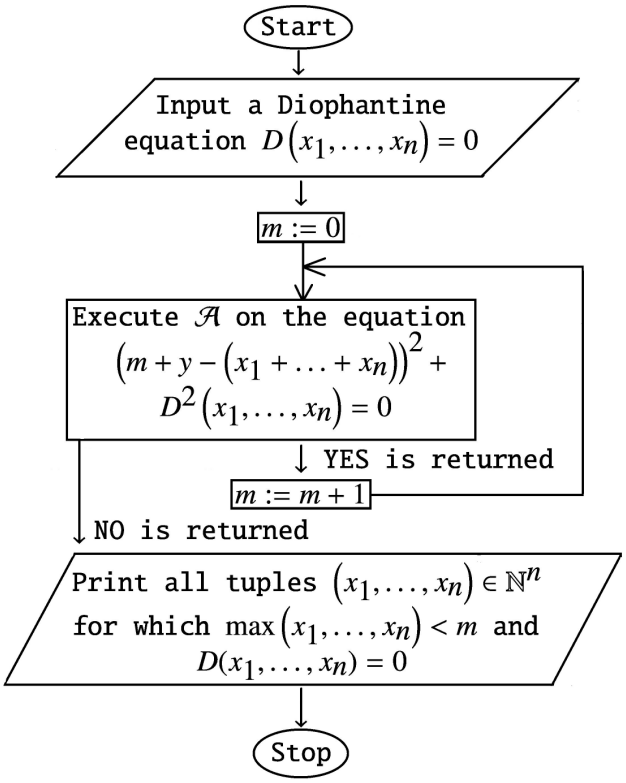
We claim that (*a*) *is equivalent to* (*b*) *and* (*a*) *implies* (*c*).

Proof. The implication $(a) \Rightarrow (c)$ is obvious. We prove the implication $(a) \Rightarrow (b)$. There is an algorithm Dioph which takes as input a positive integer m and a non-empty system $S \subseteq E_m$, and returns a Diophantine equation $\mathrm{Dioph}(m,S)$ which has the same solutions in non-negative integers $x_1,\ldots,x_m$. Item (*a*) implies that for each Diophantine equation D, if the algorithm $\mathcal{A}$ returns YES for D, then D has a solution in non-negative integers. Hence, if the algorithm $\mathcal{A}$ returns YES for $\mathrm{Dioph}(m,S)$, then we can compute the smallest non-negative integer $i(m,S)$ such that $\mathrm{Dioph}(m,S)$ has a solution in non-negative integers not greater than $i(m,S)$. If the algorithm $\mathcal{A}$ returns NO for $\mathrm{Dioph}(m,S)$, then we set $i(m,S) = 0$. The function

$$\mathbb{N} \setminus \{0\} \ni m \to \max\left\{i(m,S) : \emptyset \neq S \subseteq E_m\right\} \in \mathbb{N}$$

is computable and majorizes the function f_K. We prove the implication $(b) \Rightarrow (a)$. Let a function h majorizes f_K. By the Lemma for $\boldsymbol{K} = \mathbb{N}$, a Diophantine equation D is equivalent to a system $S \subseteq E_n$. The algorithm $\mathcal{A}$ checks whether or not S has a solution in non-negative integers $x_1,\ldots,x_n$ not greater than $h(n)$. □

The implication $(a) \Rightarrow (c)$ remains true with a weak formulation of item (*a*), where the execution of $\mathcal{A}$ may not terminate or $\mathcal{A}$ may return nothing or something irrelevant, if D has at least κ solutions in non-negative integers. The weakened item (*a*) implies that the following flowchart



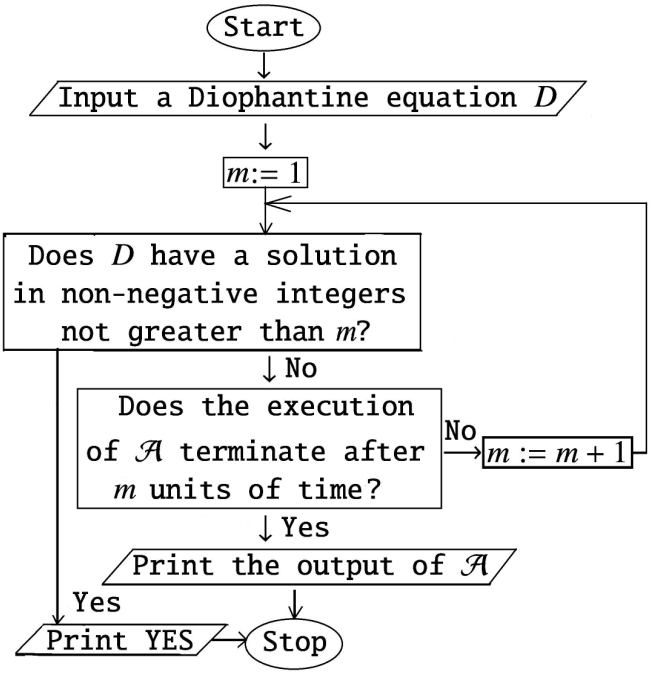
Algorithm 5
An algorithm that conditionally finds all solutions to a Diophantine equation which has less than κ solutions in non-negative integers

describes an algorithm whose execution terminates, if the set

$$Sol(D) := \{(x_1,\ldots,x_n) \in \mathbb{N}^n : \ D(x_1,\ldots,x_n) = 0\}$$

has less than κ elements. If this condition holds, then the weakened item (*a*) guarantees that the execution of the flowchart prints all elements of $Sol(D)$. However, the weakened item (*a*) is equivalent to the original one. Indeed, if the algorithm $\mathcal{A}$

satisfies the weakened item (*a*), then the flowchart below illustrates a new algorithm $\mathcal{A}$ that satisfies the original item (*a*).



Algorithm 6
The weakened item (*a*) implies the original one

The equality $f_{\omega_1} = f$ and Theorem 1 imply that item (*b*) is false for $\kappa = \omega_1$. By this and Theorem 4, we alternatively obtain a negative solution to Hilbert's Tenth Problem.

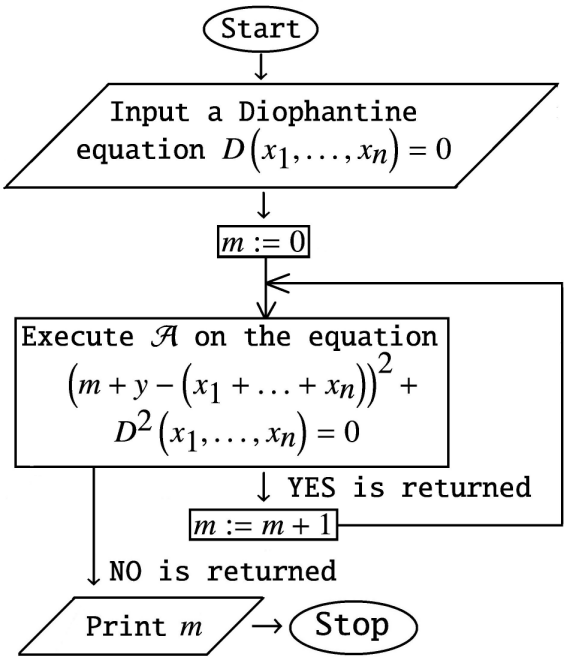
Theorem 5. *([7, Theorem 1]) If a function $h : \mathbb{N} \setminus \{0\} \to \mathbb{N}$ has a κ-fold Diophantine representation, then there exists a positive integer m such that $h(n) < f_K(n)$ for any $n \geq m$.*

By the Davis-Putnam-Robinson-Matiyasevich theorem, Theorem 1 is a special case of Theorem 5 when $\kappa = \omega_1$. Let us pose the following two questions:

Question 1. *Is there an algorithm $\mathcal{B}$ which takes as input a Diophantine equation D, returns an integer, and this integer is greater than the heights of non-negative integer solutions, if the solution set has less than κ elements? We allow a possibility that the execution of $\mathcal{B}$ does not terminate or $\mathcal{B}$ returns nothing or something irrelevant, if D has at least κ solutions in non-negative integers.*

Question 2. *Is there an algorithm C which takes as input a Diophantine equation D, returns an integer, and this integer is greater than the number of non-negative integer solutions, if the solution set is finite? We allow a possibility that the execution of C does not terminate or C returns nothing or something irrelevant, if D has infinitely many solutions in non-negative integers.*

Obviously, a positive answer to Question 1 implies the weakened item (*a*). Conversely, the weakened item (*a*) implies that the flowchart below describes an appropriate algorithm $\mathcal{B}$.



Algorithm 7
The weakened item (*a*) implies a positive answer to Question 1

Theorem 6. *A positive answer to Question 1 for $\kappa = \omega$ is equivalent to a positive answer to Question 2.*

Proof. Trivially, a positive answer to Question 1 for $\kappa = \omega$ implies a positive answer to Question 2. Conversely, if a Diophantine equation $D(x_1, \ldots, x_n) = 0$ has only finitely many solutions in non-negative integers, then the number of non-negative integer solutions to the equation

$$D^2(x_1, \ldots, x_n) + (x_1 + \ldots + x_n - y - z)^2 = 0$$

is finite and greater than $\max(a_1, \ldots, a_n)$, where $(a_1, \ldots, a_n) \in \mathbb{N}^n$ is any solution to $D(x_1, \ldots, x_n) = 0$. □

References

[1] H.-D. Ebbinghaus and J. Flum, *Finite model theory,* Springer-Verlag, Berlin, 2006.
[2] A. Janiczak, *Some remarks on partially recursive functions,* Colloquium Math. 3 (1954), 37–38.
[3] Yu. Matiyasevich, *Hilbert's tenth problem,* MIT Press, Cambridge, MA, 1993.
[4] Yu. Matiyasevich, *Towards finite-fold Diophantine representations,* Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) 377 (2010), 78–90, ftp://ftp.pdmi.ras.ru/pub/publicat/znsl/v377/p078.pdf, http://dx.doi.org/10.1007/s10958-010-0179-4.
[5] R. Murawski, *The contribution of Polish logicians to recursion theory,* in: K. Kijania-Placek and J. Woleński (eds.), *The Lvov-Warsaw School and Contemporary Philosophy,* 265–282, Kluwer Acad. Publ., Dordrecht, 1998.
[6] R. I. Soare, *Interactive computing and relativized computability,* in: *Computability: Turing, Gödel, Church, and beyond* (eds. B. J. Copeland, C. J. Posy, and O. Shagrir), MIT Press, Cambridge, MA, 2013, 203–260.
[7] A. Tyszka, *A condition equivalent to the decidability of Diophantine equations with a finite number of solutions in non-negative integers,* http://arxiv.org/abs/1404.5975.
[8] A. Tyszka, *Conjecturally computable functions which unconditionally do not have any finite-fold Diophantine representation,* Inform. Process. Lett. 113 (2013), no. 19–21, 719–722, http://dx.doi.org/10.1016/j.ipl.2013.07.004.

[9] A. Tyszka, *Does there exist an algorithm which to each Diophantine equation assigns an integer which is greater than the modulus of integer solutions, if these solutions form a finite set?* Fund. Inform. 125(1): 95–99, 2013, http://dx.doi.org/10.3233/FI-2013-854.

[10] A. Tyszka, *Four MuPAD codes,* http://www.cyf-kr.edu.pl/~rttyszka/codes.txt.

[11] A. Tyszka, *Links to an installation file for MuPAD Light,* http://www.ts.mah.se/utbild/ma7005/mupad_light_scilab_253.exe, http://caronte.dma.unive.it/info/materiale/mupad_light_scilab_253.exe, http://www.cyf-kr.edu.pl/~rttyszka/mupad_light_scilab_253.exe, http://www.cyf-kr.edu.pl/~rttyszka/mupad_light_253.exe, http://www.projetos.unijui.edu.br/matematica/amem/mupad/mupad_light_253.exe.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 631–637
DOI: 10.15439/2014F269
ACSIS, Vol. 2

On multivariate cryptosystems based on maps with logarithmically invertible decomposition corresponding to walk on graph

Vasyl Ustimenko
Maria Curie-Skłodowska University,
Institute of Mathematics,
pl. M. Curie-Skłodowskiej 1, 20-031 Lublin, Poland
Email: vasyl@hektor.umcs.lublin.pl

***Abstract*—The paper illustrates the concept of the map with logarithmically invertible decomposition. We introduce families of multivariate cryptosystems such that there security level is connected with discrete logarithm problem in Cremona group. The private key of such cryptosystem is a modification of graph based stream ciphers which use stable multivariate maps. Modified version corresponds to a stable map with single disturbance. If the disturbance (or initial condition) allows fast computation then modified version is almost as robust as original one. Methods of modification improve the resistance of such stream ciphers implemented on numerical level to straightforward linearisation attacks.**

I. Introduction

THE FORMAL concepts of multivariate map with logarithmically invertible decomposition is introduced by the author in Extended Abstracts Of Central European Conference on Cryptology, 2014. In this paper the examples of a cryptosystem based on this idea will be presented. The complexity estimates of an encryption and a decryption procedures are given. The construction uses walks on graphs $D(n,K)$ or $A(n,K)$ for purpose of Multivariate Cryptography. Such walks firstly used for the constructions of fast stream ciphers. The multivariate maps induced by such walks turn out to be fast cubical transformations of the plainspace (variety of vertices or variety of flags (see [1], [2]). It makes them useful for a design of stream ciphers and key exchange protocols. It was shown [3], [4] that the inverses of encryption maps are also cubical transformations. This fact restricts their use in public key cryptography. In [5] more general idea of multivariate map corresponding to *symbolic walk* on the graph has been introduced. Paper [6] suggests the deformation of such nonlinear map by two affine transformations and the use of deformated transformation in Multivariate Cryptography, but important questions of estimation of degrees, orders, densities are still under investigation.

Currently symbolic walks are used for the development of stream ciphers with high resistance to plaintext - ciphertext attacks of the adversary.

Current paper contains description of algorithms which allows a repetition of chosen walks on the graph $D(n,K)$ and $A(n,K)$. This rout makes the bridges towards discrete logarithm problem for cyclic subgroups of Cremona group.

Preliminaries on Multivariate Cryptography are collected in the section 2 which contains definitions of special multivariate maps. Section 3 is devoted to information on problems of Extremal Graph Theory which leads to discovery of graphs $D(n,F_q)$ and $A(n,q)$. The descriptions of graphs $D(n,K)$ and their connected components together with cryptographical applications are given in section 4. The graph based explicit construction of requested multivariate transformations is given in section 5. It comes together with the decryption of multivariate public key based on graphs $D(n,K)$.

The last section is the conclusion.

II. On multivariate cryptography and special multivariate transformations

Multivariate cryptography (see [7]) is one of the directions of Postquantum Cryptography, which concerns with algorithms resistant to hypothetic attacks conducted by Quantum Computer. The encryption tools of Multivariate Cryptography are nonlinear multivariate transformations of affine space K^n, where K is a finite commutative ring. Nowadays this modern direction of research requires new examples of algorithms with theoretical arguments on their resistance to attacks conducted by ordinary computer (Turing machine) and new tasks for cryptanalists.

Recall, that *Cremona group* $C(K^n)$ is a totality of invertible maps f of affine space K^n over a Commutative ring K into itself, such that the inverse map f^{-1} is also a polynomial one.

Let us refer to the sequence of maps $f(n)$ from $C(K^n)$, $n=1,2,\dots$ as *the family of bounded degree*, if the degree of each transformation is bounded by the finite parameter s.

Assume that a transformation $f=f(n)$ is written in the form: $x_i \to f_i(x_1,x_2,\dots,x_n)$, $i=1,2,\dots,n$, where each $f_i \in K^n$ is determined by the list of their monomial terms with respect to some chosen order.

A family of elements $f(n) \in C(K^n)$, $n>1$ is called stable if each nonidentity multiple iteration of $f(n)$ with itself has the same degree with $f(n)$. Let $|g|$ be the order of $g \in C(K^n)$.

978-83-60810-58-3/$25.00

We say, that $f(n)$ is a family of increasing order if $|f(n)|$ for n.

Let us consider the discrete logarithm problem for a stable family f^n of increasing order. We have to solve the equation $f(n)^y = b(n)$ with respect to an integer unknown y. Notice, that $\deg(f(n)) = \deg(b(n))$. It means, that studies of degrees $(f(n))^k$, $k = 1, 2, \dots$ do not bring us any new information for the task execution. If the order of an element $f(n)$ is growing fast with the growth of n, then discrete logarithm problem can be NP - hard.

We say that a family $f(n) \in C(K^n)$ has an invertible decomposition of speed d if $f(n)$ can be written as a composition of elements $f^1(n), f^2(n), \dots, f^{k(n)}(n)$ and this decomposition will allow us to compute the value of $y = f(x)$ and the re-image of given y in time $k(n)O(n^d)$ (see the authors extended abstract for Central European Conference on Cryptology 2014).

In the case $d = 1$ we say that invertible decomposition is of linear speed. The complexity of computation of the value of each $f^i(n)$ in a given point x is $O(n^d)$. We refer to the family of multivariate maps $h_{n+1} : K^{n+1} \to K^{n+1}$ as a family with logarithmically invertible decomposition of speed t with the initial function $f(x_1, x_2, \dots, x_n)$ if there exists decomposition $h_{n+1} = h_{n+1,1}h_{n+1,2} \dots h_{n+1,k(n)}$ such that the knowledge about it allows us to solve the equation

$h_{n+1}{}^{\alpha}(x_1, x_2, \dots, x_n, f(x_1, x_2, \dots, x_n)) = (b_1, b_2, \dots, b_{n+1})$ for unknowns $\alpha, x_1, x_2, \dots, x_n$ in time $k(n)O(n^t)$.

We say that function $u : Z^+ \to Z^+$ is computationally equivalent to n^s, $s \geq 0$ and write $u(n^s$ if $C_1 n^s \leq u(n) \leq C_2 n^s$ for some positive constants C_1 and C_2.

Examples of stable families $f(n) \in C(K^n)$ of bounded degree and increasing order defined in terms of algebraic graph theory are given in [4], [8], [9], [11]. An example of stable transformations of linear degree and increasing order is proposed in [12] (see also survey [10], [13] and 14] for extra examples).

III. Extremal algebraic graphs corresponding to special families of multivariate maps and their usage in symmetric cryptography

Recall, that the girth is the length of minimal cycle in the simple graph. Studies of maximal size $ex(C_3, C_4, \dots, C_{2m}, v)$ of the simple graph on v vertices without cycles of length $3, 4, \dots, 2m$, i. e. graphs of girth $> 2m$, form an important direction of Extremal Graph Theory (see [15]).

As it follows from famous the Even Circuit Theorem by P. Erdős we have inequality

$$ex(C_3, C_4, \dots, C_{2m}, v) \leq cv^{1+1/n},$$

where c is a certain constant. The bound is known to be sharp only for $n = 4, 6, 10$. The first general lower bounds of kind $ex(v, C_3, C_4, \dots C_n) = \Omega(v^{1+c/n})$, where c is some constant $< 1/2$ were obtained in the 50th by Erdős via studies of *families of graphs of large girth*, i.e. infinite families of simple regular graphs Γ_i of degree k_i and order v_i such that $g(\Gamma_i) \geq c\log_{k_i} v_i$, where c is the independent of i constant. Erdős proved the existence of such a family with arbitrary large but bounded degree $k_i = k$ with $c = 1/4$ by his famous probabilistic method.

First two explicit families of regular simple graphs of large girth with unbounded girth and arbitrarily large k appeared in 90th: the family $X(p, q)$ of Cayley graphs for $PSL_2(p)$, where p and q are primes, which has been defined by G. Margulis [10] and investigated by A. Lubotzky, Sarnak and Phillips [17] and the family of algebraic graphs $CD(n, q)$ [18]. Graphs $CD(n, q)$ appear as connected components of graphs $D(n, q)$ defined via a system of quadratic equations [19]. The best known lower bound for $d \neq 2, 3, 5$ has been deduced from the existence of above mentioned families of graphs $ex(v, C_3, C_4, \dots, C_{2d}) \geq cv^{1+2/(3d-3+e)}$ where $e = 0$ if d is odd, and $e = 1$ if d is even.

Recall, that family of regular graphs Γ_i of degree k_i and increasing order v_i is a *family of graphs of small world* if $\mathrm{diam}(\Gamma_i) \leq c\log_{\mathrm{k_i}}(v_i)$ for some independent constant c, $c > 0$, where $\mathrm{diam}(\Gamma_i)$ is a diameter of graph G_i. The graphs $X(p.q)$ form a unique known family of large girth which is a family of small world graphs at the same time. There is a conjecture known since 1995 that family of graphs $CD(n, q)$ for odd q is an other example of such kind. Currently, it is proved that the diameter of $CD(n, q)$ is bounded from above by polynomial function $d(n)$, which does not dependent from q. Expanding properties of $X(p, q)$ and $D(n, q)$ can be used in Coding Theory (magnifiers, superconcentrators, etc). The absence of short cycles and high girth property of both families can be used for the construction of LDPC codes [20]. This class of error correcting codes is an important tool of security for satellite communications. The usage of $CD(n, q)$ as Tanner graphs producing LDPC codes leads to better properties of corresponding codes in the comparison to the usage of Cayley - Ramanujan graphs (see [21]).

Both families $X(p, q)$ and $CD(n, q)$ consist of edge transitive graphs. Their expansion properties and the property to be graphs of large girth also hold for random graphs, which have no automorphisms at all. To make better deterministic approximation of random graph we can look at regular expanding graphs of large girth without edge transitive automorphism group.

Below We consider an optimization problem for simple graphs which is similar to the problem of finding maximal size for graph on v vertices with the girth $\geq d$.

Let us refer to the minimal length of a cycle, through the vertex of the given vertex of the simple graph Γ as a *cycle indicator of the vertex*. The *cycle indicator of the graph* $\mathrm{Cind}(\Gamma)$ will be defined as a maximal cycle indicator of its vertices. Regular graph will be called a *cycle irregular graph* if its indicator differs from the girth (the length of minimal cycle). The solution of the optimization problem of computation of maximal size $e = e(v, d)$ of the graph of an order v with the size greater than d, $d > 2$ has been found very recently.

It turns out that

$$e(v,d) \Leftrightarrow O(v^{1+[2/d]})$$

and this bound is always sharp (see [22] or [23] and further references).

We refer to the family of regular simple graphs Γ_i of degree k_i and order v_i as a *family of graphs of large cycle indicator*, if

$$\mathrm{Cind}(\Gamma_i) \geq c\log_{\mathrm{k_i}}(v_i)$$

for some independent constant c, $c > 0$. We refer to the maximal value of c satisfying the above inequality as *speed of growth* of the cycle indicator for a family of graphs Γ_i. As it follows from the written above evaluation of $e(v,d)$ the speed of growth of the cycle indicator for the family of graphs of constant but arbitrarily large degree is bounded above by 2.

We refer to such a family as a *family of cyclically irregular graphs of large cycle indicator* if almost all graphs from the family are cycle irregular graphs.

The following theorem was proved in [23]:

There is a family of almost Ramanujan cyclically irregular graphs of large cycle indicator with the speed of cycle indicator 2, which is a family of graphs of small word graphs.

The explicit construction of the family $A(n,q)$ like in previous statement is given in [22], [23]. Notice, that members of the family of cyclically irregular graphs are not edge transitive graphs. The LDPC codes related to new families are presented in [24], computer simulations demonstrate essential advantages of new codes in comparison to those related to $CD(n,q)$ and $D(n,q)$.

A. On the stream ciphers corresponding to special families of multivariate maps

Graphs $D(n,q)$, $A(n,q)$ and $CD(n,q)$ have been used in symmetric cryptography together with their natural analogs $D(n,K)$, $A(n,K)$ and $CD(n,K)$ over general finite commutative rings K since 1998 (see [1]). The theory of directed graphs and language of dynamical system have been very useful for studies of public key and private key algorithms based on graphs $D(n,K)$, $CD(n,K)$ and $A(n,K)$ (see [10], [25], and further references).

There are several implementations of symmetric algorithms for cases of fields (starting from [7]) and arithmetical rings ([19], in particular). Some comparison of public keys based on $D(n,K)$ and $A(n,K)$ are considered in [21].

The general scheme is the following one. We can use a family of elements $f(n)$ with invertible decomposition of speed d of increasing order for purposes of symmetric cryptography. We assume that the variety K^n is a plainspace of the encryption algorithm, the list of $(f(n,i)$, $i = 1,2,\ldots,k(n)$, is a password. Then the computation of the value c of encryption function $f(n,1)f(n,2)\ldots f(n,k(n))$ in the given plaintext $\mathrm{p} \in K^n$ and the reimage of the ciphertext c require time $O(n^d)$. Usually the parameter $k(n)$ can be chosen free. In fact, in practical cases $k(n)$ is either constant or linear function in variable n (see surveys [20]. [23], [25] on the use graph based multivariate functions as symmetric encryption functions). To hide the graph nature of $f(n)$ correspondents (Alice and Bob) can create a new encryption map $h(n)$ as a conjugation of $f(n)$ with special invertible affine transformation $\tau = \tau(n)$ (degree equals 1) of K^n. In case of private keys both correspondents know the invertible decompositions and family $\tau(n)$ of affine transformation as part of the key.

IV. On the explicit constructions

A. Description of graphs $A(n,K)$

The graph $A(n,K)$, where K is a finite commutative ring, is defined by the following way. This is a bipartite graph with the point set $P = \{x_1,x_2,\ldots,x_n)|x_i \in K\} = K^n$ and the line set $L = \{[y_1,y_2,\ldots,y_n]|y_i \in K|\} = K^n$ and such that a point $\mathrm{x} = (x_1,x_2,\ldots,x_n)$ is incident to a line $\mathrm{y} = [y_1,y_2,\ldots,y_n]$ if and only if equations $x_i - y_i = y_1x_1$ hold for even i and relations $x_j - y_j = x_1y_j$ hold for an odd j, $j \geq 3$. We identify such an incidence relation with the corresponding bipartite graph $I = A(n,K)$. We refer to the first coordinate $x_1 = \rho(\mathrm{x})$ of a point x and the first coordinate $y_1 = \rho(\mathrm{y})$ of a line y of the line as the colour of the vertex (point or line). The following property holds for the graph: there exists a unique neighbour $N_t(v)$ of a given vertex v of a given colour $t \in K$.

As it follows from the definition the projective limit of $A(n,K)$, $n \to \infty$ is well defined. The points $\mathrm{p} = (p_1,p_2,\ldots,p_n,\ldots)$ and lines $\mathrm{l} = [l_1,l_2,\ldots,l_n,\ldots]$ are tuples with finite number of nonzero coordinates. A point and a line are incident when infinite number of equations $p_2 - y_l = l_1p_1, p_3 - l_3 = p_1l_2,\ldots$ hold.

B. Description of graphs $D(n.K)$ and their connected components

We define the family of graphs $D(k,K)$, where $k > 2$ is positive integer and K is a commutative ring, such graphs have been considered in [15] for the case $K = F_q$.

Let P_D and L_D be two copies of Cartesian power K^N, where K is the commutative ring and N is the set of positive integer numbers. Elements of P_D will be called *points* and those of L_D *lines*.

To distinguish points from lines we use parentheses and brackets. If $x \in V$, then $(x) \in P_D$ and $[x] \in L_D$. It will be also advantageous to adopt the notation for co-ordinates of points and lines introduced in [30] for the case of general commutative ring K:

$$(p) = (p_{0,1},p_{1,1},p_{1,2},p_{2,1},p_{2,2},p'_{2,2},p_{2,3},\ldots,$$

$$p_{i,i},p'_{i,i},p_{i,i+1},p_{i+1,i},\ldots),$$

$$[l] = [l_{1,0},l_{1,1},l_{1,2},l_{2,1},l_{2,2},l'_{2,2},l_{2,3},\ldots,$$

$$l_{i,i},l'_{i,i},l_{i,i+1},l_{i+1,i},\ldots].$$

The elements of P and L can be thought as infinite ordered tuples of elements from K, such that only finite number of components are different from zero.

Now, we introduce a linguistic incidence structure (P_D, L_D, I_D) defined by infinite system of equations as follows. We say that the point (p) is incident with the line $[l]$, and we write $(p)I[l]$, if the following relations between their co-ordinates hold:

$$\begin{aligned} l_{i,i} - p_{i,i} &= l_{1,0}p_{i-1,i} \\ l'_{i,i} - p'_{i,i} &= l_{i,i-1}p_{0,1} \\ l_{i,i+1} - p_{i,i+1} &= l_{i,i}p_{0,1} \\ l_{i+1,i} - p_{i+1,i} &= l_{1,0}p'_{i,i} \end{aligned} \tag{6}$$

(These four relations are defined for $i \geq 1$, $p'_{1,1} = p_{1,1}$, $l'_{1,1} = l_{1,1}$). The incidence structure (P_D, L_D, I_D) we denote as $D(K)$. Now we speak of the *incidence graph* of (P_D, L_D, I_D), which has the vertex set $P_D \cup L_D$ and edge set consisting of all pairs $\{(p), [l]\}$ for which $(p)I[l]$.

For each positive integer $k \geq 2$ we obtain a symplectic quotient $(P_{D,k}, L_{D,k}, I_{D,k})$ as follows. Firstly, $P_{D,k}$ and $L_{D,k}$ are obtained from P_D and L_D, respectively, by simply projecting each vector into its k initial coordinates. The incidence $I_{D,k}$ is then defined by imposing the first $k-1$ incidence relations and ignoring all others. The incidence graph corresponding to the structure $(P_{D,k}, L_{D,k}, I_{D,k})$ is denoted by $D(k, K)$.

To facilitate notation in the future results on "connectivity invariants", it will be convenient for us to define $p_{-1,0} = l_{0,-1} = p_{1,0} = l_{0,1} = 0$, $p_{0,0} = l_{0,0} = -1$, $p'_{0,0} = l'_{0,0} = -1$, $p'_{1,1} = p_{1,1}, l'_{1,1} = l_{1,1}$) and to assume that our equations are defined for $i \geq 0$.

Notice, that for $i = 0$, the written above four conditions are satisfied by every point and line, and for $i = 1$ the first two equations coincide and give $l_{1,1} - p_{1,1} = l_{1,0}p_{0,1}$.

Let $k \geq 6$, $t = [(k+2)/4]$, and let $u = (u_\alpha, u_{11}, \cdots, u_{tt}, u'_{tt}, u_{t,t+1}, u_{t+1,t}, \cdots)$ be a vertex of $\mathrm{D}(k, K)$ ($\alpha \in \{(1,0),(0,1)\}$, it does not matter whether u is a point or a line). For every r, $2 \leq r \leq t$, let

$$a_r = a_r(u) = \sum_{i=0,r} (u_{ii}u'_{r-i,r-i} - u_{i,i+1}u_{r-i,r-i-1}),$$

and $a = a(u) = (a_2, a_3, \cdots, a_t)$. Similarly, we assume that $a = a(u) = (a_2, a_3, \cdots, a_t, \dots)$ for the vertex u of infinite graph $D(K)$.

Proposition 4.1: Let u and v be vertices from the same component of $D(k, K)$. Then $a(u) = a(v)$. Moreover, for any $t-1$ field elements $x_i \in F_q$, $2 \leq t \leq [(k+2)/4]$, there exists a vertex v of $D(k, K)$ for which
$a(v) = (x_2, \dots, x_t) = (x)$.

V. On flag systems of graphs $A(n, K)$ and $D(n.K)$, walks on them and multivariate maps

Graphs $D(n, K)$ and $A(n, K)$ have somme common properties. We refer to the first coordinate $x_{1,0} = \rho(\mathrm{x})$ ($x_1 = \rho(\mathrm{x})$ of a point x from graph $D(n, K)$ (graph $A(n, K)$, respectively) and the first coordinate $y_{1,0} = \rho(\mathrm{y})$ ($y_1 = \rho(\mathrm{y})$) of a line y as the colour of the vertex (point or line). The following property holds for the graph: there exists a unique neighbour $N_t(v)$ of a given vertex v of a given colour $t \in K$.

A flag of the incidence system $D(n, K)$ or $D(K)$ ($A(n, K)$ or $A(K)$) is an unordered pair $\{(\mathrm{x}), [\mathrm{y}]\}$ such that $(\mathrm{x})I[\mathrm{y}]$. Obviously, the totalities of flags $FD(n, K)$ or $(FA(n, K))$ of the bipartite flag $D(n, K)$ (or $A(n, K)$, respectively) are isomorphic to the variety K^{n+1}. So, flag $\{(\mathrm{x}), [\mathrm{y}]\}$ of $D(n, K)$ is defined by the tuple $(x_{10}, x_{11}, \dots, y_{01})$. Notice, that $N_{y_1}(\{\mathrm{x}\}) = [\mathrm{y}]$.

We consider an operator $NP_\alpha(\{(\mathrm{x}), [\mathrm{y}]\})$, $\alpha \in K$ mapping flag $\{(\mathrm{x}), [\mathrm{y}]\}$ of the incidence structure $G(n, K)$ (where G is D or A)) into its image $\{(\mathrm{x}'), [\mathrm{y}]\}$, where $\mathrm{x}') = N_\alpha([\mathrm{y}])$.

Similarly, an operator $NL_\alpha(\{(\mathrm{x}), [\mathrm{y}]\})$ maps $\{(\mathrm{x}), [\mathrm{y}]\}$ into $\{(\mathrm{x}), N_\alpha(\mathrm{x})\})$.

Let $\alpha_1, \alpha_2, \dots, \alpha_k$ and $\beta_1, \beta_2, \dots, \beta_k$ be chosen sequences of elements from the commutative ring K. The composition

$$E = NP_{\alpha_1}NL_{\beta_1}NP_{\alpha_2}NL_{\beta_2}\dots NP_{\alpha_k}NL_{\beta_k}$$

transforms flag $\{(\mathrm{x}), [\mathrm{y}]\}$ into the new flag $\{(\mathrm{x}'), [\mathrm{y}']\}$. The process of recurrent computations of $E(\{(\mathrm{x}), [\mathrm{y}]\} = \{(\mathrm{x}'), [\mathrm{y}']\}$ corresponds to the walk in a graph $G(n, K)$ with the original vertex (x) and the final point (x'). Notice, that $[\mathrm{y}'] = N_\alpha(\mathrm{x}')$.

Let us assume now that we have two finite families of polynomials of $K[z_1, z_2] : \phi_1(z_1, z_2), \phi_2(z_1, z_2), \dots, \phi_{k+1}(z_1, z_2)$ and $\psi_1(z_1, z_2), \psi_2(z_1, z_2), \dots, \psi_k(z_1, z_2)$. We assume that their density is restricted by independent constant d and their degree is bounded by the linear function $\alpha n + \beta$.

The transformation $\tilde{E}$ shifts a flag $\{(\mathrm{x}), [\mathrm{y}]\}$ into its image for the map

$$\begin{aligned} &NP_{\phi_1(x_1,y_1)}NL_{\psi_1(x_1,y_1)}NP_{\phi_2(x_1,y_1)}NL_{\psi_2(x_1,y_1)}\cdots \\ &\dots NP_{\phi_k(x_1,y_1)}NL_{\psi_k(x_1,y_1)}. \end{aligned}$$

Additionally, we assume that the system of equations $\phi_k(z_1, z_2) = a$, $\psi_k(z_1, z_2) = b$ has exactly one solution independently from the choice of a and b (boundary requirement). The written above condition insure that the reimage of $\{\mathrm{x}', [\mathrm{y}']$ for $\tilde{E}$ is uniquely determined. Really, parameters x_1 and y_1 are determined by the system of equations.

It allows us to compute each expression of kind $\phi_i(x_1, y_1)$ and $\psi_j(x_1, y_1)$ and to obtain the reverse walk in the graph with the origin x' and final point x. So, we get the original flag $(\mathrm{x}), [\mathrm{y}]$ with $[\mathrm{y}] = N_{y_1}(\mathrm{x})$. The code of our flag is $(x_1, x_2, \dots, x_n, y_1)$.

Let $f = f_n$ be the transformation of affine space K^{n+1} into itself which maps flag $(x_1, x_2, \dots, x_n, y_1)$ into the image for $\tilde{E}$ defined by the family of bivariate polynomials from $K[z_1, z_2]$. Assume that f_n is written in a standard form $x_i \to f_i(x_1, x_2, \dots, x_n, y_1)$, $i = 1, 2, \dots, n$, $y_1 = f_{n+1}(x_1, x_2, \dots, x_n, y_1)$.

Let ${g_n}^i : K^{n+1} \to K^{n+1}$ be the transformation moving $\mathrm{z} = (z_1, z_2, \dots, z_n, u_1)$ into $NP_{\phi_{i_{z_1,u_1}}}(\mathrm{z})$ and ${h_n}^j$ be the transformation moving z into $NL_{\psi_{j_{z_1,u_1}}}(\mathrm{z})$. Obviously, $f = {g_n}^1{h_n}^2{g_n}^2{h_n}^2\dots{g_n}^k{h_n}^k$ is the invertible decomposition of f of speed $O(n)$. Notice, that generally speaking it is not true that each ${g_n}^i$ or ${h_n}^i$ is invertible. The following statement is a

direct corollary of results [3] in the case $G(n,K)=D(n,K)$, and results of [4] in the case of $G(n,k)=A(n,K)$.

Theorem 5.1: The $G(n,K)$ graph based transformations $f_n : K^{n+1} \to K^{n+1}$ defined above for $\phi_j(z_1,z_2)=z_1+a_j$ and $\psi_j(z_1,z_2)=z_2+b_j$ where $a_j,b_j \in K$, $j=1,2,\dots,k$ are stable cubical maps.

It means that we always have $O(n^4)$ monomial terms for the map f_n. Notice that f_n is given by its invertible decomposition. The following statement is a direct corollary from the theorem.

Proposition 5.1: Let us consider the specialization $\tilde{f}_n$ of f_n given by relations $y_{0,1}=h(x_{1,0})$ ($y_1=h(x_1$ in case of graphs $A(n,K)$, respectively), where $h(x) \in K[x]$ is a polynomial expression of degree t, such that equation of kind $h(x_{1,0})=b$, $b \in K$ ($h(x_1)=b$) has no more than one solution. Then degree of $\tilde{f}_n$ is bounded by t^3.

Remark 5.1:

We can change variables $x_{1,0}$ and x_1 of the proposition for y_{01} and y_1, respectively.

Recall, that M is a multiplicative subset of commutative ring K if it is closed under multiplication and does not contain zero. Let us consider the following special choice of coefficients a_j and b_j. The following statement is proved in [23] (see also [13], [14]).

Theorem 5.2: Let $f_n : K^{n+1} \to K^{n+1}$ be $G(n,K)$ graph based transformation $f_n : K^{n+1} \to K^{n+1}$ defined for $\phi_j(z_1,z_2)=z_1+a_j$ and $\psi_j(z_1,z_2)=z_2+b_j$, where $a_j,b_j \in K$, $j=1,2,\dots,k$ in theorem 1.

Let M be a multiplicative set of K and $a_1,b_1 \in M$, $a_{i+1}-a_i \in M$, $b_{i+1}-b_i \in M$ for $i=1,2,\dots,k-1$. Then the order of a transformation f_n is going to infinity with the growth of n.

Remark 5.2: In the case of graph $D(n,K)$ we can change polynomial $h(x_{1,0}$ for the $h(x_{1,0},a_2(\mathrm{x}),a_3(\mathrm{x}),\dots,a_t(\mathrm{x}))$, where $h(z_1,z_2,\dots z_t) \in K[(z_1,z_2,\dots z_t]$, $t=[(n+2)/4]$.

We can look at f_n as function with invertible decomposition with initial relation $y_{0,1}=h(x_{1,0}$ (case of $D(n,K)$) or $y_1=h(x_1)$ (case of $A(n,K)$. Really, invertible decomposition of f_n allows to solve

$(f_n)^s(x_{1,0},h(x_{1,0},x_{1,1},\dots,)) = (c_{1,0},c_{0,1},c_{1,1},c_{2,1},\dots)$ or

$(f_n)^s(x_1,h(x_1,x_2,\dots,x_n)) = (c_1,c_1',c_2,\dots,c_n)$ can be solved fast in some special simple cases.

For simplicity of writing we assume that $G(n,K)=D(n,K)$. Let us consider the system of equation $(*)$: $x_{10}+\alpha_k s=c_{1,0}$, $h(x_{1,0})+\beta_k s=c_{0,1}$

We can eliminate parameter s: $\beta_k x_{1,0}+\alpha_k\beta_k s=\beta_k c_{1,0}$

$h(x_{1,0})\alpha_k+\alpha_k\beta_k s=c_{1,0}\alpha_k$.

So, we get an equation of kind $c_{0,1}\alpha_k-\beta_k c_{1,0}=h(x_{1,0})\alpha_k x_{1,0}\beta_k$ (*)

Let us assume that $h(x_{1,0})\alpha_k x_{1,0}\beta_k=c$ has not more than one solution for each $c \in K$.

Under this condition we can solve (*) for x_1. So, if α_k or β_k differs from 0 we can find parameter s.

Assume that characteristics of ring K is a large prime p. Let us consider the following two simple cases:

(a) $\alpha_k=0$ but β_k is a regular ring element. It is clear that in this case $x_{1,0}$ is known and we can find parameter s with arbitrarily chosen function $h(x)$.

(b) $\beta_k=0$ and equation $h(x_{1,0})=c$ has no more than one solution. In this case one can find $x_{1,0}$ and find parameter s from the first equation.

We say that multivariate map $g_n : K^{n+1}=K^{n+1}$ is symmetrical if $\deg(g_n)=\deg(g_n)^{-1}$. Obviously, each stable transformation is symmetrical. It is clear that in the case (a) we get a stable transformation of K^n into itself. In case of $\deg(g_n)\neq\deg(g_n)^{-1}$ we refer to g_n as assymetrical map.

The following cryptosystem can be used.

Alice chooses a function $h(x_{1,0},a_2(\mathrm{x}),a_3(\mathrm{x}),\dots,a_t(\mathrm{x}))$ of finite degree t and invertible affine transformation: $\tau_1 : K^n \to K^n$, which sends x onto xA+b. Assume that it will be extended till K^{n+1} via the rule $\tau_1 : z \to az+l(\tau_1(\mathrm{x}))=z'$, where l is some linear function from x. Let τ be an expanded linear transformation.

Alice takes the symbolic tuple $(x_{1,0},x_{1,1},\dots,z)$ applies τ and gets the vector $u_{1,0},u_{1,1},\dots,z'=\mathrm{u}$. She will treat this tuple as a flag from $FD(n,K)$.

She writes the equation $z'=h(x_1,x_2)$ and rewrites it in the form $z=h'(x_1,x_2)$.

Alice choses the pseudorandom strings $\alpha_1,\alpha_2,\dots\alpha_k$ and $\beta_1,\beta_2,\dots,\beta_k$ of ring elements.

She generates defined above transformation $f_n : K^{n+1} \to K^{n+1}$. Alice computes symbolically $f_n(\mathrm{u})=\mathrm{w}$ and applies τ^{-1} to w.

She forms a stable cubical transformations $E=g_n=\tau f_n\tau^{-1}$ and writes it in standard form

$$\begin{aligned} x_{10} &\to x'_{1,0}=g_{1,0}(x_{1,0},x_{1,1},\dots,z)\\ x_{11} &\to x'_{1,1}=g_{1,1}(x_{1,0},x_{1,1},\dots,z)\\ &\vdots\\ z &\to z'=g_{0,1}(x_{1,0},x_{1,1},\dots,z)\end{aligned}$$

In the case of the first n rules Alice uses the specialisation $z=h'(\mathrm{x})$ and writes $\tilde{g}'_{1,0}(\mathrm{x})=g_{1,0}(h'(\mathrm{x},x_{1,0},x_{1,1}\dots)$, $\tilde{g}'_{1,1}(\mathrm{x})=g_{1,1}(h'(\mathrm{x}),x_{1,0},x_{1,1},\dots)$, $\dots$, in a standard form. The specialisetion gives us a restriction E' of our encryption map on the point set isomorphic to K^n.

Bob gets these n rules from Alice together with initial condition $z=h'(x_{1,0},x_{1,1},\dots)$.

He takes his plaintext $(\mathrm{x})=(p_{1,0},p_{1,1},\dots)$ and applies the restricted map E' iteratively s times.

Thus, he gets consecutively $E'^i(\mathrm{p})$, $i=1,2,\dots,s$ and computes recursively

$$\begin{aligned} z_1 &= g_{0,1}(p_{1,0},p_{1,1},\dots,h'(\mathrm{p}),\\ z_2 &= g_{0,1}(E(\mathrm{p},z_1)),\\ &\vdots\\ z_s &= g_{0,1}(E^{s-1}\mathrm{p},z_{s-1}).\end{aligned}$$

He sends Alice his expanded ciphertext as a pair $\mathrm{c}=E'^s(\mathrm{p})$ and parameter z_s.

For the decryption Alice applies transformation τ to the c concateneted with z_s and gets c_1. She computes $E^{-1}(\mathrm{c}_1) = c_2$. Computation $\tau^{-1}(\mathrm{c}_2)$ gives her the plaintext p.

Let us consider some obvious properties of defined above cryptosystem in special cases (a) and (b).

(a) We can see that our encryption is of symmetrical degree. Let $\mathrm{deq}h = t$, then our map f_n has a degree bounded by t^3. If parameter t is a constant then the map E' is computable in polynomial time. Notice, that linearisation attacks are possible, they allow to compute E'^{-1}. This fact is not yet a breaking of the system, because E' is a stable map which order is growing with the growth of parameter n.

Thus, finding the solution for $E'^s = H(\mathrm{x})$ can be a difficult task. The discrete logarithm problem for cyclic subgroup of Cremona group of increasing order appears there. Notice, that only one value of $H(\mathrm{x})$ can be given for chosen by Bob parameter s. Algorithm can be used in dynamical mode: every session Alice changes encryption base and every time Bob changes parameter s.

Notice, that $s = s(n)$ can be a function from parameter n. Bob can encrypt for polynomia time $s(n)O(n^{t^3})$. Alice can decrypt because of the logarithmical invertibility of the map.

(b) Let us just consider a simple example

$$h(x_{1,0}, a_2(\mathrm{x}), \ldots, a_t(\mathrm{x})) = (d(a_2(\mathrm{x}), a_3(\mathrm{x}), \ldots, a_t(\mathrm{x}))x_{1,0} + \\ + b(a_2(\mathrm{x}), \ldots, a_t(\mathrm{x}))^r + c(a_2(\mathrm{x}), a_3(\mathrm{x}), \ldots, (a_t\mathrm{x})),$$

d, b, c are multivariate functions, r is odd and equation $x^r = \alpha$ in K has not more than one solution for each parameter α. If we skip degenerate cases, our encryption function E' will be assymetric. It means that even finding the inverse E' can be a hard task in this case.

We presnt here a well known case of the pair (r, K) which satisfies to written above property (see the description of Imai-Matsumoto method in [26]). Let $K = F_{q^n}$ be an extention of the field F_q of characteristic 2. We take r as a parameter of kind $q^\beta + 1$ for some parameter β, such that the greatest common divisor of $q^\beta + 1$ and $q^n - 1$ is 1. Then map $x \to x^r$ is one to one correspondence and equation $x^r = \alpha$ has a unique solution.

VI. Conclusions

Known methods of symmetric encryption according to chosen walks on flags of bipartite graphs $A(n, K)$ and $D(n, K)$ use special colouring of their points and lines. The increasing girth and good expansion properties of these graphs lead to good mixing properties of the stream cipher based on stable transformation. The weakness of such method is an option of cubical linearisation attacks based on the fact that decryption map is also cubical (complexity of the attack is $O(n^{10})$, so its costly, but possible. There were several implementations of such algorithms for practical use in academic networks and ORACLE based university management systems for various cases of fields and rings: [2], [27], [28] devoted to ciphers used in The University of South Pacific (Fiji), [30], [31], [35] discussed algorithms used at Sultan Qaboos University (Oman), [31], [32], [35] were used in University of Maria Curie - Sklodovska (Poland), algorithm of [29] and [34] were used in teaching process of Kiev Mohyla Academy (Ukraine) and University of British Columbia (Canada), respectively.

Private key algorithm, presented in this paper allows to modificate discussed above programs with essential increase of resistence to linearisation attack without damage of theoretical speed ($O(n)$ in the case of keys of constant length and $O(n^2)$ for passwords of length $O(n)$). We can create encryption maps of large symmetric degree or assymetrical maps with inverses of high degree.

In a public mode we introduce the multivariate cryptosystems such that their security is connected with discrete logarithm problem for large cyclic subgroups of Cremona group. We hope that a new class of multivariate cryptosystems can be an interesting objects for cryptanalitical studies.

References

[1] Ustimenko V., *Coordinatisation of Trees and their Quotients*, In the "Voronoj's Impact on Modern Science", Kiev, Institute of Mathematics, 1998, vol. 2, 125-152.

[2] Ustimenko V., *CRYPTIM: Graphs as Tools for Symmetric Encryption*, Lecture Notes in Computer Science, Springer, v. 2227, 278-287 (2001).

[3] A. Wróblewska, *On some properties of graph based public keys*, Albanian Journal of Mathematics, Volume 2, Number 3, 2008, 229-234, NATO Advanced Studies Institute: "New challenges in digital communications".

[4] Vasyl Ustimenko, Aneta Wróblevska, *On the key exchange with nonlinear polynomial maps of degree 4*, Proceedings of the conference "Applications of Computer Algebra", Vlora, Albanian Journal of Mathematics, Special Issue, December, 2010, vol .4 n 4, 161-170.

[5] Ustimenko V., Graphs with special arcs and cryptography, Acta Applicandae Mathematicae (Kluwer) 2002, 74,117-153.

[6] Ustimenko V. *Maximality of affine group and hidden graph cryptosystems*// J. Algebra Discrete Math. -2005 ., No 1,-P. 133–150.

[7] Ding J., Gower J. E., Schmidt D. S., *Multivariate Public Key Cryptosystems*, 260. Springer, Advances in Information Security, v. 25, (2006).

[8] Vasyl Ustimenko, *On the graph based cryptography and symbolic computations*, Serdica Journal of Computing, Proceedings of International Conference on Application of Computer Algebra, ACA-2006, Varna, N1 (2007).

[9] V. Ustimenko, *On the extremal graph theory for directed graphs and its cryptographical applications*, In: T. Shaska, W.C. Huffman, D. Joener and V.Ustimenko, Advances in Coding Theory and Cryptography, Series on Coding and Cryptology, vol. 3, 181-200 (2007).

[10] V. A. Ustimenko, *On the cryptographical properties of extreme algebraic graphs*, in Algebraic Aspects of Digital Communications, IOS Press (Lectures of Advanced NATO Institute, NATO Science for Peace and Security Series - D: Information and Communication Security, Volume 24, July 2009, 296 pp.

[11] V. Ustimenko, A. Wróblewska, On the key exchange with nonlinear polynomial maps of stable degree, Annalles UMCS Informatica AI X1, 2 (2011), 81-93.

[12] Vasyl Ustimenko, Aneta Wróblewska, *On some algebraic aspects of data security in cloud computing*, Proceedings of International conference "Applications of Computer Algebra", Malaga, 2013, p. 144-147.

[13] V. A. Ustimenko, U. Romańczuk, *On Dynamical Systems of Large Girth or Cycle Indicator and their applications to Multivariate Cryptography*, in "Artificial Intelligence, Evolutionary Computing and Metaheuristics ", In the footsteps of Alan Turing Series: Studies in Computational Intelligence, Volume 427/January 2013, 257-285.

[14] V. A. Ustimenko, U. Romańczuk *On Extremal Graph Theory, Explicit Algebraic Constructions of Extremal Graphs and Corresponding Turing Encryption Machines*, in "Artificial Intelligence, Evolutionary Computing and Metaheuristics ", In the footsteps of Alan Turing Series: Studies in Computational Intelligence, Vol. 427, Springer, January , 2013, 237-256.

[15] B. Bollobás, *Extremal Graph Theory*, Academic Press, London, 1978.

[16] G. Margulis, *Explicit group-theoretical constructions of combinatorial schemes and their application to desighn of expanders and concentrators*, Probl. Peredachi Informatsii., 24, N1, 51-60. English translation publ. Journal of Problems of Information transmission (1988), 39-46.
[17] A. Lubotsky, R. Philips, P. Sarnak, *Ramanujan graphs*, J. Comb. Theory., 115, N 2., (1989), 62-89.
[18] F. Lazebnik, V. A. Ustimenko and A. J. Woldar, *A New Series of Dense Graphs of High Girth*, Bull (New Series) of AMS, v.32, N1, (1995), 73-79.
[19] F. Lazebnik, V. Ustimenko, *Explicit construction of graphs with arbitrary large girth and of large size*, Discrete Applied Mathematics 60 (1995), 275-284.
[20] P. Guinand, J. Lodge, *Tanner type codes arising from large girth graphs*, Canadian Workshop on Information Theory CWIT '97, Toronto, Ontario, Canada (June 3-6 1997):5–7.
[21] D. MacKay and M. Postol, *Weakness of Margulis and Ramanujan - Margulis Low Dencity Parity Check Codes*, Electronic Notes in Theoretical Computer Science, 74 (2003), 8pp.
[22] V. Ustimenko, *On some optimisation problems for graphs and multivariate cryptography* (in Russian), In Topics in Graph Theory: A tribute to A.A. and T. E. Zykova on the ocassion of A. A. Zykov birthday, pp 15-25, 2013, www.math.uiuc.edu/kostochka.
[23] Ustimenko V. A.: On extremal graph theory and symbolic computations, Dopovidi National Academy of Sci of Ukraine, N2 (in Russian), 42-49 (2013)
[24] M. Polak, V. A. Ustimenko, *On LDPC Codes Corresponding to Infinite Family of Graphs A(n,K)*, Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS), CANA, Wroclaw, September, 2012 , pp 11-23.
[25] V. Ustimenko, *Linguistic Dynamical Systems, Graphs of Large Girth and Cryptography*, Journal of Mathematical Sciences, Springer, vol.140, N3 (2007) pp. 412-434.
[26] N. Koblitz, *Algebraic Cryptography*. Springer, 1998.
[27] Y. Khmelevsky, V. Ustimenko, *Walks on graphs as symmetric and asymmetric tools for encryption*, South Pacific Journal of Natural Studies, 2002, vol. 20, 23-41.
[28] Y. Khmelevsky, V. Ustimenko, *Practical aspects of the Informational Systems reengineering*, The South Pacific Journal of Natural Science, volume 21, 2003, p.75-21.
[29] V. Ustimenko, A. Tousene, *CRYPTALL - a System to Encrypt All types of Data*, Notices of Kiev - Mohyla Academy , v. 23, 2004,pp 12-15.
[30] A. Touzene, V. Ustimenko, *Graph Based Private KeyCrypto System*, International Journal on Computer Research, Nova Science Publisher, volume 13 (2006), issue 4, 12p.
[31] A. Touzene, V. Ustimenko, *Private and Public Key Systems Using Graphs of High Girth*, In "Cryptography Research Perspectives", Nova Publishers, Ronald E. Chen (the editor), 2008, pp.205-216
[32] J. Kotorowicz, V. Ustimenko, *On the implementation of cryptoalgorithms based on algebraic graphs over some commutative rings*, Condenced Matters Physics, 2006, 11 (no. 2(54)) (2008) 347–360.
[33] V, Ustimenko, S. Kotorowicz *On the properties of Stream Ciphers Based on Extremal Directed graphs*, In "Cryptography Research Perspectives", Nova Publishers, Ronald E. Chen (the editor), 2008, 12pp.
[34] Y. Khmelevsky, Gaetan Hains, E. Ozan, Chris Kluka, D. Syrotovsky, V. Ustimenko, *International Cooperation in SW Engineering Research Projects*, Proceedings of Western Canadien Conference on Computing Education, University of Northen British Columbia, Prince George BC, May 6-7, 2011, 14pp.
[35] A. Touzene, V. Ustimenko, Marwa AlRaisi, Imene Boudelioua, *Performance of Algebraic Graphs Based Stream-Ciphers Using Large Finite Fields*, Annalles UMCS Informatica AI X1, 2 (2011), 81-93.
[36] M.Klisowski, V. A. Ustimenko, *On the Comparison of Cryptographical Properties of Two Different Families of Graphs with Large Cycle Indicator*, Mathematics in Computer Science, 2012, Volume 6, Number 2, Pages 181-198.

7th International Symposium on Multimedia Applications and Processing

ORGANIZED BY

SOFTWARE Engineering Department, Faculty of Automation, Computers and Electronics, University of Craiova, Romania "Multimedia Applications Development" Research Centre

Background and Goals

Multimedia information has become ubiquitous on the web, creating new challenges for indexing, access, search and retrieval. Recent advances in pervasive computers, networks, telecommunications, and information technology, along with the proliferation of multimedia mobile devices - such as laptops, iPods, personal digital assistants (PDA), and cellular telephones - have stimulated the development of intelligent pervasive multimedia applications. These key technologies are creating a multimedia revolution that will have significant impact across a wide spectrum of consumer, business, healthcare, educational and governmental domains. Yet many challenges remain, especially when it comes to efficiently indexing, mining, querying, searching, retrieving, displaying and interacting with multimedia data.

The Multimedia - Processing and Applications 2014 (MMAP 2014) Symposium addresses several themes related to theory and practice within multimedia domain. The enormous interest in multimedia from many activity areas (medicine, entertainment, education) led researchers and industry to make a continuous effort to create new, innovative multimedia algorithms and applications.

As a result the conference goal is to bring together researchers, engineers, developers and practitioners in order to communicate their newest and original contributions. The key objective of the MMAP conference is to gather results from academia and industry partners working in all subfields of multimedia: content design, development, authoring and evaluation, systems/tools oriented research and development. We are also interested in looking at service architectures, protocols, and standards for multimedia communications - including middleware - along with the related security issues, such as secure multimedia information sharing. Finally, we encourage submissions describing work on novel applications that exploit the unique set of advantages offered by multimedia computing techniques, including home-networked entertainment and games. However, innovative contributions that don't exactly fit into these areas will also be considered because they might be of benefit to conference attendees.

Call for Papers

MMAP 2014 is a major forum for researchers and practitioners from academia, industry, and government to present, discuss, and exchange ideas that address real-world problems with real-world solutions.

The MMAP 2014 Symposium welcomes submissions of original papers concerning all aspects of multimedia domain ranging from concepts and theoretical developments to advanced technologies and innovative applications. MMAP 2014 invites original previously unpublished contributions that are not submitted concurrently to a journal or another conference.

Papers acceptance and publication will be judged based on their relevance to the symposium theme, clarity of presentation, originality and accuracy of results and proposed solutions.

Topics

Topics of interest are related to Multimedia Processing and Applications including, but are not limited to the following areas:

- Audio, Image and Video Processing
- Animation, Virtual Reality, 3D and Stereo Imaging
- Multimedia File Systems and Databases: Indexing, Recognition and Retrieval
- Machine Learning, Data Mining, Information Retrieval in Multimedia Applications
- Multimedia in Internet and Web Based Systems:
  - E-Learning, E-Commerce and E-Society Applications
- Human Computer Interaction and Interfaces in Multimedia Applications
- Multimedia in Medical Applications
- Entertainment and games
- Security in Multimedia Applications: Authentication and Watermarking
- Distributed Multimedia Systems
- Network and Operating System Support for Multimedia
- Mobile Network Architecture
- Intelligent Multimedia Network Applications

Steering Committee

Ioannis Pitas, University of Thessaloniki, Greece
Costin Badica, University of Craiova, Romania
Borko Furht, Florida Atlantic University, USA
Harald Kosch, University of Passau, Germany
Vladimir Uskov, Bradley University, USA
Thomas M. Deserno, Aachen University, Germany

Publicity Chair

Badica, Amelia, University of Craiova, Romania
Burlea Schiopoiu, Adriana, University of Craiova, Romania

Organizing Committee

Dumitru Dan Burdescu, University of Craiova, Romania
Costin Badica, University of Craiova, Romania

Marius Brezovan, University of Craiova, Romania
Liana Stanescu, University of Craiova, Romania
Cristian Marian Mihaescu, University of Craiova, Romania

Event Chairs

Brezovan, Marius, University of Craiova
Burdescu, Dumitru Dan, University of Craiova, Romania

Program Committee

Badica, Amelia, University of Craiova, Romania
Böszörmenyi, Laszlo, Klagenfurt University, Austria
Burlea Schiopoiu, Adriana, University of Craiova
Camacho, David, Universidad Autonoma de Madrid, Spain
Cano, Alberto, University of Cordoba, Spain
Cardoso, Jaime S., Universidade do Porto, Portugal
Cretu, Vladimir, Politehnica University of Timisoara, Romania
Debono, Carl James, University of Malta, Malta
Fabijańska, Anna, Lodz University of Technology, Poland - Institute of Applied Computer Science, Poland
Fomichov, Vladimir, National Research University Higher School of Economics, Moscow, Russia., Russia
Giurca, Adrian, Brandenburg University of Technology, Germany
Grosu, Daniel, Wayne State University, United States
Groza, Voicu, University of Ottawa, Canada
Grundspenkis, Janis, Riga Technical University, Latvia
Kabranov, Ognian, Cisco Systems, United States
Kannan, Rajkumar, Bishop Heber College Autonomous, India
Korzhik, Valery, State University of Telecommunications, Russia
Kotenko, Igor, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Science, Russia
Kriksciuniene, Dalia, Vilnius University, Lithuania
Lamas, David, Tallin University, Estonia
Lau, Rynson, City University of Hong Kong, Hong Kong S.A.R., China
Lloret, Jaime, Polytechnic University of Valencia, Spain
Logofatu, Bogdan, University of Bucharest, Romania
Luna, Jose, University of Cordoba, Spain
Mangioni, Giuseppe, DIEEI - University of Catania, Italy
Mihaescu, Cristian, University of Craiova
Miyata, Hitoshi, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Japan
Mocanu, Mihai, University of Craiova, Romania
Morales-Luna, Guillermo, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Mexico
Ogiela, Marek, AGH University of Science and Technology, Poland
Ohzeki, Kazuo, Shibaura Institute of Technology, Japan
Õunapuu, Enn, Tallinn University of Technology, Estonia
Paltoglou, Georgios, University of Wolverhampton, United Kingdom
Popescu, Dan, CSIRO, Sydney, Australia, Australia
Querini, Marco, Department of Civil Engineering and Computer Science Engineering
Rutkauskiene, Danguole, Kaunas University of Technology
Salem, Abdel-Badeeh M., Ain Shams University, Egypt
Sari, Riri Fitri, University of Indonesia, Indonesia
Smedberg, Asa, Stockholm University, Sweden
Stanescu, Liana, University of Craiova
Tejera, Mario Hernández, University of Las Palmas de Gran Canaria, Spain
Trausan-Matu, Stefan, Politehnica University of Bucharest, Romania
Trzcielinski, Stefan, Poznan University of Technology, Poland
Tsihrintzis, George, University of Piraeus, Greece
Vega-Rodríguez, Miguel A., University of Extremadura, Spain
Velastin, Sergio, Kingston University, United Kingdom
Watanabe, Toyohide, University of Nagoya
Wotawa, Franz, Technische Universitaet Graz, Austria
Zurada, Jacek, University of Louisville, United States

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 641–650
DOI: 10.15439/2014F136
ACSIS, Vol. 2

Gaussian-Based Approach to Subpixel Detection of Blurred and Unsharp Edges

Anna Fabijańska
Lodz University of Technology
Institute of Applied Computer Science
18/22 Stefanowskiego Str., 90-924 Lodz, Poland
Email: anna.fabijanska@p.lodz.pl

***Abstract*—In this paper the problem of edge detection with subpixel accuracy is considered. In particular, the precise detection of significantly blurred edges is regarded. A new method for subpixel edge detection is introduced. The method attempts to reconstruct image gradient function at the edge using the Gaussian function. The results of subpixel edge detection in the artificially created and the real images obtained by the introduced approach are presented and compared with the results of previously proposed methods. In particular, the moment based methods, the gravity center method and the parabola fitting method are considered in the comparison. The presented results prove the robustness of the introduced approach against the averaging and the Gaussian blur. Additionally, the comparison shows, that the introduced approach outperforms the existing state-of-art methods for subpixel edge detection.**

I. Introduction

EDGE detection is the problem of crucial importance in image processing. Edges define location and geometric features of objects present in the scene. Therefore, in a typical vision system, edge detection is performed during low level processing and provides information for operations performed in the following stages, such as quantitative analysis, target recognition or image coding etc.

Recently, the requirements for edge detection accuracy rapidly increase. Satellite remote sensing, telemetry, photogrammetry, medical image analysis, industrial inspection, geometrical measurement and other applications where accuracy is at premium require precision of tenths or even hundredths of pixel.

The traditional, well-established methods for edge detection such as gradient operators (Sobel, Prewitt, Roberts), Canny edge detector, operator LoG etc. all belong to pixel level. Therefore, they are mostly insufficient for practical applications of modern machine vision. Due to their low precision of edge location and extracted wider edges, these approaches more and more often have difficulties in meeting the actual accuracy requirements of vision systems. Therefore, the development of subpixel techniques for edge detection has become one of the hotspots of the current research in image processing. Some work has already been done on this problem. However, the major methods for subpixel edge detection are still to be developed. It should be also underlined that while there has been substantial work performed on the detection of clear and well defined edges at subpixel accuracy, a little has been done on the subpixel edge detection in low contrast images containing blurred, noisy and unsharp edges [1][2].

This paper presents a new method which is a step forward through introducing subpixel analysis into edge detection. In particular, it considers precise edge detection of significantly blurred edges. As the already proposed methods deal mostly with sharp, regular and well defined edges the introduced approach can be considered like a novelty.

This paper is organized as follows. Firstly, in Section 2 background on subpixel edge detection is given. Then, in Section 3 the proposed method for subpixel edge detection is introduced. Section 4 presents results of the new algorithm obtained for synthetic images. Resistance of the method to Gaussian blur and averaging is tested. Next, in Section 5 results of edge detection obtained for real images are shown. Finally, Section 6 concludes the paper.

II. Background on Subpixel Edge Detection

The main idea behind subpixel edge detection is to divide pixel into classes by determining edge location inside a pixel. Such approach is fundamentally different to the classical image processing where pixel is the basic and indivisible image component which is fully qualified to one class.

Subpixel edge detection is a challenging task. The discrete structure of a pixel grid causes irreversible loss in image intensities, influences shape of the objects present in the image and reduces edge information. Therefore, edge position inside a pixel can only be estimated with some probability.

The need of edge detection at subpixel level was firstly mentioned by the researchers in the late 70's [3]. Since then the issue of image processing with subpixel accuracy gained an interest of scientists and several approaches to this problem were proposed. Recently, methods for subpixel edge detection can be qualified into three main groups:

This work was not supported by any organization

1) curve fitting methods;
2) moment based methods;
3) reconstructive methods.

They are briefly characterized in the following subsections.

A. Curve fitting methods

Curve fitting methods determine subpixel edges by fitting various curves into edge points determined with a pixel accuracy. Firstly, for edge detection at pixel level the traditional edge detectors are used. Then, fitting is performed in an image plane in order to obtain continuous border.

This methodology was used by Yao and Ju [4] who fitted cubic splines into spatial data points provided by Canny operator or by Breder [5] who applied B-spline interpolation. Similar approach was also proposed by Kisiworo [6] who used deformable models to obtain subpixel edge position.

The accuracy of subpixel edge obtained using curve fitting methods is strongly limited by the accuracy of edge detection at pixel level. These methods are also sensitive to badly defined edge points which can deform the resulting shape of the object. Therefore, curve fitting methods yields reasonable results only in applications where shape of the object is known a priori and edge is accurately located at pixel.

B. Moment based methods

Moment based approaches determine edge position by relating image moments into parameters of subpixel edge. Methods which use image intensity moments (regarding only pixel intensities) and spatial moments (regarding both pixel intensities and spatial information about pixel neighborhood) have been proposed.

History of the moment based approaches dates back to 80's when Machuca and Gilbert [7] proposed the first method using image moments to determine edge position. The method integrates region containing the edge in order to determine its position using moments found within the integrated region. The moments are defined based on properties of vector from the given pixel to the gravity center of a pixel square neighborhood. Although Machuca and Gilbert's method has no ability to determine edge with subpixel accuracy it was an inspiration for the following approaches to the considered problem.

Tabatabai and Mitchell [8] proposed a method for subpixel edge detection which fits three intensity moments into the ideal step edge. In their approach the ideal edge is defined as sequence of one intensity followed by the sequence of the second intensity. The moments are defined as a sum of pixel intensity powers and do not consider any spatial information. The main drawback to this method is that it determines edges only in non-decreasing or non-increasing intensity sequences.

Geometric moment approach developed by Lyvers [9] fits moments into 2D model of an ideal edge. This model is described by four parameters which indicate subpixel edge position. The relation between the parameters of an ideal edge and image moments is established; then edge subpixel position is determined. This procedure requires evaluation of six moments by convolving an image with circular masks. As a result, the method is computationally complex. Additionally, geometrical moments proposed by Lyvers are not orthogonal, what makes the method lacks optimality in information redundancy. Approach proposed by Ghosal and Mehrotra [10] eliminates this weakness by fitting orthogonal Zernike moments into Lyvers' edge model. Additionally, the complexity of the method is decreased as only three masks are required. Zernike moments have difficulty in describing small objects, however they are most commonly applied for subpixel edge detection. Recently, Bin [11] put forward orthogonal Fourier-Mellin moments (OFMM) proposed by Sheng and Shen [12] into the Lyvers' edge model. However, determination of subpixel edge position using OFMMs requires application of seven circular masks what causes complexity of the method.

The main drawback to moment-based approaches is lack of clear criteria for classifying pixels as edge or non-edge. Moreover, they produce response (i.e. parameters of subpixel edge) for every set of pixels containing change in image intensity and work properly only in a close neighborhood of the edge pixel. Therefore in the current form they can mostly be used to refine position of the properly defined coarse edges.

C. Reconstructive methods

Reconstructive approaches to subpixel edge detection attempt to restore continuous information about an edge from the discrete intensity sample values. These sample values are provided by the traditional methods to edge detection such as Sobel, Canny or LoG. Next, different interpolation, approximation and extrapolation techniques are applied.

The continuous image information can be reconstructed independently in the vertical and the horizontal direction or simultaneously in both directions. In the first case one dimensional image intensity functions are retrieved in every direction and the final result is a superposition of results obtained in each direction. For reconstruction performed in all directions simultaneously two-dimensional function is found. In both cases the coordinates of the characteristic points of the reconstructed image function (i.e. local extremes, zero crossings, inflection points, etc.) indicate edge position with subpixel accuracy. In order to diminish the complexity of edge detection, image intensity function is often reconstructed in some neighborhood of a coarse border. Therefore, firstly, standard feature selection is applied in order to determine the coarse edge. Then this location is refined to subpixel level by adapting local feature pattern in the closest neighborhood.

The reconstructive approaches to subpixel edge detection can be divided into following groups:

1. ***methods reconstructing image intensity function*** [13] which determine subpixel edge position based on properties of function modeling image intensity at the edge; these methods however are in minority, due to lack of characteristic points of image intensity function at the edge;
2. ***methods reconstructing image first derivative function*** [14], [15] which retrieve image gradient function at the edge based on gradient sample values provided by operators like Sobel, Prewitt [14] or Canny [15] - most commonly second order polynomial is fitted into gradient sample values in a small (3 - 5 pixels) neighborhood; several approaches using wavelet transform instead of image first derivative have also been proposed [19], [20];
3. ***methods reconstructing image second derivative function*** [16]–[18] which reconstruct continuous image 2nd derivative function at the edge based on sample values provided by operator LoG; most commonly image derivative function is linearly interpolated in the neighborhood where the 2nd image derivative function changes its sign [16], [17] then coordinates of the zero-crossings of the reconstructed derivative function determine edge position with subpixel accuracy.

D. Other methods

There are also several subpixel edge detection methods which do not meet the classification presented in the previous subsections. One of them is approach used by Stanke [23] or Ji [24] where subpixel edge position is indicated by center of gravity of a gradient peak. Bie and Liu [25] applied quad-tree decomposition to divide pixels into subpixels while Kisworo [6] determined subpixel edge using image energy computed based on image intensities and their Hilbert transform. Some methods based on curvelets [21][22] have also been proposed.

Regarding the classification presented in the preceding subsections, the method introduced in this paper is a combination of reconstructive and curve fitting methods. More detailed description of the method is given in the following part of this paper.

III. THE PROPOSED METHOD

The proposed method attempts to retrieve continuous edge information at the edge from the discrete image data. Reconstruction is performed only in the neighborhood of the edge. Therefore, the method starts from defining the coarse edge location. Then edge position is refined to subpixel level. Finally, continuous edge is obtained via cubic spline interpolation. The detailed description of the above mentioned steps is given in the following subsections.

A. Coarse edge determination

For the coarse edge determination Sobel gradient masks are applied. Input image L is convolved with the horizontal h_x and the vertical h_y gradient masks in accordance with Equation (1).

$$\nabla L \approx \sqrt{\left(h_x \otimes L\right)^2 + \left(h_y \otimes L\right)^2} \tag{1}$$

where $\otimes$ denotes convolution. The gradient image ∇L is next thresholded with a global threshold T. The value of T is determined using ISODATA algorithm [26]. The applied threshold selection method is an iterative approach which starts from assigning an arbitrary initial threshold. Then mean intensities of pixels above the initial threshold and below the initial threshold are computed and the new threshold is obtained as their average. The procedure is repeated based upon a new threshold as long as the threshold value changes.

When value of T is determined thresholding is performed in accordance with Equation (2). The operation produces binary image corresponding with the region of the highest gradient.

$$\nabla L'(x) = \begin{cases} 1 & for \;\; \nabla L \geq T \\ 0 & for \;\; \nabla L < T \end{cases} \tag{2}$$

Finally, the coarse edge is obtained as a result of skeletisation performed on the binary image $\nabla L'(x)$ in accordance with Equation (3) [27].

$$\partial L = \bigcup_{n=0}^{N} \left[\left(\nabla L' \Theta nH'\right) - \left(\nabla L' \Theta nH'\right) \circ H \right] \tag{3}$$

where: H denotes structuring element (see Fig. 1), Θ and $\oplus$ denote erosion and dilation respectively, $\circ$ denotes morphological opening and:

$$N = \max\{n \,|\, \nabla L'(x) \Theta nH' \neq \emptyset\} \tag{4}$$

$$H' = \{-h \,|\, h \in H\} \tag{5}$$

$$nH = \begin{cases} \overbrace{\{\bar{0}\} \oplus H \oplus \ldots \oplus H}^{n} & for \;\; n = 1,2,\ldots \\ \{\bar{0}\} & for \;\; n = 0 \end{cases} \tag{6}$$

Equation (3) is iterated until the convergence with the structuring elements shown in Figure 1 and all their 90° rotations.

0	0	0
X	1	X
1	1	1

X	0	0
1	1	0
X	1	X

Fig. 1 Structuring elements used for skeletisation.

Successive steps of the coarse border determination in a sample image are presented in Figure 2. Particularly, Figure 2a presents input image. In Figure 2b gradient image is shown. The result of global ISODATA thresholding is presented in Figure 2c. Finally, Figure 2d presents the coarse border.

B. Refining edge to subpixel level

In this step of the algorithm, reconstruction of gradient profile at the edge is performed. Gaussian function given by Equation (7) is fitted along the normal direction of edges into a gradient sample values provided by Sobel operator.

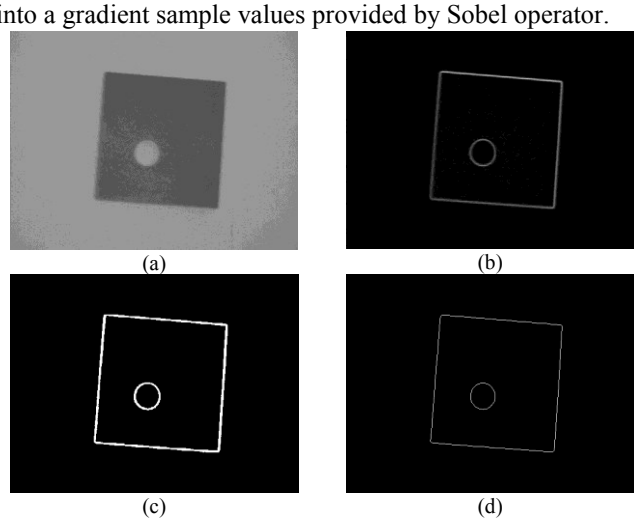


Fig. 2. Successive steps of coarse border determination; (a) input image; (b) gradient image; (c) image after ISODATA thresholding; (d) coarse edge.

$$f(x) = Ae^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{7}$$

where *f(x)* denotes gradient value at location *x*. The fitting is performed in a neighbourhood of every pixel from the coarse edge. It's linear neighbourhood in the gradient direction (the horizontal or the vertical) is considered. Several pixels on each side of the coarse edge are used.

The main idea of the proposed method is presented in Figure 3. In particular, in Figure 3a direction of a linear neighbourhoods used for gradient reconstruction are indicated. Figure 3b explains the idea of Gaussian function fitting along the normal direction of the edge. The figure presents 3D surface plot where gradient intensity is represented as a third dimension.

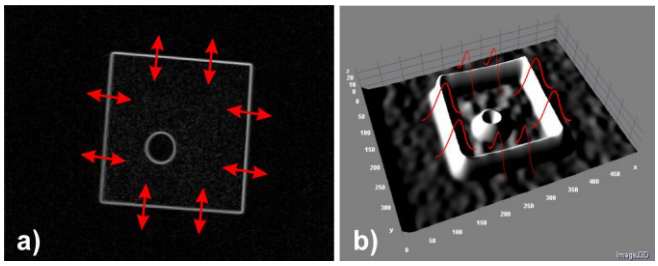


Fig. 3. The main idea of the proposed method; (a) direction of linear neighborhoods used for gradient reconstruction; (b) an idea of Gaussian function fitting along the normal direction of the edge presented in 3D surface plot.

Parameters of the Gaussian function fitted into a gradient sample values determine properties of the edge. Specifically:

– σ - describes a blur level of the edge;
– A - corresponds with gradient maximum value at the edge;
– μ - determines subpixel position of the edge pixel.

An example Gaussian function fitted into the discrete gradient sample values in a neighbourhood of pixel at location $x=27$ is presented in Figure 4. Empty circles correspond with gradient sample values shown under the graph. The coordinate μ of the maximum of the approximating Gaussian function indicates the edge location with subpixel accuracy.

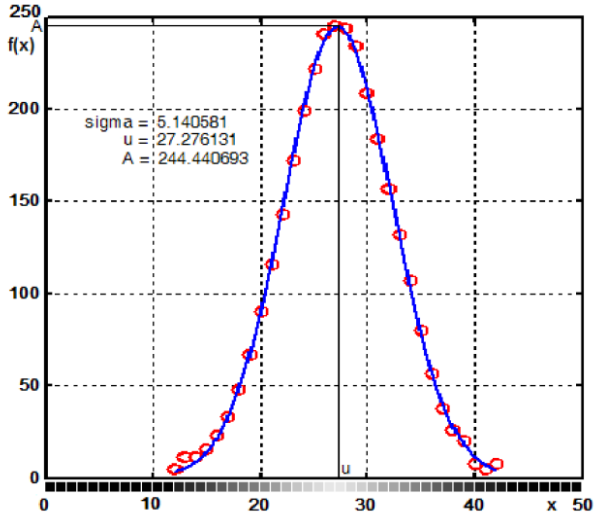


Fig. 4. Gaussian function fitted into gradient sample values.

Parameters of the Gaussian function are obtained via multidimensional unconstrained nonlinear minimization using Nelder-Mead algorithm [28]. This is an effective and computationally compact, simplex based method for finding a local minimum of a function of several variables. The algorithm works iteratively and uses only function values, without any derivative information. Each iteration of the simplex-based direct search begins with a simplex (i.e. a generalized triangle in n dimensions), specified by its n_s+1 vertices and the associated function values. One or more test points are computed, along with their function values, and the iteration terminates with bounded level sets.

The initial estimates for the fitting are as follows:

– for A – the maximum gradient value in the current neighbourhood;
– for μ – x coordinate of the central pixel in the regarded neighbourhood in case of the horizontal fitting and y coordinate of the central pixel in the regarded neighbourhood in case of the vertical fitting;
– for σ – size of a neighbourhood used for gradient function reconstruction.

C. Edge linking

In the last step of the algorithm edge linking is performed. It aims at obtaining continuous border from the subpixel positions of the coarse edge points determined in the previous step.

Firstly, the coarse edge is represented by means of chain code i.e. connected sequence of straight line segments of a specified length and direction; 8-connectivity segments are considered [29]. Despite the information about the succeeding pixel, each node of the chain contains information about the corresponding subpixel position.

Next, Cubic spline interpolation is performed over the subpixel positions corresponding with the consecutive pixels in the chain code in accordance with Equation (8).

$$F(x) = \sum F_i(x) \tag{8}$$

where: $x \in [x_i, x_{i+1}]$, $x_{p'0}=x_0<x_1<...<x_{n-1}<x_n=x_{p'k}$ and:

$$F_i(x) = a_i(x-x_i)^3 + b_i(x-x_i)^2 + c_i(x-x_i) + d_i \tag{9}$$

$$F_i(x_i) = y_i, F_i(x_{i+1}) = y_{i+1} \tag{10}$$

$$F'_{i-1}(x_i) = F'_i(x_i), F''_{i-1}(x_i) = F''_i(x_i) \tag{11}$$

$$F''_0(x_0) = 0, F''_{i-1}(x_n) = 0 \tag{12}$$

IV. Tests on Simulated Data

Firstly, the proposed method for subpixel edge detection was tested on the simulated data to verify if it works correctly. Specifically, the robustness of the method against the Gaussian blur and the averaging was investigated. In order to present robustness of Gaussian function in determining blurred edge position, the third step of the algorithm (i.e. edge linking) was not performed on the presented results.

Geometrically created 8-bit grayscale image of a circle was used to test the performance of the proposed approach. The circle of radius 50 pixels was centered at the position (75.0, 75.0). The intensity of the background was 52 while the intensity of the circle was 255.

The assessment of edge detection quality was made by means of:

- coordinates of the determined circle center;
- an average radius of the determined circle;
- standard deviation of the radius of the determined circle.

Coordinates of the circle center were defined as a center of gravity of the determined subpixel edge points and computed in accordance with the Equation (13).

$$\begin{cases} x_c = \dfrac{1}{k}\sum_{i=1}^{k} x_i \\ y_c = \dfrac{1}{k}\sum_{i=1}^{k} y_i \end{cases} \tag{13}$$

where k is number of subpixel edge points and x_i, y_i denote coordinates of i-th subpixel edge point. Radius was defined as an average distance of subpixel edge points from the real circle center (i.e. (75.0, 75.0)). This is expressed by Equation (14).

$$\bar{r} = \frac{1}{k}\sum_{i=1}^{k}\sqrt{(75-x_i)^2 + (75-y_i)^2} \tag{14}$$

The test image was distorted by:

- Gaussian filter of an increasing radius (from 1 to 10);
- an average filter of an increasing size (from 1 to 10).

Results provided by the proposed method (series: *gauss*) were compared with the previously proposed approaches to subpixel edge detection, such as:

- Tabatabai and Mitchell's method [8] (series: *tabatabai*);
- Zernike moments approach [10] (series: *zm*);
- parabola fitting approach [14][15] (series: *par*);

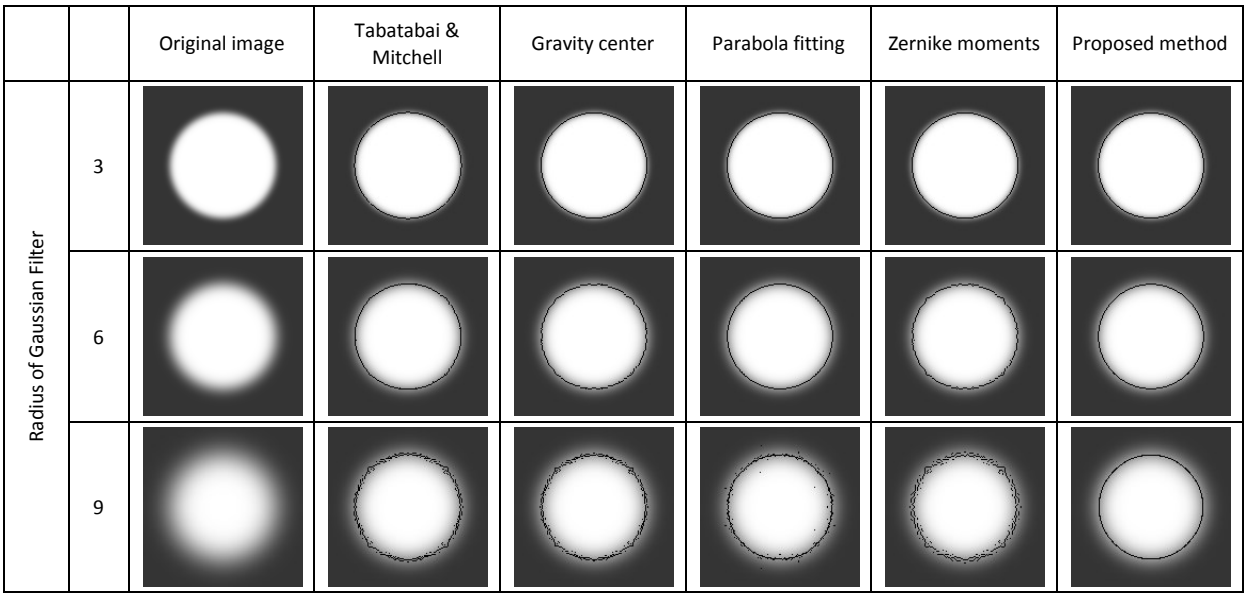


Fig. 5. Results of edge detection at sub-pixel level in images distorted by Gaussian blur.

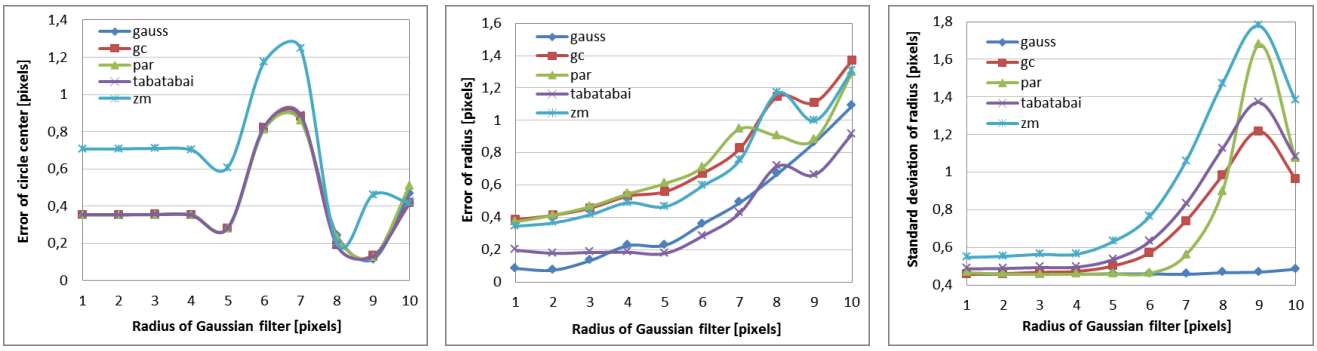


Fig. 6. Results of edge detection at sub-pixel level in images distorted by Gaussian filter.

– gravity center approach [23][24] (series: *gc*).

In all the cases the neighborhood equal to the diameter of the filter used to distort the image was regarded while determining edge position.

The results of edge detection in images distorted by the Gaussian blur are presented in Figures 5 and 6. Specifically, Figure 5 shows edges obtained with subpixel accuracy rounded to the closest pixel. Method used for edge detection is indicated above each column. Radius of the Gaussian filter used for blurring is given at the beginning of each row.

Figure 6 presents the comparison of edge detection results at subpixel level. Specifically, Figure 6a shows the error of circle center determination in function of radius of Gaussian filter used for blurring. The error is expressed by means of Euclidean distance between the real and the determined (in accordance with Eq. (13)) circle center. Figure 6b presents the error of circle radius determination in function of radius of Gaussian filter used for blurring. The error is expressed by means of the difference in length between the real and the determined (in accordance with Eq. (14)) radius. Finally, Figure 6c presents standard deviation of the circle radius in function of radius of Gaussian filter used for blurring.

Results of edge detection in images distorted by an average filter are presented in Figures 7 and 8. As previously, Figure 7 shows edges obtained with subpixel accuracy rounded to the closest pixel while Figure 8 presents comparison of edge detection results at subpixel level. The error of circle center (Fig. 8a), the error of radius (Fig. 8b) and the standard deviation of radius (Fig. 8c) are presented in function of size of an average filter used for blurring.

The results presented in Figures 5-9 prove the robustness of the proposed method against the Gaussian blur and the averaging.

Firstly, based on visual assessment (Fig. 5, Fig. 7), it should be underlined that only the proposed Gaussian fitting approach produces continuous and regular edges for a wide range of blur corruption. This is in the case of both: the Gaussian blur and the averaging for all regarded dimensions of filter used for image corruption. The other regarded approaches to subpixel edge detection produce continuous and regular edges only for low level of blur. With increasing blur, increases irregularity and discontinuity of the determined edge. This is also proved by graphs on Figures 6c and 8c showing standard deviations of the determined radius in function of radius of blurring filter.

Considering the comparison at subpixel level (Fig. 6,

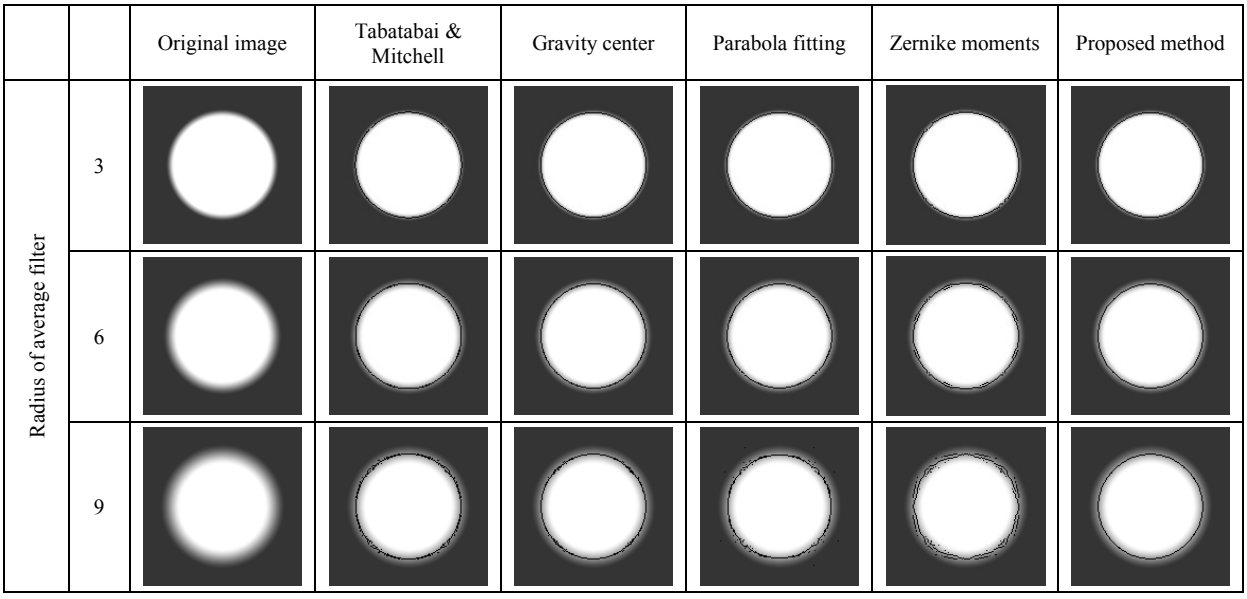


Fig. 7. Results of edge detection at sub-pixel level in images distorted by an average filter.

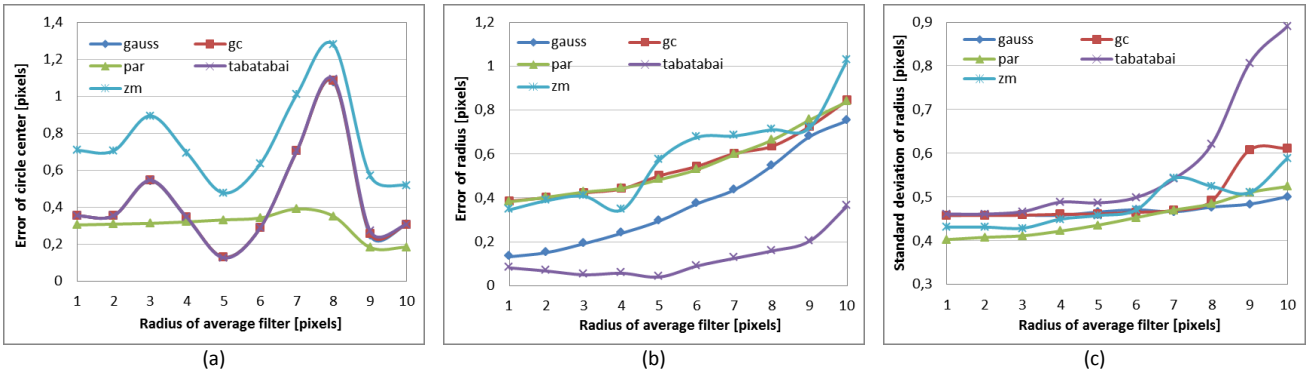


Fig. 8. Results of edge detection at sub-pixel level in images distorted by an average filter.

Fig. 8) it is clear that the best results are provided by the proposed method. In the case of Gaussian, blur the error of circle center determination is similar for all tested methods (see Fig. 8a). However, the smallest errors of an average radius provide the Gaussian fitting approach and Tabatabai and Mitchell's method (see Fig. 8b). Having in mind irregularity of edges provided by the second method, Gaussian fitting approach is superior in the case of Gaussian blur corruption. In the case of averaging, the error of an average radius provided by the proposed method is indeed higher than the error of Tabatabai and Mitchell's method, however, due to a very high standard deviation of radius of the latter method again Gaussian fitting approach can be regarded superior.

It should be also underlined, that Zernike moments approach fails when applied to blurred images (by both: Gaussian blur and averaging) - specifically for high level of blur the method produces the most irregular and discontinuous edges from all tested methods. Parabola fitting approach yields reasonable results for low and medium level of blur corruption, but for very blurry images the method has some problems with stability and produces edge points which significantly outstand from the border. Gravity center approach is always stable and for low level of blur produces continuous edges. However for increasing blur the method changes object shape. It can be observed that for large blur the determined edges become squarer.

Here, it should be concluded, that results obtained for the synthetic images prove correctness of the introduced method, its robustness against blurred edges and its superiority over other approaches to subpixel edge detection.

V. Tests on Real Data

In the next step, the tests on real images were performed in order to define the scope of applicability of the introduced method for subpixel edge detection. Specifically, images of heat-emitting specimens of metals and alloys obtained from the computerized system for high temperature measurements of surface properties [30][31] were considered. Due to the intense thermal radiation, usage of gas protective atmosphere and application of infrared filters the images are characterized by low contrast and blurred edges. Sample images obtained from the regarded system are shown in Figure 9. They present specimens of: copper at 853°C (Fig. 9a), steel at 797°C (Fig. 9b), copper at 1265°C (Fig. 9c) and steel at 1104°C (Fig. 9d).

Results of subpixel edge detection in sample images from Figure 9 are presented in Figure 10. Ten pixels at each side of the coarse edge pixel was regarded while refining edge position. Edges provided by the proposed approach are compared with results provided by other approaches. Specifically, the coarse edges are presented in the first row. The second row shows results of refining edge position using the proposed Gaussian fitting approach. The following rows presents edges provided by the gravity center approach, the parabola fitting approach, the Zernike moments approach and Tabatabai and Mitchell's method respectively. The results are rounded to the closest pixel. Interesting regions of specimen border are highlighted by red rectangles and magnified. In order to present robustness of the Gaussian function in determining blurred edge position, the third step of the algorithm (i.e. edge linking) was not performed on the presented results.

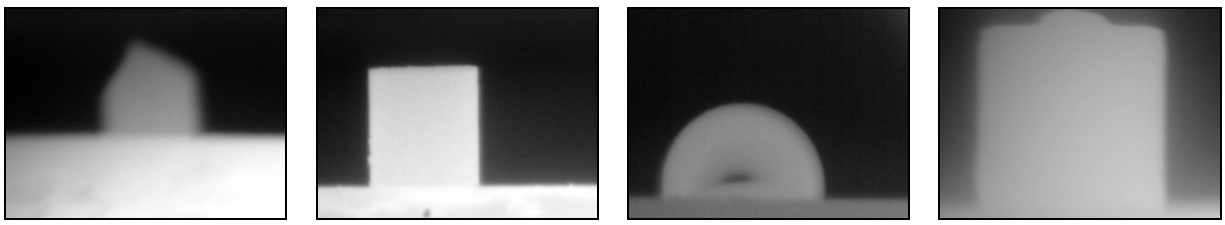


Fig. 9. Sample images of heat-emitting specimens of metals and alloys; (a) copper, 853°C; (b) steel, 797°C; (c) copper, 1265°C; (d) steel, 1104°C.

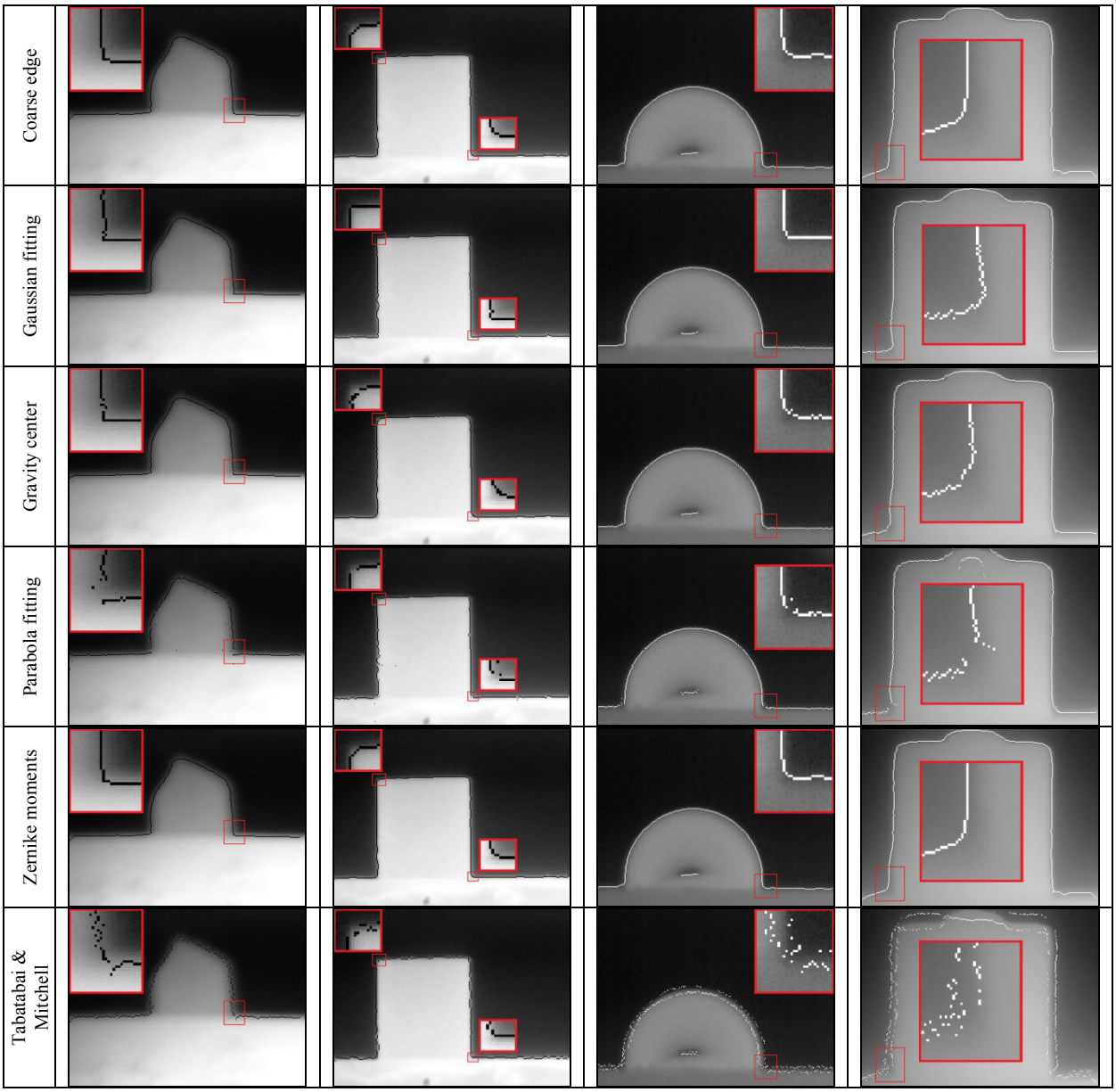


Fig. 10. Results of subpixel edge detection in images of heat-emitting specimens of metals and alloys. The original images are shown in Figure 9. Method used for edge detection is indicated at the beginning of each row.

Results shown in Figure 10 clearly show, that the proposed method significantly improves quality of edge detection in images of heat-emitting specimens of metals and alloys and outperforms other regarded approaches.

Firstly, it should be underlined that Zernike moments approach and Tabatabai and Mitchell's method fail when applied to the images of heat-emitting specimens. Subpixel edge provided by Zernike moments approach is at pixel level identical with the coarse edge. At subpixel level the differences between the coarse and the refined edge are negligible. In the case of the considered class of images Tabatabai and Mitchell's method becomes unstable and provides jagged, discontinuous, irregular and ambiguous edges what is unacceptable. Also parabola fitting approach has some problems with the continuity and the stability as it determines subpixel position of singular edge points visibly outstanding from the edge location. It is especially in the case when the supposed edge position is not in the center of the neighborhood of the coarse edge regarded while refining its position.

Both: the gravity center approach and the proposed Gaussian fitting approach are stable and provide continuous edges. However, the gravity center approach is sensitive to blur - it has problems in describing corners and rounds them off. In the examples shown in Figure 10 it is especially visible in contact of specimen, base and background or in case of image of steel at 797°C (Fig. 9b). Moreover, the gravity center approach tends to move border from the object when big neighborhood is used while refining the edge position.

The most significant increase in edge detection accuracy can be observed when the proposed Gaussian fitting

approach is used. Subpixel edges produced by the introduced method are regular. Moreover, they are continuous what diminishes number of edge information to be guessed during edge linking and makes the results more unequivocal. Additionally, improved edge fits closely specimen shape. Corners are sharp and well defined. This makes the method adequate for the considered class of images.

VI. CONCLUSIONS

In this paper problem of edge detection at subpixel level was considered. Specifically, precise edge detection in blurred images was regarded. The reconstructive method for subpixel edge detection was introduced. The method firstly determines the coarse edge using Sobel gradient masks, thresholding and skeletisation. Then it attempts to reconstruct continuous image gradient function at the coarse edge using Gaussian function. Position of the maximum of the reconstructing Gaussian function indicates edge position with subpixel accuracy.

The correctness and robustness of the method was proven by tests performed on geometrically created synthetic images under a wide range of blur corruption. Obtained results clearly show that the proposed method significantly improves quality of edge detection. The refined edges are continuous and much more regular than those provided by the previously proposed approaches to subpixel edge detection. The more blurred is the edge, the difference is more significant. The advantage of the Gaussian function based method can be seen both: in accuracy and the stability of the obtained subpixel edge position.

Tests performed on the real images obtained from the high temperature industrial vision system proved that the introduced method can be particularly useful in case of low contrast images with blurred and unsharp edges. In consequence it can be successfully applied in a wide spectrum of machine vision applications where accuracy is at premium.

REFERENCES

[1] Shan, Y., and Boon, G.W.: 'Sub-pixel location of edges with non-uniform blurring: a finite closed-form approach', Image Vision Comput., 2000, 18, (13), pp. 1015-1023 http://dx.doi.org/10.1016/S0262-8856(00)00040-8

[2] Ye, J., Fu, G., and Poudel, U.P.: 'High-accuracy edge detection with Blurred Edge Model', Image Vision Comput., 2005, 23, (5), pp. 453–467, http://dx.doi.org/10.1016/j.imavis.2004.07.007

[3] Nevtia, R., and Babu, K.: 'Linear feature extraction and description', Comput. Vision Graph., 1980, 13, (3), pp. 257–269, http://dx.doi.org/10.1016/0146-664X(80)90049-0

[4] Yao, Y., and Ju, H.: 'A sub-pixel edge detection method based on Canny operator', Proc. 6th Int. Conf. Fuzzy Systems and Knowledge Discovery, Tianjin, China, August 2009, pp. 97-100, http://dx.doi.org/10.1109/FSKD.2009.573

[5] Breder, R., Estrela, V., V., and de Assis, J. T.: 'Sub-pixel accuracy edge fitting by means of B-spline', Proc. IEEE Int. Workshop Multimedia Signal Processing, Rio de Janeiro, Brazil, October 2009, pp. 1-5, http://dx.doi.org/10.1109/MMSP.2009.5293265

[6] Kisworo, M., Venkatesh, S., and West, G.: '2-D edge feature extraction to subpixel accuracy using the generalized energy approach', Proc. IEEE Region 10th Int. Conf. EC3-Energy, Computer, Communication and Control Systems, New Delhi, India, August 1991, pp. 344-348, http://dx.doi.org/10.1109/TENCON.1991.753898

[7] Machuca, R., and Gilbert, A. L.: 'Finding edges in noisy scenes', IEEE Trans PAMI, 1981, 3, pp. 103-111, http://dx.doi.org/10.1109/TPAMI.1981.4767057

[8] Tabatabai, A. J., and Mitchell, O. R.: 'Edge location to sub-pixel values in digital imagery', IEEE Trans. PAMI, 1984, 6, (2), pp. 188-20l, http://dx.doi.org/10.1109/TPAMI.1984.4767502

[9] Lyvers, E. P., Mitchell, O. R., Akey M. L., and Reeves, A. P.: 'Subpixel measurements using a moment-based-edge operator', IEEE Trans. PAMI, 1989, 11, (12), pp. 1293-1309, http://dx.doi.org/10.1109/34.41367

[10] Ghosal, S., and Mehrotra, R.: 'Orthogonal moment operators for subpixel edge detection', Pattern Recogn. Lett., 1993, 26, (2), pp. 295-305, http://dx.doi.org/10.1016/0031-3203(93)90038-X

[11] Bin, T. J., Lei, A., Jiwen, C., Wenjing, K., and Dandan, L.: 'Subpixel edge location based on orthogonal Fourier-Mellin moments', Image Vision Comput., 2008, 26, (4), pp. 563-569, http://dx.doi.org/10.1016/j.imavis.2007.07.003

[12] Sheng, Y., and Shen, L.: 'Orthogonal Fourier-Mellin moments for invariant pattern recognition', J. Opt. Soc. Am., 1994, 11, (6), pp. 1748-1757, http://dx.doi.org/10.1364/JOSAA.11.001748

[13] Xu, G. S.: 'Sub-pixel edge detection based on curve fitting', Proc. 2nd Int. Conf. Information and Computing Science, Manchester, UK, May 2009, pp. 373-375, http://dx.doi.org/10.1109/ICIC.2009.205

[14] Bailey, D. G.: 'Sub-pixel profiling', Proc. 5th Int. Conf. Information Communications and Signal Processing, Bangkok, Thailand , December 2005, pp. 1311-1315, http://dx.doi.org/10.1109/ICICS.2005.1689268

[15] Rocket, P.: 'The accuracy of sub-pixel localization in the Canny edge detector', Proc. British Machine Vision Conf., Nottingham, UK, September 1999, http://www.bmva.ac.uk/bmvc/1999/papers/39.pdf, http://dx.doi.org/10.5244/C.13.39

[16] MacVicar-Whelan, P. J., and Binford, T. O.: 'Line finding with subpixel precision', Proc. DARPA Image Understanding Workshop, USA, April 1981, pp. 26-31, http://dx.doi.org/10.1117/12.965750

[17] MacVicar-Whelan, P. J., and Binford, T. O.: 'Intensity discontinuity location to subpixel precision', Proc. DARPA Image Understanding Workshop, USA, April 1981, pp. 752-754

[18] Jin, J. S.: 'An adaptive algorithm for edge detection with subpixel accuracy in noisy images', Proc. IAPR Workshop on Machine Vision Applications, Tokyo, Japan, November 1990, pp. 249-252

[19] Liu, C., Xia, Z., Niyokindi, S., Pei, W., Song, J., and Wang, L.: 'Edge location to sub-pixel value in color microscopic images', Proc. Int. Conf. Intelligent Mechatronics and Automation, Chengdu, China, August 2004, pp. 548-551, http://dx.doi.org/10.1109/ICIMA.2004.1384255

[20] Xu, G. S.: 'Linear array CCD image sub-pixel edge detection based on wavelet transform', Proc. 2nd Int. Conf. Information and Computing Science, Manchester, UK, May 2009, pp. 204-206, http://dx.doi.org/10.1109/ICIC.2009.160

[21] Tamrakar, A., and Kima B. B.: 'Combinatorial grouping of edges using geometric consistency in a lagrangian framework', Proc. Conf. Computer Vision and Pattern Recognition Workshop, New York, USA, June 2006, pp. 189, http://dx.doi.org/10.1109/CVPRW.2006.56

[22] Gebäck, T., and Koumoutsakos, P.: 'Edge detection in microscopy images using curvelets', BMC Bioinformatics, 2009, 10, (75), available on-line at: http://www.biomedcentral.com/1471-2105/10/75, http://dx.doi.org/10.1186/1471-2105-10-75

[23] Stanke, G., Zedler, L., Zorn, A., Weckend, F., and Weide, H. G.: 'Sub-pixel accuracy by optical measurement of large automobile components', Proc. 24th Conf. of IEEE Industrial Electronics Society, Aachen, Germany, August-September 1998, pp. 2431-2433, http://dx.doi.org/10.1109/IECON.1998.724107

[24] Ji, X., Wang, K., and Wei, Z.: 'Structured light encoding research based on sub-pixel edge detection', Proc. Int. Conf. Information Engineering and Computer Science, Wuhan, China, December 2009, pp. 1-4, http://dx.doi.org/10.1109/ICIECS.2009.5365408

[25] Bie, H. X., and Liu, C. Y.: 'Edge-directed sub-pixel extraction and still image super-resolution', Proc. 2nd Int. Congress Image and Signal Processing, Tianjin, China, October 2009, pp. 1-4, http://dx.doi.org/10.1109/CISP.2009.5301055

[26] Ridler, T., and Calvard, S.: 'Picture thresholding using an iterative selection method', IEEE Trans. Syst. Man Cyb., 1978, 8, pp. 630-632, http://dx.doi.org/10.1109/TSMC.1978.4310039

[27] Sidiropoulos, N. D., Baras, J. S., and Berenstein, C. A.: 'Discrete random sets: an inverse problem, plus tools for the statistical inference of the discrete boolean model', Proc. SPIE, 1992, 1769, pp. 32-43, http://dx.doi.org/10.1117/12.60630

[28] Lagarias, J. C., Reeds, J. A., Wright, M. H., and Wright, P. E.: 'Convergence properties of the Nelder-Mead simplex method in low dimensions', SIAM J. Optim., 1998, 9, (1), pp. 112-147, http://dx.doi.org/10.1137/S1052623496303470

[29] Gonzalez, R. C., and Woods, R. E.: 'Digital image processing', Prentice Hall, 2007

[30] Sankowski, D., Strzecha, K., and Jeżewski, S.: 'Digital image analysis in measurement of surface tension and wettability angle', Proc. Int. Conf. Modern Problems of Telecommunications, Computer Science and Engineers Training, Lviv-Slavsko, Ukraine, February 2000, pp. 129-130

[31] Fabijańska, A., and Sankowski, D.: 'Computer vision system for high temperature measurements of surface properties', Mach. Vision Appl., 2009, 20, (6), pp. 411-421, http://dx.doi.org/ 10.1007/s00138-008-0135-1

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 651–658
DOI: 10.15439/2014F365
ACSIS, Vol. 2

Computer Aided Assessment of Linear and Quadratic Function Graphs Using Least-squares Fitting

Sebastian Stoliński, Wojciech Bieniecki
Lodz University of Technology
Institute of Applied Computer Science
al. Politechniki 11, 90-924 Lodz, Poland
Email: {sstolin, wbieniec}@kis.p.lodz.pl

Magdalena Stasiak-Bieniecka
Lodz University of Technology
Department of Electrical Apparatus
ul. Stefanowskiego18/24, 90-924 Lodz, Poland
Email: stasiak@p.lodz.pl

***Abstract*—In this paper an image processing algorithm for automatic evaluation of scanned examination sheets is described. The discussed image contains selected function graphs sketched on a prepared sheets. This type of task is characteristic of final high school exams of natural sciences. Our challenge was to develop an evaluation algorithm, which works with a precision comparable to the teacher. If the image contains the correct solution, the algorithm should husk it from a set of random lines, deletions, amendments, drafts, bearing in mind, that lines were drawn by hand. In addition, the algorithm should calculate scores for partially correct solutions. An essential part of our proposal, which is image segmentation and identification, is based on least-squares fitting combined with 1-NN classification. The proposed solution is flexible and can be extended to other types of tasks such as drawing geometrical figures.**

I. Introduction

ELECTRONIC marking (e-marking), also known as Computer Assisted Assessment (CAA) or Computer Based Assessment (CBA) is relatively a new idea in the field of teaching. Its main advantage is facilitating the laborious process of design, delivery, collection, scoring and analysis of the assessments [10]. Other advantages of CAA are easier schedule and administration of assessments, the immediacy of results, their increased objectivity and security, the possibility of monitoring students and suitability for distance learning [6].

Students also seemed to consider CBA as being more promising, credible, objective, fair, interesting, fun, fast and less difficult or stressful, while they stated that they preferred computerized versus written assessment [7], [5]. In [11] it has been shown, that introduction of CAA allows to keep original accuracy of the exam and increase its reliability and even improve exam quality.

E-marking has been widespread in Great Britain and the USA. The experience gained by Examination Boards like AQA, OCR and EDEXCEL in Great Britain and ETS in the USA suggests that introducing e-marking improves the quality and reliability of the exams.

CAA and e-marking systems described in the literature have been designed for automatic evaluation of the exams carried out at the computer, which means, they use analytical or lexical form immediately.

The prevalence of this form of examination in the case of final examinations in primary and secondary schools can be difficult, due to the need to build IT infrastructure capable to handle a large number of people using the system at the same time.

We anticipate that still the dominant number of examinations will be carried out on the paper sheets and will be checked by humans.

Designing a system that will identify and evaluate the content of the answer sheets based on image processing and understanding algorithms can solve this problem, and in addition will be the value of research in the field of artificial intelligence.

Among the available literature and documentation we have not found any CAA systems that rely on image analysis and could be compared with ours, although there exist possible useful for our problem applications of:

- optical character recognition / intelligent character recognition (OCR/ICR) [18];
- lexicography analysis [13];
- image understanding techniques [14];
- neural networks to OCR/ICR and text identification algorithms [4], [12];
- Hough transform to object identification [9].

Our method of evaluation sketched function graphs relies on

- conversion the sketched shapes to its analytical form (coefficients an equation of a straight line or a parabola)
- merging of graph fragments basing on the evaluated coefficients
- comparison of the sketched graphs to model graphs by comparison of the evaluated coefficients to required values

In contrast to our previous approaches we do not use any reference image (possibly sketched by a teacher).

In practice each image should be segmented into individual primitives before the comparison. In our previous works [15],

978-83-60810-58-3/$25.00

[16], [17] we utilized cross-correlation [3] and Generalized Hough Transform (GHT) for this purpose. Unfortunately, the methods, we utilized, did not prove to be flexible and requires major redesign of the algorithm and for assessing new tasks. On the contrary, the least squares method can be used to identify most of the lines described analytically (through equation). The difficulty lies only in the transformation of the figure equation to the form of which the iterative process of fitting is convergent.

II. Description of the Task

The students have been asked to draw two graphs. First one is a function that has two points of discontinuity(eq. 1).

$$f(x) = \begin{cases} -4, & \text{for } x \leq -4 \\ -0.5x + 3, & \text{for } x \in (-4; 4) \\ -x + 9, & \text{for } x \geq 4 \end{cases} \quad (1)$$

Its graph (Fig. 1) consists of three line segments:

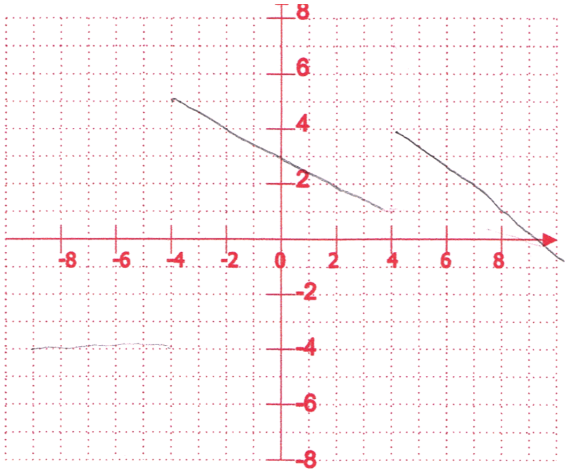


Fig. 1. An example of correctly sketched graph of the linear function

Second one is a parabola given by the formula (eq. 2):

$$f(x) = x^2 - 6 \cdot x + 5 \quad (2)$$

It has zeros in $x_1 = 1$ and $x_2 = 5$ and the minimum in the point $(3, -4)$ (see Fig. 2)

III. Fitting Shapes using Least-Squares

In the literature one can find a few examples of the use of fitting methods for finding the unknown parameters of geometric figures or function graphs. The publications are related to circles and ellipses [2], spheres, ellipses, hyperbolas, and parabolas [8]. Authors of these studies often adopt a two-phase method: first phase - algebraic fitting, second phase geometrical fitting.

Algebraic fit consists in solving the equation:

$$F(x, z) = \theta \quad (3)$$

where z is a vector of n parameters, x are points in l-dimensional (for example $l = 2$) space. To calculate the

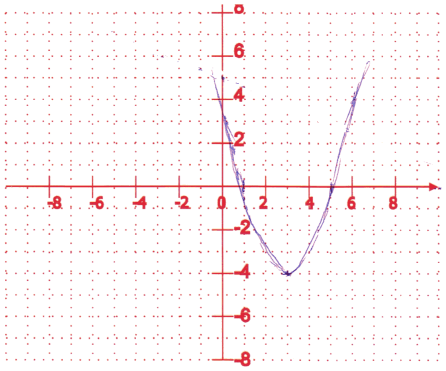


Fig. 2. An example of correctly sketched graph of the parabola

parameters for an analytical form of the function using fitting we must create the matrix B for which:

$$[B] = \theta B = \begin{bmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_{k-1}(x_1) & 1 \\ f_1(x_2) & f_2(x_2) & \cdots & f_{k-1}(x_2) & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ f_1(x_m) & f_2(x_m) & \cdots & f_{k-1}(x_m) & 1 \end{bmatrix} \quad (4)$$

Then, for each point $x_p, p = 1 \ldots m$, we have

$$f_1(x_p) \cdot a_1 + f_2(x_p) \cdot a_2 + \ldots + f_{k-1}(x_p) \cdot a_{k-1} + a_k = 0 \quad (5)$$

The algebraic fit usually does lead to the exact solution as it consists in solution of the overdetermined system ($m \gg k$) and choice the approximate solution to minimize the mean square error. In our approach we have chosen Singular Value Decomposition.

The idea of decomposition of the matrix $[B], (m \times k)$ is creation of a column orthogonal matrix $[U], (m \times m)$, a diagonal matrix $[W](m \times k)$ with zero or non-zero elements and a square, orthogonal matrix $[V](k \times k)$. We obtain the following equation:

$$[B] = [U] \cdot [W] \cdot [V]^T \quad (6)$$

The condition of orthogonality is

$$[U]^T \cdot [U] = [V]^T \cdot [V] = 1 \quad (7)$$

Evaluated matrix V corresponding to smallest singular value of $[B]$ contains a vector of parameters z in its last column.

$$[V] = \begin{bmatrix} v_{11} & \cdots & z_1 \\ v_{21} & \cdots & z_2 \\ \vdots & \ddots & \vdots \\ v_{k1} & \cdots & z_k \end{bmatrix} \quad (8)$$

This solution is taken as the starting point for the iterative process of geometric fit method, which allows a more accurate

approximation of the sought parameters. We also control the convergence of the iteration process, and at each step we can estimate the error. This in turn give grounds for considering whether it is possible to fit the selected function to given set of points.

The objective of the geometric fit is to minimize the geometric square error.

$$\epsilon = \sum \|x - x_0\| \tag{9}$$

In this study the Gauss-Newton method is used.

IV. Proposed algorithms

The proposed method of sketched graphs assessment is based on the analysis of all connected components in the image, obtained in the process of preprocessing. In the phase of image preprocessing, all printed content of the examination sheet is erased - only sketched lines remain. However, the process of overprints removal as well as the way in which the student draws a graph causes some defects - the filtered lines may not remain connected, for example, the circle becomes a dozen of arcs. Next the lines are thinned to a single pixel.

Assuming that the exercise was to draw a graph of a linear function, which is a straight line, the algorithm will work as follows:

1) For each connected component, run fitting procedure that finds a, b, c parameters for the equation $a \cdot x + b \cdot y + c = 0$. If the process is not convergent – it means that the component is not a line segment and reject it.
2) Calculated vectors $z = (a, b, c)$ form a feature linear space. Using clustering, find the most similar vectors. This means that the corresponding sets of points belong to one line.
3) for the union of the components obtained in the previous step re-do the fitting to accurately determine the parameters of the line.

This method can be generalized to simultaneously search for several straight lines. Then, the obtained vectors (a, b, c) will be subjected to clustering to indicate several groups of similar vectors.

In case the student task is to draw a number of different geometric shapes, we will try to adjust the parameters of different functions to each of them, looking for the best fit.

A. Fitting a straight line

In the first phase, the algorithm will minimize the algebraic error. Assume that a simple algebraic representation a straight line in the plane is given by

$$\begin{aligned} &F(x) = A^T \cdot x + c = 0, \\ &A = (a, b) \in \mathbb{R}^2, \\ &x \in \mathbb{R}^2, \\ &c \in \mathbb{R} \end{aligned} \tag{10}$$

The aim is to find values of a, b, c for given points x. Substituting the coordinates of the points in the above equation we obtain the system of equations. $[B] \cdot z = 0$, for the parameters $z = (a, b, c)$, where B is in the form:

$$[B] = \begin{bmatrix} x_{11} & x_{12} & 1 \\ x_{21} & x_{22} & 1 \\ \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & 1 \end{bmatrix} \tag{11}$$

Assuming, that $m > 3$, B is a rectangular, the system is overdetermined and probably inconsistent. The solution is approximated with Singular Value Decomposition.

Denote the obtained solution by $z = (a_0, b_0, c_0)$.

In the case of good fit such a solution is sufficient. However, if there are points x lying far from the approximated line, the bias arises, and therefore a second phase of the algorithm – geometric fit – must be launched.

Since there are many combinations of (a_0, b_0, c_0) corresponding to one line, one of the co-ordinates should be eliminated. Choose $z_M = \max(|a_0|, |b_0|, |c_0|)$ and assume, that $z = (1, b_0/z_M, c_0/z_M)$ or $z = (a_0/z_M, 1, c_0/z_M)$ or $z = (a_0/z_M, b_0/z_M, 1)$

The Gauss-Newton method involves the iteration which consists of two operations:

- solution of the system $-[J] \cdot h = f$ with the unknown vector h;
- correction of the solution $z = z + h$.

In the above system f is the objective function. It is a vector of distances of each point to the fitted line. J is the Jacobian - contains derivatives of the coordinates of the vector z (sought parameters). Formally:

$$\begin{aligned} &f = (f_1, f_2, \ldots, f_m); \quad f_i = \tfrac{|ax_{1i} + bx_{2i} + c|}{\sqrt{a^2 + b^2}}; \\ &J = \begin{bmatrix} \frac{\partial f_1}{\partial a} & \frac{\partial f_1}{\partial b} & \frac{\partial f_1}{\partial c} \\ \frac{\partial f_2}{\partial a} & \frac{\partial f_2}{\partial b} & \frac{\partial f_2}{\partial c} \\ \vdots & \vdots & \vdots \\ \frac{\partial f_m}{\partial a} & \frac{\partial f_m}{\partial b} & \frac{\partial f_m}{\partial c} \end{bmatrix}; \\ &J_{i1} = \frac{\operatorname{sgn}(ax_{1i} + bx_{2i} + c) \cdot x_{2i} - |ax_{1i} + bx_{2i} + c| \cdot \frac{a}{\sqrt{a^2 + b^2}}}{a^2 + b^2} \\ &J_{i2} = \frac{\operatorname{sgn}(ax_{1i} + bx_{2i} + c) \cdot x_{2i} - |ax_{1i} + bx_{2i} + c| \cdot \frac{b}{\sqrt{a^2 + b^2}}}{a^2 + b^2} \\ &J_{i1} = \frac{\operatorname{sgn}(ax_{1i} + bx_{2i} + c)}{\sqrt{a^2 + b^2}} \end{aligned} \tag{12}$$

The condition of convergence is calculated by the relative difference between the current and the previous solution.

$$\Delta = \frac{\|h\|_\infty}{\|z\|_\infty} \tag{13}$$

B. Fitting a parabola

In our discussion we will consider only the parabola which symmetry axis is parallel to the y-axis in the coordinate system. Accordingly, the parabola is defined by the algebraic equation:

$$F(x) = a \cdot x_1^2 + b \cdot x_1 + c \cdot x_2 + d = 0 \tag{14}$$

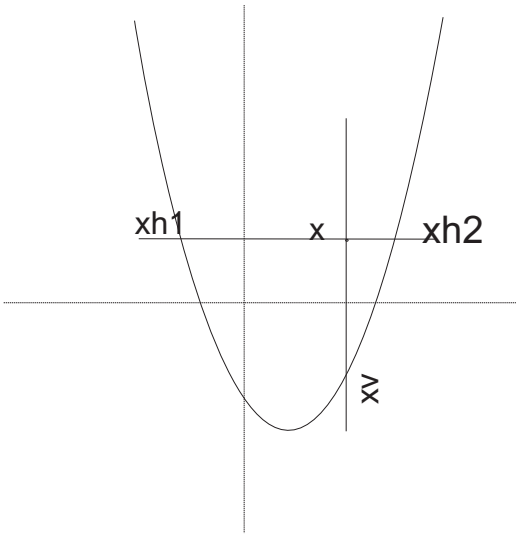


Fig. 3. Approximation of the distance of the point from the parabola

Denote the searched value $z = (a, b, c, d)$. To find the algebraic fitting for the equation denote:

$$[B] = \begin{bmatrix} x_{11}^2 & x_{11} & x_{12} & 1 \\ x_{21}^2 & x_{21} & x_{22} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1}^2 & x_{m1} & x_{m2} & 1 \end{bmatrix} \tag{15}$$

Just as in the case of a straight line, the system of equations is solved by SVD giving the value of $z_0 = (a_0, b_0, c_0, d_0)$. Then the geometric fit is carried out.

Similarly, as in the case of a straight line, in order to avoid ambiguity (and, consequently, the divergence of the process) we eliminate one of the parameters. Given the assumed position of the parabola it will be a parameter c.

$$F(x) = a/c_0 \cdot x_1^2 + b/c_0 \cdot x_1 + x_2 + d/c_0 = 0 \tag{16}$$

Iterative process will be conducted the same way as in the case of a straight line. However, due to a difficulty in determining the distance from the point to the parabola, some simplification will be used.

To calculate the distance of a point x to the parabola together with its partial derivatives, we derive a straight line from this point, parallel to the x-axis. It may cross the parabola at points X_{H1} and X_{H2}. We also derive a line from x, parallel to the y-axis, which always intersects the parabola (the point X_V). The points of intersection are calculated straight out of the equation of the parabola. Denote:

$$\begin{aligned} f_x &= \min(\rho(x, x_{h1}); \rho(x, x_{h2})) \\ f_y &= \rho(x, x_v) \end{aligned} \tag{17}$$

If x_{h1} and x_{h2} do not exist, assume that $f_x = f_y$.

Assuming that x_{H1} is a closer point, we calculate the partial derivatives of its distance.

$$\begin{aligned} \frac{\partial f_x}{\partial a} &= \text{sgn}(x_1 - x_{h1,1}) \cdot b \cdot (b + \sqrt{\Delta} + 2a(x_2 + c)/\sqrt{\Delta})/2a^2 \\ \frac{\partial f_x}{\partial b} &= \text{sgn}(x_1 - x_{h1,1}) \cdot (-1 - b/\sqrt{\Delta})/2a \\ \frac{\partial f_x}{\partial c} &= \text{sgn}(x_1 - x_{h1,1}) \cdot \sqrt{\Delta} \\ \frac{\partial f_y}{\partial a} &= \text{sgn}(x_2 - x_{v,2}) \cdot x_1^2 \\ \frac{\partial f_y}{\partial b} &= \text{sgn}(x_2 - x_{v,2}) \cdot x_1 \\ \frac{\partial f_y}{\partial c} &= \text{sgn}(x_2 - x_{v,2}) \end{aligned} \tag{18}$$

As an approximate distance of the point to the parabola we use the geometric average of calculated values f_x, f_y.

$$sf = f_x^2 + f_y^2 \qquad f = \frac{f_x \cdot f_y}{\sqrt{sf}} \tag{19}$$

Then, the Jacobian is defined by equations.

$$\begin{aligned} J_1 &= \left(\left(\frac{\partial f_x}{\partial a} f_y + \frac{\partial f_y}{\partial a} f_x \right) \sqrt{sf} - f \left(\frac{\partial f_x}{\partial a} f_x + \frac{\partial f_y}{\partial a} f_y \right) \right) / sf \\ J_2 &= \left(\left(\frac{\partial f_x}{\partial b} f_y + \frac{\partial f_y}{\partial b} f_x \right) \sqrt{sf} - f \left(\frac{\partial f_x}{\partial b} f_x + \frac{\partial f_y}{\partial b} f_y \right) \right) / sf \\ J_3 &= \left(\left(\frac{\partial f_x}{\partial c} f_y + \frac{\partial f_y}{\partial c} f_x \right) \sqrt{sf} - f \left(\frac{\partial f_x}{\partial c} f_x + \frac{\partial f_y}{\partial c} f_y \right) \right) / sf \end{aligned} \tag{20}$$

V. The main algorithm for image processing

Due to the fact that each image may have a different content and may include various types of function graph we decided to identify the graph using least-squares method for each line segment found in the image.

The preprocessing phase includes separation of the drawing from the rest of scanned examination sheet. Our method described in [16] has been replaced by more efficient color discrimination. For this purpose the coordinate system had to be printed red, while students use blue or black pen.

The color filtration condition is presented by the formula

for all pixel p IN I_l **do**
  if $|red(p) - green(p)| > 35 AND |red(p) - blue(p)| > 20$ **then**
    $p = (255, 255, 255)$
  end if
end for

Next steps of the preprocessing algorithm:

- image binarization using Otsu method [1];
- using morphological filters thin the lines, remove isolated points;

$$interval = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{21}$$

- using hit-miss transform detect and remove all crossings – trench the crossing lines;

$$interval = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{22}$$

- segment the image – label all connected components. We obtain I_l image containing $1 \ldots l$ components.

The following sections present our methods of analyzing and assess individually: discontinuous linear function and quadratic function.

A. Analyzing straight lines

The main part of the image processing is summarized in the following algorithms

Algorithm 1 Processing of connected components of I_l

$count \leftarrow 0$; $tolerance \leftarrow 1e-5$;
while $d > tolerance$ AND $count < 10$ **do**
  for all $line$ IN I_l **do**
    if $size(line) > 20$ **then**
      $(a_0, b_0, c_0) \leftarrow$ AlgebraicFitLine $(line)$
      $W_i \leftarrow (a, b, c, x_{\min}, x_{\max}, y_{\min}, y_{\max})_i \leftarrow$ GeometricFitLine $(line, a_0, b_0, c_0)$
    end if
  end for
  $d \leftarrow \max(res)$
  $count \leftarrow count + 1$
  $lines \leftarrow$Merge$(lines)$
end while

Here is the explanation of the Alg. 1. Repeat the operations until the error of fit is greater than expected: For each connected component, greater than 20 pixels do the operations: (if the number of iterations reaches 10, the procedure breaks regardless the obtained result)

1) algebraic fit (see sec. IV-A). A result is a vector (a_0, b_0, c_0)
2) geometric fit (see sec. IV-A). A result is a vector (a, b, c) and the convergence error res. The procedure of geometric fit is iterative and the iteration breaks if the expected convergence is reached ($tol < 1e-3$) or the number of iterations exceeds the established number ($maxiter > 10$).
3) find two nearest connected components and if the distance and merge them (assign the same label)

From the set of recognized lines we choose these, that lie nearest to the model lines. This is required, when the graph to be drawn consists of a few line segments (the function is not continuous). Model line parameters are obtained from the analytical form of the function. The comparison is carried out using 1-Nearest Neighbor statistical classifier, with the feature space identical to W_i vector in Alg. 1.

B. Obtaining a score

Each of line segments of the discontinuous function which is to be drawn is assessed independently, so if there are three segments of the graph, the maximum mark is 3. We take into account A, B, C coefficients of the linear function and the minimum and maximum x coordinate of the found line.

C. Analyzing a parabola

The initial phase of the algorithm is similar to Alg. 1 for processing straight lines

Algorithm 2 Processing of connected components of I_l

$count \leftarrow 0$; $tolerance \leftarrow 1e-1$;
while $d > tolerance$ AND $count < 10$ **do**
  for all $line$ IN I_l **do**
    if $size(line) > 20$ **then**
      $(a_0, b_0, c_0, d_0) \leftarrow$ AlgebraicFitParabola $(line)$
      $W_i \leftarrow (a, b, c, x_{\min}, x_{\max}, y_{\min}, y_{\max})_i \leftarrow$ GeometricFit $(line, a_0, b_0, c_0, d_0)$
    end if
  end for
  $d \leftarrow \max(res)$
  $count \leftarrow count + 1$
  $lines \leftarrow$Merge$(lines)$
end while

The main loop is repeated until we reach a required tolerance but no more than 10 times. As for straight lines, the algebraic fit followed by geometric fit are calculated. Again, as for straight lines, we merge the neighboring curves. The merge is done conditionally. For each pair the fitting is performed as for one set of points. If the convergence does not increase more than three times, the curves may be merged. After completion the process of fitting and merging the curves we obtain one or more curves that are possibly parabolas.

For the process of the assessment we take into account:

- (a, b, c) coefficients for the parabola equation $a \cdot x^2 + b \cdot x + c = 0$
- $(x_{\min}, x_{\max})$ - a position of the curve in the coordinate set.
- $(y_{\min}, y_{\max})$ - minimum or maximum value of the drawn function
- a count of pixels, that are not assigned as parabolas (these are possibly amendments)

If for the examined graph true are the statements:

1) values $(x_{\min}, x_{\max})$ and $(y_{\min}$ or $y_{\max})$ fall within a specified range
2) two of (a, b, c) parameters fall within a specified range
3) the count of amendment pixels is less than a specified threshold

the student receives 1 point. Moreover if all of (a, b, c) parameters fall within a specified range, the student receives a maximum score - 2 points.

If the count of amendment pixels exceeds a specified threshold, the graph is assigned as unrecognized.

VI. Teaching the algorithms

The algorithm for graphs classification runs in a supervised manner (with teaching). For each new type of the image (modified print-out, different task, different scanner) the step of teaching must be repeated. The teaching phase consists of evaluation acceptable ranges in the feature space. Some of them are calculated from the formula of the task:

- a, b, c parameters (both lines and parabola)
- x_{min} and x_{max} values

- y_{min} or x_{max} value (only for a parabola)

but their tolerances must be evaluated experimentally. Other parameters are:

- a threshold for detection of amendments
- a threshold for minimal length of a line (only for lines)

The teaching is carried out by comparing the proper scores (given by a teacher) for several test images containing graph sketches with the scores calculated by algorithms.

The aim of the tune-up is to minimize the overall error (number of different scores)

For the experiments 57 sheets with Task 1 and 72 sheets with task 2 have been used. All the sheets have been assessed by teachers for comparison. The examination sheets have been scanned in color mode with resolution 300 DPI. With this resolution each image containing extracted coordinate set with sketches has an area about 1 Megapixel.

Table I contains results of manual assessment of Task 1 for 11 exemplary works and the parameters obtained by an algorithm.

TABLE I
TASK 1: 11 EXEMPLARY SHEETS - ASSESSED BY A TEACHER

Sample	score for a segment 1	2	3	Total Score	Notes
1	1	1	1	3	
2	0	0	0	0	
3	0	0	0	0	
4	1	1	1	3	additional lines
5	1	1	1	3	
6	0	0	1	1	
15	0	0	0	0	
18	1	0	1	2	strike-throughs
19	1	1	1	3	
20	1	1	1	3	strike-throughs
21	1	1	1	3	strike-throughs

To the process of algorithm teaching 30 of 57 sheets have been randomly drawn (summarized in Table II).

TABLE II
TASK 1: THE PROCESS OF TEACHING

Property	Value
samples – in total	57
samples in a training	30
samples in a testing set	27
Segment 1 A	0 ± 0.0009
Segment 1 B	0.0015 ± 0.0007
Segment 1 C	1 ± 0.05
Segment 1 $x_{\min}$	45 ± 45
Segment 1 $x_{\max}$	340 ± 25
Segment 2 A	0.018 ± 0.0179
Segment 2 B	0.045 ± 0.043
Segment 2 C	1 ± 0.05
Segment 2 $x_{\min}$	340 ± 35
Segment 2 $x_{\max}$	770 ± 35
Segment 3 A	-0.0017 ± 0.0005
Segment 3 B	-0.0017 ± 0.0015
Segment 3 C	1 ± 0.05
Segment 3 $x_{\min}$	750 ± 25
Segment 3 $x_{\max}$	1050 ± 50

Note, that the values of parameters presented in Table II are expressed in pixels rather than units.

Tables III and IV present corresponding data for Task 2

TABLE III
TASK 2: 11 EXEMPLARY SHEETS - ASSESSED BY A TEACHER

Sample	score	Notes
10	1	
11	2	
12	2	
13	2	
14	1	
15	0	
16	1	
17	0	
18	0	additional objects
19	0	
20	1	

TABLE IV
TASK 2: THE PROCESS OF TEACHING

Property	Value
total samples	72
training set cardinality	36
testing set cardinality	36
Parameter A	-0.02 ± 0.015
Parameter B	28 ± 4
Parameter C	-10000 ± 5500
Parameter $x_{\min}$	530 ± 120
Parameter $x_{\max}$	880 ± 150
Parameter $maxpix$	< 600

VII. THE RESULTS OF THE EXPERIMENT

In Table V the best results (for parameters presented in Tables II and IV) for training sets have been summarized.

TABLE V
THE OF RESULTS FOR TRAINING SETS COMPARED TO TEACHER SCORES

Item	Test 1	Test 2
Samples	36	36
Unrecognized	3	1
Underestimated score	1	0
Overestimated score	2	1
Compliant score	30	34
Recognized compliant ratio	83%	94%

In Figs 4 – 8 exemplary correct and incorrect results are presented.

Table V summarizes the results obtained for testing sets.

TABLE VI
THE OF RESULTS FOR TESTING SETS COMPARED TO TEACHER SCORES

Item	Test 1	Test 2
Samples	36	36
Unrecognized	6	0
Underestimated score	2	1
Overestimated score	1	5
Compliant score	27	30
Recognized compliant ratio	75%	83%

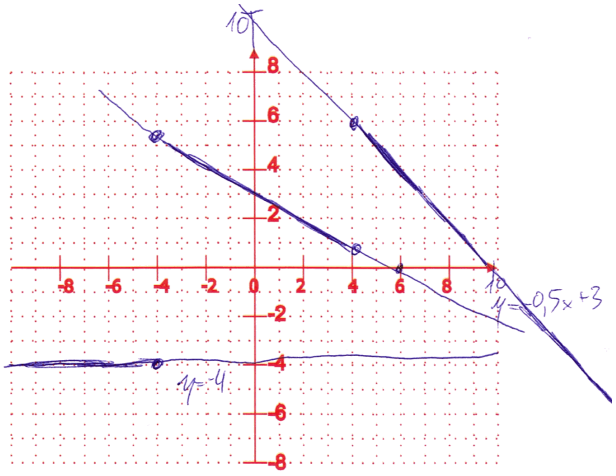


Fig. 4. Task 1, sample 4. A correct solution - but the algorithm did not recognize the sketches

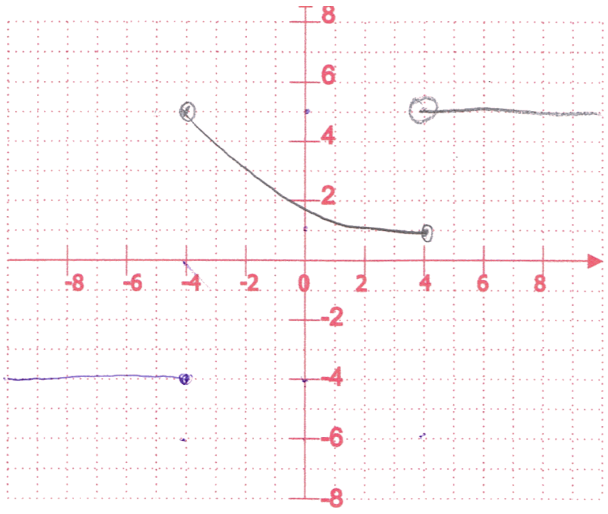


Fig. 6. Task 1, sample 32. An incorrect solution - the central line is not straight. The algorithm qualified this line as correct.

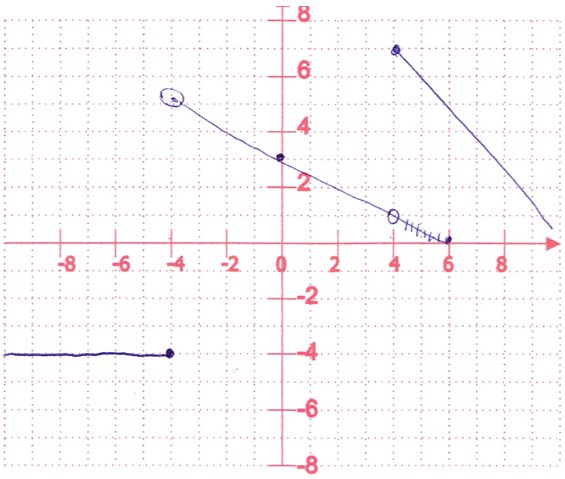


Fig. 5. Task 1, sample 29. A correct solution - underestimated by the algorithm. The central line has been disqualified.

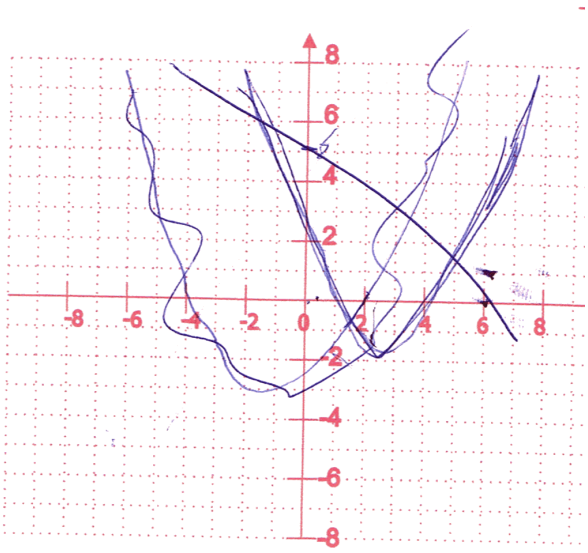


Fig. 7. Task 2, sample 18. Unrecognized sketches

VIII. Conclusions

According to this report (Table VI) the algorithm works properly for the case of the tested task. The errors occurred for the cases when the solutions contained amendments, strike-throughs. Tasks which were not recognized, contained additional objects or the drawings were done careless. Parameters of the algorithm were chosen so as to minimize the amount of erroneous assessments among the training set, even at the cost of increase of the number of unresolved tasks.

A general drawback of this approach is a necessity to train the algorithm before an assessment of a new task, but some of the parameters may be read from the analytical form of the function. Other parameters, which are tolerance, can be expressed as a percentage of the size of the unit.

Our future research will include the detection and assessment of graphs of trigonometric, exponential and rational functions. We'll try to extract multiple types of function graphs from one sketch (the task may include drawing a graphical solution of the set of inequalities)

Furthermore the algorithm of identification acceptable and redundant objects will be improved.

References

[1] Otsu, Nobuyuki: A Threshold Selection Method from Gray-level Histograms. IEEE Transactions on Systems, Man and Cybernetics **9**(1), 62–66 (1979). DOI 10.1109/TSMC.1979.4310076

[2] Gander, Walter and Golub, GeneH. and Strebel, Rolf: Least-squares fitting of circles and ellipses. BIT Numerical Mathematics **34**(4), 558–578 (1994). DOI 10.1007/BF01934268. URL http://dx.doi.org/10.1007/BF01934268

[3] J P Lewis: Fast normalized cross-correlation (1995)

[4] Charalambos Strouthopoulos and Nikos Papamarkos: Text identification for document image analysis using a neural network. Image Vision Comput. **16**(12–13), 879–896 (1998). DOI 10.1016/S0262-8856(98)00055-9

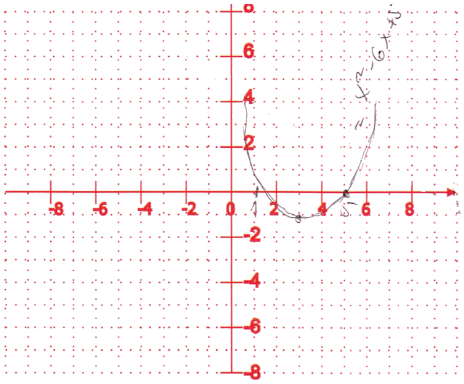


Fig. 8. Task 2, sample 39. Wrong solution (teacher scored 0 point), algorithm scored 1 point (partially correct).

[5] K Sambell and A Sambell and G Sexton: 19, , pp.179–192. Kogan Page, London (1999)

[6] M. Thelwall: Computer-based assessment: a versatile educational tool. Computers and Education **34**(1), 37–49 (2000). DOI 10.1016/S0360-1315(99)00037-8

[7] A C Croft and M Danson and B R Dawson and J P Ward: Experiences of using computer assisted assessment in engineering mathematics. Computers and Education **37**(1), 53–66 (2001). DOI 10.1016/S0360-1315(01)00034-3

[8] Sung Joon Ahn and Wolfgang Rauh and Hans-Jürgen Warnecke: Least-squares orthogonal distances fitting of circle, sphere, ellipse, hyperbola, and parabola. Pattern Recognition **34**(12), 2283 – 2303 (2001). DOI http://dx.doi.org/10.1016/S0031-3203(00)00152-7. URL http://www.sciencedirect.com/science/article/pii/S0031320300001527

[9] Sewisy, Adel A.: Graphical techniques for detecting lines with the hough transform. Int. J. Comput. Math. **79**(1), 49–64 (2002). DOI 10.1080/00207160211911

[10] Gavin Sim and Phil Holifield and Martin Brown: Implementation of computer assisted assessment: lessons from the literature. Research in Learning Technology **12**(3) (2004). DOI 10.1080/0968776042000259546

[11] D Fowles C Adams: How does assessment differ when e-marking replaces paper-based marking? 31st International Association for Educational Assessment Conference, Abuja, Nigeria, 4-9 September 2005 (2005)

[12] Shashank Araokar: Visual character recognition using artificial neural networks. CoRR **abs/cs/0505016** (2005)

[13] Prusa, D. and Hlavac, V.: Ninth International Conference on Document Analysis and Recognition, 2007. ICDAR 2007. pp., 849 –853 (2007). DOI 10.1109/ICDAR.2007.4377035

[14] Ryszard Tadeusiewicz and Marek R. Ogiela and Piotr S. Szczepaniak: Notes on a linguistic description as the basis for automatic image understanding. Applied Mathematics and Computer Science **19**(1), 143–150 (2009). DOI 10.1.1.390.8222

[15] Sebastian Stoliński and Wojciech Bieniecki and Jacek Stańdo: Automatic detection and evaluation of the spline function plot. Automatyka **14/3/1**, 879–896 (2010). DOI http://dx.doi.org/10.7494/automat

[16] Wojciech Bieniecki and Jacek Stańdo and Sebastian Stoliński: Automatic evaluation of examination tasks in the form of function plot. pp., 140–143. Polytechnic National University (2010)

[17] Sebastian Stoliński and Wojciech Bieniecki: The algorithms for automatic evaluation of selected examination tasks from the geometry. Automatyka **15/3**, 551–560 (2011). DOI http://dx.doi.org/10.7494/automat

[18] Sebastian Stoliński and Wojciech Bieniecki: Application of ocr systems to preprocessing and digitalization of paper documents. pp., 102–111. WULS Press, Warszawa (2011)

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 659–668
DOI: 10.15439/2014F174
ACSIS, Vol. 2

Efficient Volumetric Segmentation Method

Dumitru Dan Burdescu, Liana Stanescu, Marius Brezovan, Cosmin Stoica Spahiu
Computers and Information Technology Department
Faculty of Automation, Computers and Electronics, University of Craiova,
Craiova, Dolj, Romania
Address: Bvd. Decebal, Nr. 107, 200440, Tel./Fax: +40-251 438198
dburdescu@yahoo.com; lia_stanescu@yahoo.com

***Abstract* -- In this paper we extend our previous work for planar images by adding a new step in the volumetric segmentation algorithm that allows us to determine regions closer to it. There are huge of papers for planar images and segmentation methods and most of them are graph-based for planar images and very few papers for volumetric segmentation methods. However, even if image segmentation is a heavily researched field, extending the algorithms to spatial has been proven not to be an easy task. A true volumetric segmentation remains a difficult problem to tackle due to the complex nature of the topology of spatial objects, the huge amount of data to be processed and the complexity of the algorithms that scale with the new added dimension. The problem of partitioning images into homogenous regions or semantic entities is a basic problem for identifying relevant objects. Visual segmentation is related to some semantic concepts because certain parts of a scene are pre-attentively distinctive and have a greater significance than other parts. A number of approaches to segmentation are based on finding compact regions in some feature space. A recent technique using feature space regions first transforms the data by smoothing it in a way that preserves boundaries between regions. The key to the whole own algorithm of volumetric segmentation is the honeycomb cells. The pre-processing module is used mainly to blur the initial RGB spatial image in order to reduce the image store and to make algorithms to be efficient. Then the volumetric segmentation module creates virtual cells of prisms with tree-hexagonal structure defined on the set of the image voxels of the input spatial image and a volumetric grid graph having tree-hexagons as cells of vertices. Early graph-based methods use fixed thresholds and local measures in finding a volumetric segmentation.**

Index terms- Volumetric Segmentation; Graph-based segmentation; Color segmentation; Syntactic segmentation

I. Introduction and Related Works

THERE is a wide range of computational vision problems for planar images that could use of segmented images. The problem of partitioning images into homogenous regions or semantic entities is a basic problem for identifying relevant objects. Higher-level problems such as object recognition and image indexing can also make use of segmentation results in matching, to address problems such as figure-ground separation and recognition by parts. In both intermediate level and higher-level vision problems, contour detection of objects in real images is a fundamental problem. However the problems of planar image segmentation and grouping remain great challenges for computer vision. As a consequence we consider that a spatial segmentation method can detect visual objects from images if it can detect at least the most objects. We develop a visual feature-based method which uses a virtual spatial graph constructed on cells of prisms with tree-hexagonal structure containing half of the image voxels in order to determine a forest of spanning trees for connected component representing visual objects. Thus the spatial image segmentation is treated as a spatial graph partitioning problem. In addition our spatial segmentation algorithm produces good results from both from the perspective perceptual grouping and from the perspective of determining homogeneous in the input images. Early graph-based methods use fixed thresholds and local measures in finding a spatial segmentation.

In [1] one determined the normalized weight of an edge by using the smallest weight incident on the vertices touching that edge. Other methods for planar images [2], [3] use an adaptive criterion that depends on local properties rather than global ones. In contrast with the simple graph-based methods, cut-criterion methods capture the non-local cuts in a graph are designed to minimize the similarity between pixels that are being split [4] [5]. The normalized cut criterion [5] takes into consideration self similarity of regions. An alternative to the graph cut approach is to look for cycles in a graph embedded in the image plane. In [6] the quality of each cycle is normalized in a way that is closely related to the normalized cuts approach. Other approaches to planar image segmentation consist of splitting and merging regions according to how well each region fulfills some uniformity criterion. Such methods [7] use a measure of uniformity of a region. In contrast [2] and [3] use a pair-wise region comparison rather than applying a uniformity criterion to each individual region. Complex grouping phenomena can emerge from simple computation on these local cues [8]. A number of approaches to segmentation are based on finding compact regions in some feature space [9]. A recent technique using feature space regions [10] first transforms the data by smoothing it in a way that preserves boundaries between regions. Our previous works [11] and [12] are related to the works in [2] and [3] in the sense of pair-wise comparison of region similarity. In these papers we extend our previous work by adding a new step in the spatial segmentation algorithm that allows us to determine regions closer to it.

II. Constructing a Virtual Tree-Hexagonal Structure

The pre-processing module is used mainly to blur the initial RGB spatial image in order to reduce the image noise [13] and to apply the spatial segmentation algorithm. Then the segmentation module creates virtual cells of prisms with tree-hexagonal structure defined on the set of the image voxels of the input spatial image and a spatial triangular grid graph having tree-hexagons as cells of vertices. In order to allow a unitary processing for the multi-level system at this level we store, for each determined component C, the set of the tree-hexagons contained in the region associated to C and the set of tree-hexagons located at the boundary of the component. In addition for each component the dominant color of the region is extracted. This color will be further used in the post-processing

module if any. The contour extraction module determines for each segment of the image its boundary. The boundaries of the de determined visual objects are closed contours represented by a sequence of adjacent tree-hexagons. At this level a linked list of points representing the contour is added to each determined component. The post-processing module (if any) extracts representative information for the above determined visual objects and their contours in order to create an efficient index for a semantic image processing system.

A spatial image processing task contains mainly three important components: acquisition, processing and visualization. After the acquisition stage an image is sampled at each point on a three dimensional grid storing intensity or color information and implicit location information for each sample. The grid is the most dominant of any grid structure in image processing and conventional acquisition devices acquire square sampled images. An important advantage of using this grid is the fact that the visualization stage uses directly the voxels of the digitized image. We do not use a hexagonal lattice model because of the additional actions involving the double conversion between square and tree-hexagonal voxels. However we intent to use some of the advantages of the tree-hexagonal grid such as uniform connectivity. This implies that there will be less ambiguity in defining boundaries and regions [14]. As a consequence we construct a virtual tree-hexagonal structure over the grid voxels of an input image, as presented in Figure 1. This virtual tree-hexagonal grid is not a tree-hexagonal lattice because the constructed hexagons are not regular.

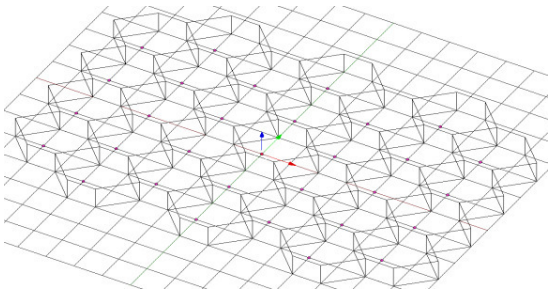


Fig. 1. Virtual Tree-Hexagonal structure constructed on the image voxels

Let I be a spatial initial image having the dimension h×w× z (e.g. a matrix having 'h' rows, 'w' columns and 'z' deep of matrix voxels). In order to construct a tree-hexagonal grid on these voxels we retain an eventually smaller image with

$$\begin{aligned} h' &= h-(h-1) \bmod 2, \\ w' &= w-w \bmod 4, \\ z' &= z- z \bmod 4. \end{aligned} \tag{1}$$

In the reduced image at most the last line of voxels and at most the last three columns and deep of matrix of voxels are lost, assuming that for the initial image h > 3 and w > 4 and z > 4, that is a convenient restriction for input spatial images.

Each tree-hexagon from the tree-hexagonal grid contains sixteen voxels: such twelve voxels from the frontier and four interior frontiers of voxels. Because tree-hexagons voxels from an image have integer values as coordinates we select always the left up voxel from the four interior voxels to represent with approximation the gravity center of the tree-hexagon, denoted by the pseudo-gravity center. We use a simple scheme of addressing for the tree-hexagons of the tree-hexagonal grid that encodes the spatial location of the pseudo-gravity centers of the tree-hexagons as presented in Figure 1.

Let h× w× z the three dimension of the initial spatial image verifying the previous restriction (e.g. h mod 2 = 1, w mod 4 = 0, z mod 4 = 0, h ⩾ 3 and w⩾4 and z⩾4). Given the coordinates (l,c,d) of a voxel p' from the input spatial image, we use the linearized function, $ip_{h,w,z}(l,c,d) = (l-1)w+c+d$, in order to determine an unique index for the voxel.

Let 'ps' be the sub-sequence of the voxels from the sequence of the voxels of the initial spatial image that correspond to the pseudo-gravity center of tree-hexagons, and' hs', 'ws' and 'zs' the sequence of tree-hexagons constructed over the voxels of the initial spatial image. For each voxel 'p' from the sequence ps having the coordinates (l,c,d), the index of the corresponding tree-hexagon from the sequence hs, ws and zs are given by the following relation:

$$f h_{h,w,z}(l,c,d) = [(l-2)w+c+d-2l]/4 +1 \tag{2}$$

In this case the following relation holds:

$$f h_{h,w,z}(l,c,d) = i. \quad . \tag{3}$$

Moreover it is easy to verify that the function 'f h' defined by the relation (2) is bijective. Its inverse function is given by:

$$f h^{-1} h, w, z (k) = (l,c,d) \tag{4}$$

where:

$$l = (2+ 4(k-1)/w \text{ if } h < w,$$

$$l = 2+ 4(k-1)/w +tw \text{ if } h \geqslant w, \text{ and } h = tw+h' , \tag{5}$$

$$c = 4(k-1)+2l-(l-2)w, \tag{6}$$

$$d = 4(k-1)+2l-(l-2)w. \tag{7}$$

Relations (4), (5), (6) and (7) allow us to uniquely determine the coordinates of the voxel representing the pseudo-gravity center of a tree-hexagon specified by its index (its address). In addition these relations allow us to determine the sequence of coordinates of all sixteen voxels contained into a tree-hexagon with an address 'k'.

The sub-sequence 'ps' of the voxels representing the pseudo-gravity center and the function 'fh' defined by the relation (2) allow to determine the sequence of the tree-hexagons 'Hs' that is used by the segmentation and contour detection algorithms. After the processing step the relations (4), (5), (6), (7) allow to up-date the voxels of the spatial initial spatial image for the visualization step.

Each tree-hexagon represents an elementary item and the entire virtual tree-hexagonal structure represents a triangular grid graph, G = (V,E), where each tree-hexagon 'H' in this structure has a corresponding vertex v ∈V. The set E of edges is constructed by connecting tree-hexagons that are neighbors in a 20-connected sense. The vertices of this graph correspond to the pseudo-gravity centers of the hexagons from the tree-hexagonal grid and the edges are straight lines connecting the pseudo-gravity centers of the neighboring hexagons, as presented in Figure 2.

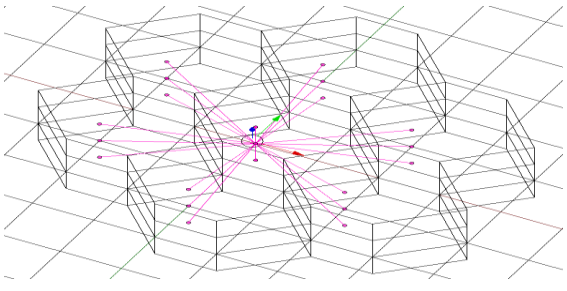


Fig.2. Triangular grid graph constructed on the pseudo-gravity centers of the tree-hexagonal grid

There are two main advantages when using tree-hexagons instead of all voxels as elementary piece of information:

• The amount of memory space associated to the graph vertices is reduced. Denoting by 'np' the number of voxels of the initial spatial image, the number of the resulted tree-hexagons is always less than np/4, and thus the cardinal of both sets V and E is significantly reduced;

• The algorithms for determining the visual objects and their contours are much faster and simpler in this case.

We associate to each tree-hexagon 'H' from V two important attributes representing its dominant color and the coordinates of its pseudo-gravity center, denoted by c(h) and g(h). The dominant color of a tree-hexagon is denoted by c(h) and it represents the color of the voxel of the tree-hexagon which has the minimum sum of color distance to the other twenty voxels. Each tree- hexagon 'H' in the tree-hexagonal grid is thus represented by a single point, g(h), having the color c(h). By using the values g(h) and c(h) for each tree-hexagon information related to all voxels from the initial image is taken into consideration by the spatial segmentation algorithm.

III. Volumetric Segmentation Algorithm

Let $V = \{h_1, \ldots, h_{|V|}\}$ be the set of virtual tree-hexagons constructed on the input spatial image voxels as presented in previous section and $G = (V,E)$ be the undirected spatial grid-graph, with E containing pairs of honey-beans cell (tree-hexagons) that are neighbors in a 20-connected sense. The weight of each edge $e = (h_i,h_j)$ is denoted by w(e), or similarly by w(hi,hj), and it represents the dissimilarity between neighboring elements 'hi' and 'hj' in a some feature space. Components of a spatial image represent compact regions containing voxels with similar properties. Thus the set V of vertices of the graph G is partitioned into disjoint sets, each subset representing a distinct visual object of the initial image.

As in other graph-based approaches [15] we use the notion of segmentation of the set V. A segmentation, S, of V is a partition of V such that each component $C \in S$ corresponds to a connected component in a spanning sub-graph

$GS = (V,ES)$ of G, with $ES \subseteq E$.

The set of edges $E - ES$ that are eliminated connect vertices from distinct components. The common boundary between two connected components $C', C'' \in S$ represents the set of edges connecting vertices from the two components:

$$cb(C',C'') = \{(hi,hj) \in E \mid hi \in C', hj \in C''\} \quad (8)$$

The set of edges $E-ES$ represents the boundary between all components in S. This set is denoted by *bound(S)* and it is defined as follows:

$$bound(S) = \cup_{C',C'' \in S} cb(C',C''). \quad (9)$$

In order to simplify notations throughout the paper we use Ci to denote the component of a segmentation S that contains the vertex $hi \in V$.

We use the notions of segmentation too fine and too coarse as defined in [2] that attempt to formalize the human perception of salient visual objects from an image. A segmentation S is too fine if there is some pair of components $C', C'' \in S$ for which there is no evidence for a boundary between them. S is too coarse when there exists a proper refinement of S that is not too fine. The key element in this definition is the evidence for a boundary between two components.

The goal of a spatial segmentation method is to determine a proper segmentation, which represent visual objects from an image.

Definition 1 Let $G = (V,E)$ be the undirected spatial graph constructed on the virtual tree-hexagonal structure of an image, with $V = \{h1, \ldots, h|V|\}$. A proper segmentation of V, is a partition S of V such that there exists a sequence $[Si, Si+1, \ldots, S\ f-1, S\ f]$ of segmentations of V for which:

• $S = S^f$ is the final segmentation and S_i is the initial segmentation,

• S_j is a proper refinement of S_{j+1} (i.e., $S_j \subset S_{j+1}$) for each $j = i, \ldots, f-1$,

• segmentation S_j is too fine, for each $j = i, \ldots, f-1$,

• any segmentation S_l such that $S_f \subset S_l$, is too coarse,

• segmentation S^f is neither too coarse nor too fine.

Let $C',C'' \in Sa$ be two components obtained by splitting a component $C \in Sb$. In this case C' and C'' have a common boundary, $cb(C',C'') \neq \emptyset$.

Our segmentation algorithm starts with the most refined segmentation, $S_0 = \{\{h_1\}, \ldots, \{h_{|V|}\}\}$ and it constructs a sequence of segmentations until a proper segmentation is achieved. Each segmentation S j is obtained from the segmentation S_{j-1} by merging two or more connected components for there is no evidence for a boundary between them. For each component of a segmentation a spanning tree is constructed and thus for each segmentation we use an associated spanning forest.

The evidence for a boundary between two components is determined taking into consideration some features in some model of the spatial input image. When starting, for a certain number of segmentations the only considered feature is the color of the regions associated to the components and in this case we use a color-based region model. When the components became complex and contain too much tree-hexagons, the color model is not sufficient and geometric features together with color information are considered. In this case we use a syntactic based with a color-based region model for regions. In addition syntactic features bring supplementary information for merging similar regions in order determine salient objects.

For the sake of simplicity we will denote this region model as syntactic-based region model.

As a consequence, we split the sequence of all segmentations,

$$S_i^f = [S_0,S_1, \ldots, S_{k-1},S_k], \quad (10)$$

in two different subsequences, each subsequence having a different region model,

$$S_i = [S_0,S_1, \ldots, S_{t-1},S_t],$$
$$S^f = [S_t, S_{t+1}, \ldots, S_{k-1},S_k], \quad (11)$$

where Si represents the color-based segmentation sequence, and Sf represents the syntactic-based segmentation sequence.

The final segmentation St in the color-based model is also the initial segmentation in the syntactic-based region model.

For each sequence of segmentations we develop a different algorithm. Moreover we use a different type of spanning tree in each case: a maximum spanning tree in the case of the color-based segmentation, and a minimum spanning tree in the case of the syntactic-based segmentation. More precisely our method determines two sequences of forests of spanning trees,

$$F_i = [F_0,F_1, \ldots, F_{t-1},F_t],$$
$$F^f = [F_{t'}, F_{t'+1}, \ldots, F_{k'-1},F_{k'}], \quad (12)$$

each sequence of forests being associated to a sequence of segmentations.

The first forest from Fi contains only the vertices of the initial graph, F0 = (V, Ø), and at each step some edges from E are added to the forest $F_l = (V,E_l)$ to obtain the next forest, $F_{l+1} = (V,E_{l+1})$. The forests from Fi contain maximum spanning trees and they are determined by using a modified version of Kruskal's algorithm, where at each step the heaviest edge (u,v) that leaves the tree associated to 'u' is added to the set of edges of the current forest.

The second subsequence of forests that correspond to the subsequence of segmentations Sf contains forests of minimum spanning trees and they are determined by using a modified form of Boruvka's algorithm. This sequence uses as input a new graph,

$G' = (V', E')$, which is extracted from the last forest, Ft , of the sequence Fi. Each vertex 'v' from the set V' corresponds to a component Cv from the segmentation St (i.e. to a region determined by the previous algorithm). At each step the set of new edges added to the current forest are determined by each tree T contained in the forest that locates the lightest edge leaving T. The first forest from F^f contains only the vertices of the graph G' , $Ft' = (V', \emptyset)$.

We focus on the definition of a logical predicate that allow us to determine if two neighboring regions represented by two components, Cl' and Cl'', from a segmentation Sl can be merged into a single component Cl+1 of the segmentation Sl+1. Two components, Cl' and Cl'' , represent neighboring (adjacent) regions if they have a common boundary:

$$ad\,j(Cl', Cl'') = true \quad if \quad cb(Cl', Cl'') \neq \emptyset,$$
$$ad\,j(Cl', Cl'') = false \quad if \quad cb(Cl', Cl'') = \emptyset \qquad (13)$$

We use a different predicate for each region model, color based and syntactic-based respectively.

$$PED(e,u) = [w_R(R_e - R_u)^2 + w_G(G_e - G_u)^2 + w_B(B_e - B_u)^2\,]^{1/2} \qquad (14)$$

where the weights for the different color channels, wR, wG, and wB verify the condition wR +wG +wB = 1. Based on the theoretical and experimental results on spectral and real world data sets, Gijsenij et al. [16] is concluded that the PED distance with weight-coefficients (wR =0.26, wG = 0.70, wB =0.04) correlates significantly higher than all other distance measures including the angular error and Euclidean distance.

In the color model regions are modeled by a vector in the RGB color space. This vector is the mean color value of the dominant color of tree-hexagons belonging to the regions.

The evidence for a boundary between two regions is based on the difference between the internal contrast of the regions and the external contrast between them [2] and [15]. Both notions of internal contrast and external contrast between two regions are based on the dissimilarity between two colors.

Let h_i and h_j representing two vertices in the graph G =(V,E), and let $w_{col}(h_i,h_j)$ representing the color dissimilarity between neighboring elements hi and hj, determined as follows:

$$w_{col}(h_i,h_j) = PED(c(h_i),c(h_j)) \quad if\ (h_i,h_j) \in E,$$
$$w_{col}(h_i,h_j) = \infty \quad otherwise, \qquad (15)$$

where PED(e,u) represents the perceptual Euclidean distance with weight-coefficients between colors 'e' and 'u', as defined by Equation (14), and c(h) represents the mean color vector associated with the tree-hexagon 'H'. In the color-based segmentation, the weight of an edge (hi,hj) represents the color dissimilarity,

$$w(h_i,h_j) = w_{col}(h_i,h_j).$$

Let Sl be a segmentation of the set V. We define the internal contrast or internal variation of a component $C \in S_l$ to be the maximum weight of the edges connecting vertices from C:

$$IntVar(C) = \max_{(h_i,h_j) \in C}(w(h_i,h_j)). \qquad (16)$$

The internal contrast of a component C containing only one tree-hexagon is zero:

$$IntVar(C) = 0, \quad if\ |C| = 1.$$

The external contrast or external variation between two components, $C', C'' \in S$ is the maximum weight of the edges connecting the two components:

$$ExtVar(C',C'') = \max_{(h_i,h_j) \in cb(C',C'')}(w(h_i,h_j)). \qquad (17)$$

We chosen the definition of the external contrast between two components to be the maximum weight edge connecting the two components and not to be the minimum weight, as in [2] because: (a) it is closer to the human perception (in the sense of the perception of the maximum color dissimilarity), and (b) the contrast is uniformly defined (as maximum color dissimilarity) in the two cases of internal and external contrast.

The maximum internal contrast between two components, $C', C'' \in S$ is defined as follows:

$$IntVar(C',C'') = \max(IntVar(C'), IntVar(C'')), \qquad (18)$$

The comparison predicate between two neighboring components C' and C'' (i.e., ad j(C',C'') = true) determines if there is an evidence for a boundary between C' and C'' and it is defined as follows:

$$diffcol(C',C'') = true, \ if$$
$$ExtVar(C',C'') > IntVar(C',C'') + th^{kg}(C',C''),$$

$$diffcol(C',C'') = false, \ if$$
$$ExtVar(C',C'') \leqslant IntVar(C',C'') + th^{kg}(C',C''), \qquad (19)$$

with the the adaptive threshold $th^{kg}(C',C'')$ given by

$$th^{kg}(C',C'') = th^{kg} / \min(|C'|, |C''|) \ , \qquad (20)$$

where |C| denotes the size of the component C (i.e. the number of the tree-hexagons contained in C) and the threshold 'th^{kg}' is a global adaptive value defined by using a statistical model.

The predicate 'diffcol' can be used to define the notion of segmentation too fine and too coarse in the color-based region model.

Definition 2 Let G = (V,E) be the undirected spatial graph constructed on the tree-hexagonal structure of a spatial input image and S a color-based segmentation of V. The segmentation S is too fine in the color-based region model if there is a pair of components $C', C'' \in S$ for which

$$ad\,j(C',C'') = true \wedge diffcol(C',C'') = false.$$

Definition 3 Let G = (V,E) be the undirected spatial graph constructed on the tree-hexagonal structure of a spatial input image and S a segmentation of V. The segmentation S is too coarse if there exists a proper refinement of S that is not too fine.

We decided to use the RGB color space because it is efficient and no conversion is required.

Let G= (V,E) be the initial graph constructed on the virtual tree-hexagonal structure of a spatial image. The proposed segmentation algorithm will produce a proper segmentation of V according to the Definition 1. The sequence of segmentations, Si f , as defined by Equation (10), and its associated sequence of forests of spanning trees, Fi f , as defined by Equation (12), will be iteratively generated as follows:

- The color-based sequence of segmentations, Si, as defined by Equation (11), and its associated sequence of forests, Fi, as defined by Equation (12), will be generated by using the color-based region model and a maximum spanning tree construction method based on a modified form of the Kruskal's algorithm [17].
- The syntactic-based sequence of segmentations, Sf, as defined by Equation (11), and its associated sequence of forests, F f, as defined by Equation (12), will be generated by using the syntactic-based model and a minimum spanning tree construction method based on a modified form of the Boruvka's algorithm.

The general form of the segmentation procedure is presented in Algorithm 1

Algorithm 1 Segmentation algorithm
1: procedure SEGMENTATION (l,c,d,P,H,Comp)
2: Input l, c, d, P
3: Output H, Comp
4: H ←CREATEHEXAGONALSTRUCTURE(l, c, d, P)
5: G←CREATEINITIALGRAPH(l, c, d, P,H)
6: CREATECOLORPARTITION (G,H,Bound)
7: G' ←EXTRACTGRAPH (G,Bound, th^{kg})
8: CREATESYNTACTICPARTITION(G,G', th^{kg})
9: Comp ←EXTRACTFINALCOMPONENTS(G')
10: end procedure

The input parameters represent the image resulted after the pre-processing operation: the array P of the spatial image voxels structured in 'l' lines, 'c' columns and 'd' depths. The output parameters of the segmentation procedure will be used by the contour extraction procedure: the tree-hexagonal grid stored in the array of tree-hexagons H, and the array Comp representing the set of determined components associated to the salient objects in the input spatial image. The global parameter th^{kg} is the thresholds.

The color-based segmentation and the syntactic-based segmentation are determined by the procedures CREATECOLORPARTITION and CREATESYNTACTICPARTITION respectively.

The color-based and syntactic-based segmentation algorithms use the tree-hexagonal structure H created by the function CREATEHEXAGONALSTRUCTURE over the voxels of the initial spatial image, and the initial triangular grid graph G created by the function CREATEINITIALGRAPH. Because the syntactic-based segmentation algorithm uses a graph contraction procedure, CREATESYNTACTICPARTITION uses a different graph, G', extracted by the procedure EXTRACTGRAPH after the color-based segmentation finishes.

Both algorithms for determining the color-based and syntactic based segmentation use and modify a global variable (denoted by CC) with two important roles:

• to store relevant information concerning the growing forest of spanning trees during the segmentation (maximum spanning trees in the case of the color-based segmentation, and minimum spanning trees in the case of syntactic based segmentation),

• to store relevant information associated to components in a segmentation in order to extract the final components because each tree in the forest represent in fact a component in each segmentation S in the segmentation sequence determined by the algorithm.

In addition, this variable is used to maintain a fast disjoint set-structure in order to reduce the running time of the color based segmentation algorithm. The variable CC is an array having the same dimension as the array of hexagons 'H', which contains as elements objects of the class Tree with the following associated fields:

(isRoot, parent, compIndex, frontier, surface, color)

The field 'isRoot' is a boolean value specifying if the corresponding tree-hexagon index is the root of a tree representing a component, and the field 'parent' represents the index of the tree-hexagon which is the parent of the current tree-hexagon. The rest of fields are used only if the field 'isRoot' is true. The field 'compIndex' is the index of the associated component.

The field 'surface' is a list of indices of the tree-hexagons belonging to the associated component, while the field 'frontier' is a list of indices of the tree-hexagons belonging to the frontier of the associated component. The field 'color' is the mean color of the tree- hexagon colors of the associated component.

The procedure EXTRACTFINALCOMPONENTS determines for each determined component C of Comp, the set sa(C) of tree-hexagons belonging to the component, the set sp(C) of tree-hexagons belonging to the frontier, and the dominant color c(C) of the component.

IV. Color-Based Region Algorithm

Let G = (V,E) be the undirected spatial graph constructed on the tree-hexagonal structure of a spatial image. The proposed color-based segmentation algorithm will produce a proper segmentation of V according to the Definition 1, where the notion of segmentation too fine is given by the Definition 2.

The sequence of segmentations, $(S_0,S_1, \ldots ,S_{t-1},S_t)$, and its associated sequence of growing forests, $(F_0,F_1, \ldots ,F_{t-1},F_t)$, will be iteratively generated, based on a maximum spanning tree construction method. We use a modified form of the Kruskal's algorithm [17] presented in Algorithm 2, where the trees generated at each step represent the connected components of spatial segmentation.

The input parameters of the color-based segmentation algorithm are the initial spatial graph 'G' and the array 'H' of the tree-hexagons from the tree-hexagonal grid. The output parameter is the list 'Bound' of edges representing the boundary of the final spatial segmentation. The global parameter threshold 'th^{kg}' is determinate by using Algorithm 1.

This value is used at the line 19 of Algorithm 2, where the expression th^{kg} (ti, t j) is given by the relation (20), t_i and t_j representing the components Ct_i and Ct_j respectively.

Because we use maximum spanning trees instead of minimum spanning trees the list of the edges E(G) is sorted in non-increasing edge weight. The forest of spanning trees is initialized in such a way each element of the forest contains exactly one tree-hexagon.

Algorithm 2 Color-based segmentation

1: **procedure CREATECOLORPARTITION(G,H, Bound)
2: Input G = (V,E), H = {h1, . . .,h|V|}
3: Output Bound
4: th^{kg} ←*DETERMINETHRESHOLD(G)
5: Bound ←hi ◅ Initialize Bound
6: for all i←1, |V| do
7: *MAKESET(hi) ◅ Initialize the disjoint set data structures
8: end for
9: ◅ At this point l ←0
10: ◅ and S0 ←{{h1}, . . . ,{h|V|}}
11: *SORT(E,$E\pi$)
12: ◅ $E\pi$ = ($e_{\pi 1}$, . . ., $e_{\pi|E|}$) is the sorting of E
13: ◅ in order of non-increasing weight
14: **for all** k←1, |E| **do**
15: ◅ Let $e_{\pi k}$ = (h_i ,h_j) be the current edge in $E\pi$
16: t_i ←*FINDSET(h_i)
17: t_j ←*FINDSET(h_j)
18: **if** $t_i \neq t_j$ **then**
19: if w(hi ,hj) ⩽ INTVAR(ti, t j)+ th^{kg} (ti, t j) then
20:* UNION(ti, t j,w(hi,hj))
21: ◅ l ←l+1
22: ◅ Sl ←Sl−1−{{Cti},{Ct j }} ∪ {Cti ∪Ct j }
23: else
24:* Add the edge (hi ,hj) the the list Bound
25: ◅ bound(Sl)←bound(Sl−1) ∪ {(hi,hj)}
26: end if
27: else
28: ◅ Do nothing, ti ∈ Ct j
29: end if
30: end for
31: end procedure

The expression th^{kg} (ti, t j) = th^{kg} /min(|Cti |,|Cj $_j$ |) at the line 19 of Algorithm 2 is very important at the beginning of the algorithm because initially the components considered contains only one tree-hexagon and in this case

IntVar(Cti ,Cj $_j$)=0, and th^{kg} /min(|Cti |,|Cj $_j$ |) = th^{kg}. In order to consider an edge (hi,hj) to belonging to the non-boundary class of edges and in consequence to merge the components Cti and Cj $_j$ corresponding to 'h_i' and 'h_j' respectively, it is necessary that $w(h_i,h_j) < th^{kg}$.

When the components grow and both components Ct_i and $C_{j\,j}$ contain more than one tree-hexagon, the external variation between Cti and $Cj_{\ j}$ decreases, and in this case the decision for merging or non-merging Cti and $Cj_{\ j}$ is affected more by their size than by the global threshold 'th^{kg}'.

For each segmentation Sl determined by Algorithm 2 and for each connected component C of the corresponding spanning graph Gl there is a unique maximum spanning tree, Fl(C), that maximize the sum of edge weights for this component.

The forest of all maximum spanning trees associated to the segmentation Sl is

$Fl = \cup C \in Sl\ \ Fl(C)$,

and algorithm makes greedy decisions about which edges to add to Fl . Every time when an edge is added to the maximum spanning tree a union of the two partial spanning trees containing the two vertices of the edge is made. In this way the sequence of the edges contained in the forest Fl of spanning trees is implicit determined at the line 14 of Algorithm 2.

Conversely for each spatial tree T from the forest Fl , the set of all vertices of the initial graph contained in the tree T is denoted by Set(T) and it represents the connected component of Sl associated to maximum spanning tree T:

$T = Fl(Set(T))$.

The functions MAKESET, FINDSET and UNION used by the segmentation algorithm implement the classical MAKESET, FIND-SET and UNION operations for disjoint set data structures with union by rank and path compression [17]. In addition the function call, UNION(ti, t j ,w(hi,hj)), performs the following operation, assuming that ti is the root of the new spanning tree resulted by combining the spanning trees represented by 'ti' and 't j' :

• determining CC[ti].surface as the concatenation of the lists CC[ti].surface and CC[tj].surface,

• determining CC[ti].frontier as a list of indices of tree-hexagons belonging to the frontier of the new component

$\{Cti \cup Ct\,j\}$,

• determining CC[ti].color as the value (ni*ci+nj*cj)/(ni+nj) , where ci = CC[ti].color, and 'ni' represents the number of elements in the tree CC[ti].

V. Syntactic-Based Region Algorithm

The syntactic-based region model uses some geometric properties of regions together with color information. We use a subset of syntactic features advocated [18] including homogeneity, compactness and regularity.

The region model contains the area of the region and the region boundary. As presented in the previous Subsection, for each region C the segmentation algorithm determines the set $sa(C)$ containing the tree-hexagons forming the region, and the set $sp(C)$ containing the tree-hexagons located at the boundary of the region. Because for each tree-hexagon 'H' we determine its dominant color $c(h)$ and its pseudo-gravity center $g(h)$, for each region C the following information can be further determined:

- the mean color of the region, $c(C)$, the area of the region, $a(C)$, and the length of the contour of the region, $p(C)$. In addition, for each pair of regions, Ci and Cj , the length $p(Ci,Cj)$ of the common boundary between these region can be determined.

In order to reduce the time complexity of the segmentation algorithm we estimate the area $a(C)$ and the perimeter $p(C)$ of a region C in function of the length of the sets $sa(C)$ and $sp(C)$ respectively. Assuming that the distance between two neighboring voxels situated on axis Ox, Oy or Oz has the value 1, the area of a tree-hexagon is 12 and thus the area of a region C is given by the following relation:

$$a(C) = 12\times|sa(C)|, \tag{21}$$

where $|sa(C)|$ represents the cardinal of the set $sa(C)$.

In order to determine a good final segmentation and to discover the salient objects from the input image, the syntactic based sequence of segmentations, Sf , as defined by Equation (11), can decomposed into several subsequences, each subsequence being determined by a modified form of the Boruvka's algorithm.

Let $i1 < i2 < \ldots < ix < ix+1$ be a sequence of indices, with $i1 = t$ and $ix+1 = k$, that allows a decomposition of the sequence Sf as follows:

$$\begin{aligned} Sf = (\ & Si1, Si1+1, \ldots, Si2-1, Si2, \\ & Si2+1, Si2+2, \ldots, Si3, \\ & \ldots \\ & Six+1, Six+2, \ldots, Six+1\ ). \end{aligned} \tag{22}$$

As presented in Algorithm 1 the procedure CREATESYNTACTICPARTITION implements the syntactic based segmentation, while the function GENERATEPARTITION is used to generate the subsequences of segmentations, $Sf1$, . . . ,Sfx , each subsequence of the form,

$$Sf\,j = (Si\,j, Si\,j+1, \ldots, Si\,j+1-1, Si\,j+1\ ), \tag{23}$$

being determined by the function GENERATEPARTITION at the j-th call. The last segmentation of the subsequence $Sf\,j$ generate by GENERATEPARTITION is also the input sequence of the $(\,j\,+1)$-th call of GENERATEPARTITION. The first input segmentation $Si1$ is the final segmentation St of the color based segmentation algorithm. The function DETERMINEWEIGHTS determines the set A of weights as defined by following relation.

The construction of A is realized as following:

1. Let SB = $[b_1,b_2,b_3,b_4]$ be the sequence contained the same elements as the set B in non-decreasing order. For this reasoning we choose another set of weight values, which is related to the initial set B;
2. Let 'r' be the lowest common divisor of the numbers (b2− b1), (b3−b2), and (b4−b3),
3. Let s = (b4−b1)/r ,
4. The set of weights that we use are:

$$A = \{a_0, a_1, \ldots, a_s\}, \tag{24}$$

where $a_0 = b_1$, $a_k = b_4$, $a_i = a_0$+i*r, for i = 1, . . . , s, and in addition $b_2, b_3 \in A$.

Algorithm 3 Syntactic-based Segmentation

1: ****procedure** CREATESYNTACTICPARTITION(G,G', th^kg)
2: **Input** G, G', th^kg
3: **Output** G'
4: A←*DETERMINEWEIGHTS(G')
5: $count$ ←0
6: **repeat**
7: G' ← *GENERATEPARTITION(G,G', th^kg,$newPart$)
8: **if** $newPart$ **then**
9: $count$ ←0
10: k←[$a0$ $a0$ $a0$ $a0$]^T
11: **end if**
12: th^kg ←*MODIFYWEIGHTS(G', k)
13: $count$ ←$count$ +1

14: *NEXTKVECTOR(k)
15: **until** $count = |A|^4$
16: **end procedure**

More formally, the j-th call of the function GENERATEPARTITION, for which the output parameter 'newPart' has the value 'true', is associated to the non-empty subsequence $Sf\ j$ of segmentations and it generates a sequence of graphs,

$$Gi\ j = (G^{i,j}i\ j\ ,G^{i,j}i\ j+1, \ldots ,G^{i,j}i\ j+1-1,G^{i,j}i\ j+1\ ), \quad (25)$$

and a sequence of associated forests of minimum spanning trees,

$$Fi\ j = (F^{i,j}i\ j\ ,F^{i,j}i\ j+1, \ldots ,F^{i,j}i\ j+1-1,F^{i,j}i\ j+1\ ), \quad (26)$$

such that the last forest is empty, $F^{i,j}i\ j+1 = \emptyset$. For each graph $G^{i,j}l$ from the sequence $Gi\ j$, $F^{i,j}l$ represents the forest of minimum spanning trees of $G^{i,j}l$, and $G^{i,j}l+1$ is the contraction of $G^{i,j}l$ over all the edges that appear in $F^{i,j}l$, as presented in Algorithm 4.
Because the last graph, $G^{i,j}i\ j+1$, of the sequence $Gi\ j$ cannot be further contracted the dissimilarity vectors of functions associated to the edge weights, $d(C(vi),C(vj))$, are not modified, and thus the edge weights, $w(vi,vj)$, as defined by the function GRAPH EXTRACTION are not modified. In order to restart the process for determining the new subsequence,

$$Sf\ j+1 = (Si\ j+1\ ,Si\ j+1+1, \ldots ,Si\ j+2\ ), \quad (27)$$

the first graph, $G^{i,j+1}i\ j+1$ of the sequence $Gi\ j+1$ differs from the last graph, $G^{i,j}i\ j+1$, of the sequence $Gi\ j$ by modifying only the weighted vector $\mathbf{k} \in \mathsf{K}$. The function MODIFYWEIGHTS of Algorithm 2 realizes this modification and recalculates the new global weighted threshold. In this case the values for the weighted vector $\mathbf{k}$ are sequential determined in the lexicographic order, generated by the procedure NEXTKVECTOR.

This constraint is necessary in order to realize a stopping criterion for the algorithm: the last graph cannot be modified and for all distinct values of the weighted vectors $\mathbf{k} \in \mathsf{K}$ and thus another partition cannot be determined. Each time when GENERATEPARTITION generates a non-empty sequence of segmentations, the output parameter 'newPart' became 'true' and the first vector of the set K is generated.

When GENERATEPARTITION generates an empty sequence of segmentations, 'newPart' is 'false' and the next vector in lexicographic order is generated by the procedure NEXTKVECTOR.

When sequentially for all distinct weighted vectors $\mathbf{k} \in \mathsf{K}$ (e.g. $|A|^4$ distinct vectors, with the set A specified by the relation (24)) generated in lexicographic order the function GENERATEPARTITION generates a empty sequence of segmentations, the procedure GCREATESYNTACTICPARTITION finishes.
Between the last graph, $G^{i,j}i\ j+1$, of the sequence $Gi\ j$ and the first graph, $G^{i,j+1}i\ j+1$ of the sequence $Gi\ j+1$, there is a sequence of graphs that differ only by the edge weights,

$$b\ Gi\ j = (\ b\ G^{i,j}1\ ,\ b\ G^{i,j}2\ , \ldots ,\ b\ G^{i,j}bni\ j\ ), \quad (28)$$

such that $b\ G^{i,j}1 = G^{i,j}i\ j$ and $b\ G^{i,j}bni\ j = G^{i,j+1}i\ j+1$. This sequence is obtained when the function GENERATEPARTITION generates an empty sequence of segmentations, with $bni\ j < |A|^4$.

As presented in Algorithm 4 the function GENERATEPARTITION generates at the j-th call the sequence of graphs $Gi\ j$ defined by Equation (25), and the sequence of forests of minimum spanning trees defined by Equation (26), where:

- the first graph of the sequence $Gi\ j$ is the input graph of the function (i.e. the parameter G'),
- the last graph of this sequence is the graph returned by the function.

The function GENERATEPARTITION is a generalized Greedy algorithm for constructing minimum spanning trees, as presented in [19]. At each iteration, 'l', of the function GENERATEPARTITION, the contraction of the tree $G^{i,j}l$ over all the edges that appear in the minimum spanning tree $F^{i,j}l$ is performed by the function CONTRACTGRAPH.

Algorithm 4 Generate a new sequence of partitions
1: ****function** GENERATEPARTITION(G,G', th^kg, *newPartition*)
2: **Input** G, G', th^kg, G' ◅ $G' = G^{i,j}i\ j$ is the input graph
3: **Output** *newPartition*
4: *newPartition*← *false* ◅ $l \leftarrow 0$
5: **repeat**
6: $k \leftarrow 0$
7: **for** $i \leftarrow 1,G'.n$ **do**
8: **if** $G'.adjEdges[i] \neq ()$ **then**
9: Determine the lightest edge 'e' adjacent to $G'.V[i]$
10: ◅ Let $ei \in G'.adjEdges[i]$ such that
11: ◅ $e = G'.E[ei] = (vi\ ,\ vj)$ is the lightest edge
12: $th^kl \leftarrow$ *DETERMINETHL(vi, vj)
13: **if** $e.w \leq \min(th^kg, th^kl)$ **then**
14: ◅ Determination of the MST $F^{i,j}i\ j+l$
15: $k \leftarrow k+1$
16: $e.inMST \leftarrow true$
17: **end if**
18: **end if**
19: **end for**
20: **if** $k > 0$ **then**
21: $G' \leftarrow$*CONTRACTGRAPH(G,G', th^kg)
22: ◅ Determination of the graph $G' = G^{i,j}i\ j+l+1$
23: ◅ $l \leftarrow l+1$
24: *newPartition*←*true*
25: **end if**
26: **until** $k = 0$
27: **return** G' ◅ $G' = G^{i,j}i\ j+1$ is the output graph
28: **end function**

The function DETERMINETHL returns the local weighted threshold th^{hl} associated to the components Cvi and Cvj , as presented in the following relations:

- the local weighted threshold associated with the weighted vector $\mathbf{k} \in \mathsf{K}$ and with the adjacent components C' and C'' of the segmentation Sl is denoted by $\mathbf{th}^{kl}(C',C'')$ and it is determined by considering the average of dissimilarity functions for anly adjacent components with C' and C'' from the segmentation Sl ,

$$\mathbf{th}^{kl}(C',C'') = b\mathbf{k}T\ \mathbf{l}(C',C''), \quad (29)$$

where the components of the vector $\mathbf{l}(C',C'')$ are determined, for

$i = 1,2,3,4$, as follows:

$$li(C',C'') = [\sum p(C',C'',Ca,Cb)edi(C',C'')]/[\sum p(C',C'',Ca,Cb)\ 1]\ , \quad (30)$$

where the predicate $p(C',C'',Ca,Cb)$ is defined as

$$p(C',C'',Ca,Cb) = ((Ca,Cb) \in Sl) \wedge (ad\,j(C',Ca) = true) \wedge (ad\,j(C'',Cb) = true). \quad (31)$$

The function implementing the contraction procedure, CONTRACTGRAPH, is similarly to the function EXTRACTGRAPH with the following differences:

- It detects the connected components specified by the edges marked as MST in the GENERATEPARTITION, and assigns to each vertex of the new generated graph the component it belongs to. The function DETERMINECOMPONENTS implements a *Depth-First-Search* traversal method on the input graph in order to enumerate the connected components.
- As in the color-based segmentation algorithm (see Algorithm 2), for each edge from the minimum spanning tree a union of the two partial spanning trees containing the two vertices of the edge is made by using the procedure UNION. In this way it is realized a reunion of the components associated to the vertices from each connected component of the input graph:

$$C(v) = \cup_{u \in Set(Tv)} C(u), \quad (32)$$

where '*Tv'* denotes the minimum spanning tree from the input graph associated to the connected component that represents the new created vertex in the output graph, and *Set*(*Tv*) represents the connected component associated to '*Tv'*.

- The weights of the new created edges and also the weighted threshold of the output graph use a weighted vector $\mathbf{k} \in K$ such that its components have a value random chosen from the set $A = \{a_0, a_1, \ldots, a_s\}$ by using the procedure ALEAKCHOOSE. This is an important aspect of the syntactic based segmentation algorithm and in this way the distribution of the weights of the four dissimilarity functions tends to became uniform.

The sequence $F\,f$ of forests of minimum spanning trees as defined by Equation (12) can be decomposed as the sequence Sf of segmentations as follows:

$$\begin{aligned} F\,f = (\ & Fi'1, Fi'1+1, \ldots, Fi'2-1, \\ & Fi'2, Fi'2+1, \ldots, Fi'3-1, \\ & \ldots \\ & Fi'x, Fi'x+1, \ldots, Fi'x+1-1\ ). \end{aligned} \quad (33)$$

Because the graph $G^{i\ j}i\ j+l$ and its corresponding minimum spanning tree $F^{ij}i\ j+l$, for $j = 1, \ldots, x$ and $l = 0, \ldots, i_{j+1} - i_j - 1$, share the same set of vertices, from algorithm of graph contraction one can see that each subsequence of forests determined at the jth call of the function GENERATEPARTITION,

$$F^{f}j = (Fi'j, Fi'j+1, \ldots, Fi'j+1-1, Fi'j+1\ ), \quad (34)$$

can determined for each $l = 0, \ldots, i_{j+1} - i^{j} - 1$ as follows:

$$E'i_{j}+l+1 = E'i_{j}+l \cup_{e \in F^{ij}i\,j+l} Orig(e), \quad (35)$$

where $E'u$ represents the set of the edges associated to the forest $F'u = (V',E'u\ )$, and $Orig(e)$ represents the edge from the initial graph G corresponding to the edge 'e' from the current graph $G^{ij}i\ j+l$.

The call of the procedure UNION at the line 22 of graph contraction allows the determination of the sequence of the segmentations $S\,f$ as defined by Boruvka's algorithm.

$$S^{i\,j+l+1} = \{Set(T) \mid T \in Fi\,j+l+1\} = \{C(v) \mid v \in G^{ij}i\,j+l+1\}, \quad (36)$$

for each $j = 1, \ldots, x$ and $l = 0, \ldots, i\,j+1-i\,j-1$. This relation specifies the fact that there is a bijective mapping between the components from the segmentations $Si\ j+l+1$ (or equivalently between the trees from the forests $Fi\,j+l+1$) and the vertices of the contracted graphs $G^{ij}i\,j+l+1$.

At *j-th* call of the function GENERATEPARTITION, each call of the function CONTRACTGRAPH generates a new segmentation, $S^{i\,j+l+1}$, with $l = 0, \ldots, i\,j+1 - i\,j - 1$, which tends to merge the components of the previous segmentation until regions closer to salient objects are detected.

Algorithm 5 Graph contraction

1: ****function** CONTRACTGRAPH(G,G', th^kg)
2: **Input** G, G' ⊲ $G' = G^{ij}i\,j+l$ is the input graph
3: **Output** th^kg
4: $n'' \leftarrow$ *DETERMINECOMPONENTS(G', *cIndex*)
5: ⊲ Determine connected components of G'
6: ⊲ Let n'' the number of connected components
7: ⊲ Assign to each component an index in the array *cIndex*
8: $G'' \leftarrow$ *CREATEGRAPH(n'', *cIndex*)
9: ⊲ Create a new graph with one vertex for each
10: ⊲ connected component in G', i.e., $G''.n = n''$
11: Initialize two arrays of bins, B' and B'', of dimension n''
12: **for** $i \leftarrow 1, G'.m$ **do** ⊲ Let $G'.E[i] = e = (vi\ , vj)$
13: $cj \leftarrow G'.V[vj\ ].comp$
14: Add i to the bin $B'[cj\ ]$
15: **if** $e.inMST$ **then**
16: $ei0 \leftarrow e.origEdge$
17: $(hi,hj) \leftarrow (G.E[ei0].vi, G.E[ei0].vji)$
18: ⊲ $(hi\ ,hj)$ is the original edge from G
19: ⊲ corresponding to the current edge $(vi\ , vj)$
20: $(ti, t\,j) \leftarrow$ (FINDSET(hi,CC), FINDSET(hj ,CC))
21: **if** ti 6= $t\,j$ **then**
22: *UNION(ti, $t\,j$, $e.w$,CC)
23: ⊲ Determination of the MST $Fi\,j+l+1$
24: ⊲ and of the segmentation $Si\,j+l+1$:
25: ⊲ $Fi\,j+l+1 \leftarrow Fi\,j+l \cup \{Orig(e)\}$,
26: ⊲ $Si\,j+l+1 \leftarrow Si\,j+l - \{\{Cti\},\{Ct\,j\ \}\} \cup$
27: ⊲ $\cup \{Cti \cup Ct\,j\ \}$
28: **end if**
29: **end if**
30: **end for**
31: **for** $i \leftarrow 1, n''$ **do**
32: **for all** $ei \in B'[i]$ **do** ⊲ Let $(vi, vj) = G'.E[ei]$
33: $ci \leftarrow G'.V[vi].comp$
34: Add ei to the bin $B''[ci]$
35: **end for**
36: **end for**
37: *ALEAKCHOOSE(k)
38: **for** $i \leftarrow 1, n''$ **do**
39: **if** $B''[i]$ 6= hi **then**

40: Determine the lightest edge from the bin *B*''[*i*]
41: ◅ Let *ei* ∈ *B*''[*i*] such that
42: ◅ *G*'.*E*[*ei*] = (*vi* , *vj*) is the lightest edge
43: *ei*0 ←*G*'.*E*[*ei*].*origEdge*
44: (*hi*,*hj*)←(*G.E*[*ei*0].*vi*,*G.E*[*ei*0].*vj*i)
45: (*ti*, *t j*)←(FINDSET(*hi*,*CC*), FINDSET(*hj* ,*CC*))
46: *dist* ← *COLORDIST(*ti*, *t j* ,*CC*)
47: *w*← *WEIGHT(*dist*, *ti*, *t j*,*CC*, *k*)
48: h*ci*, *cj*i←h*G*'.*V*[*vi*].*comp*,*G*'.*V*[*vj*].*comp*i
49: *ADDEDGE(*G*''*, ci*, *cj* ,*w*, *ei*0)
50: **end if**
51: **end for**
52: *th*^{k}*g* ←*DETERMINETHG(*G*'', *k*)
53: **return** *G*'' ◅ *G*'' = *G*$^{i\,j}$*i j*+*l*+1 is the output graph
54: **end function**

VI. Segmentation Results and Quantitative Evaluation

These modalities produce high-resolution voxel based datasets which are in fact data points on a regularly spaced three dimensional grid.

Because sampling data points from the real world is performed slice by slice the existing spatial segmentation techniques are often planar in nature, applying existing planar algorithms to the volume data slice by slice. The results are inferior to native volumetric based solution because these algorithms ignore the interaction between adjacent slices [20], [21], [22], [23].

However, even if image segmentation is a heavily researched field, extending the algorithms to spatial has been proven not to be an easy task. A true volumetric segmentation remains a difficult problem to tackle due to the complex nature of the topology of volumetric objects, the huge amount of data to be processed and the complexity of the algorithms that scale with the new added dimension.

Martin thesis [24] states that human segmentation can be used as the ground-truth reference in benchmarking segmentations produced by different methods. On the other hand, one may argue that human segmentation is subjective and will produce different segmentations for the same image but in most cases they will differ only in certain regions of local refinement. This idea has been considered in [25], [26] as a method of avoiding penalizing segmentations that are coarser or more refined than others.

In pattern recognition and information retrieval, Precision-Recall method has received a world-wide acceptance and it's considered as a standard measure because it offers good results for relevance [26].

In the general case, precision (or confidence) is defined as the fraction of retrieved cases that are relevant, while recall (or sensitivity) is the fraction of relevant cases that are retrieved. In other words, in the context of classification, the precision for a class is equivalent with the true positives accuracy which is the number of true positives (i.e. the number of cases that are correctly labeled as belonging to that class) divided by the total number of cases labeled as belonging to that class (including false positives, which are cases that were incorrectly labeled as belonging to the class).

$$\text{Precision} = TP/(TP + FP) \quad (37)$$

Also in this context, recall is equivalent with the true positives rate which is defined as the number of true positives divided by the total number of cases that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are cases that were not labeled as belonging to the positive class but should have been).

$$\text{Recall} = TP/(TP + FN) \quad (38)$$

The terms: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) compare the classifier's prediction against apriority external information that is considered as the ground truth (observation). These are synthesized in the contingency table (or confusion matrix), expressed in Table I.

TABLE I. PRECISION-RECALL CONTINGENCY TABLE

	Observation	
Prediction	TP – Correct result	FP – Unexpected positive result
	TN - Correct absence of result	FN – Missing negative result

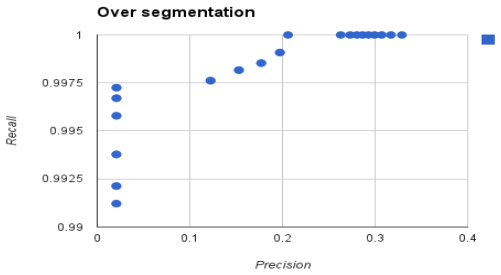


Fig. 3. Experiment Results

As said before, for image segmentation algorithms, Martin [24] proposes a method that outputs Precision-Recall curves as a mean to evaluate segmentation consistency. The curve offers a rich descriptor where both axes are sensitive and intuitive and the inherent trade-off between these two quantities can be easily analyzed.

Recall is defined as the proportion of boundary pixels/voxels in the ground truth that were successfully detected by the automatic segmentation, while precision is the proportion of boundary pixels/voxels in the automatic segmentation that correspond to the true boundary pixels. Precision is in fact a measure of the amount of noise in the classifier's result. The segmentation method used for the experimental results is based on simple hysteresis threshold. All voxels with the density within a specified threshold 'tk^{gh}' will be treated as boundary voxels while the others as empty space [27], [28], [29].

The results are as expected: the over-segmented volume has high recall and low precision (see figure 3), while the under-segmented image has low recall because it fails to find salient features for the volume, and also low precision (since because many boundary pixels remain unmatched).

VII. Conclusion

In this paper we present original and efficient volumetric segmentation methods. The major concept used in graph-based volumetric segmentation method is the concept of homogeneity of regions and thus the edge weights are based on color distance. Our previous works for planar images are related to other works in the sense of pair-wise comparison of region similarity. The key to the whole algorithm of volumetric segmentation is the honeycomb cells.

Here we presented only Color-based Segmentation, Syntactic-based Segmentation and Generate New Sequence of Partitions with Graph Contraction algorithms besides general algorithm of volumetric segmentation due to the entire space. Of course we have many procedures into general algorithm of volumetric segmentation methods. We have presented the original and efficient algorithm of volumetric segmentation methods and honeycomb cells used is the first run in volumetric segmentation algorithm. Then we can use the graph facilities and their related algorithms and computational complexity can be viewed as slow as the fundamental graph algorithms. Our original algorithms for Color-based Segmentation and Syntactic-based Segmentation are linear. Enhancement and generalization of this method is possible in several further directions. First, it could be modified to handle open curves for the purpose of medical diagnosis. Second, research direction is the using of composed shape indexing for both semantic and geometric image reasoning.

VIII. REFERENCES

[1] R. Urquhar, Graph theoretical clustering based on limited neighborhood sets. Pattern Recognition, 15(3), 173–187, 1982.

[2] P. Felzenszwalb, W. Huttenlocher, Efficient graph-based image segmentation. International Journal of Computer Vision, 59(2), 167–181, 2004.

[3] L. Guigues, L. Herve, L.P. Cocquerez, The hierarchy of the cocoons of a graph and its application to image segmentation. Pattern Recognition Letters, 24(8), 1059–1066, 2003.

[4] Y. Gdalyahu, D. Weinshall, M. Werman, Self-organization in vision: stochastic clustering for image segmentation, perceptual grouping, and image database organization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(10), 1053–1074, 2001.

[5] J. Shi, J. Malik, Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), 885–905, 2000.

[6] I. Jermyn, H. Ishikawa, Globally optimal regions and boundaries as minimum ratio weight cycles. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(8), 1075–1088, 2001

[7] M. Cooper, The tractibility of segmentation and scene analysis. International Journal of Computer Vision, 30(1), 27–42, 1998

[8] J. Malik, S. Belongie, T. Leung, J. Shi, Contour and texture analysis for image segmentation. International Journal of Computer Vision, 43(1), 7–27, 2001.

[9] D. Comaniciu, P. Meer, Robust analysis of feature spaces: color image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(5), 603–619, 2002.

[10] D. Comaniciu, P. Meer, Mean shift analysis and applications. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin, pp. 1197–1203, 1999.

[11] M. Brezovan, D. Burdescu, E. Ganea, L. Stanescu, An Adaptive Method for Efficient Detection of Salient Visual Object from Color Images. In Proceedings of the 20th International Conference on Pattern Recognition, Istambul, Turkey, pp. 2346–2349, 2010.

[12] D. Burdescu, M. Brezovan, E. Ganea, L. Stanescu, A new method for segmentation of images represented in a HSV color space. Lecture Notes in Computer Science, 5807, 606–616, 2009

[13] R. Gonzales, P. Wintz, Digital Image Processing. Reading, MA: Addison-Wesley, 1987.

[14] L. Middleton, J. Sivaswamy, Hexagonal Image Processing; A Practical Approach (Advances in Pattern Recognition). Springer-Verlag, 2005.

[15] L. Stanescu, D. Burdescu, M. Brezovan, CR. G. Mihai, Creating New Medical Ontologies for Image Annotation, Springer-Verlag New York Inc. ISBN 13: 9781461419082, ISBN 10: 1461419085", 2011

[16] A. Gijsenij, T. Gevers, M. Lucassen, A perceptual comparison of distance measures for color constancy algorithms, European Conference on Computer Vision, Marseille, France, pp. 208–221, 2008.

[17] T. Cormen, C. Leiserson, R. Rivest, Introduction to algorithms, Cambridge, MA: MIT Press, 1990.

[18] Bennstrom, C., Casas, J., Binary-partition-tree creation using a quasi-inclusion criterion. In Proceedings of the Eighth International Conference on Information Visualization, London, UK, pp. 259–294, 2004.

[19] Gabow, H.N., Galil, Z., Spencer, T., Tarjan, R.E., Efficient algorithms for finding minimum spanning trees in undirected and directed graphs. Combinatorica, 6, pg. 109–122., 1986

[20] P. Arbelaez, Pont-Tuset, J., Barron, J., Marqués, F., and Malik, J., Multiscale Combinatorial Grouping, in Computer Vision and Pattern Recognition (CVPR), 2014.

[21] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus, Indoor segmentation and support inference from RGBD images, in ECCV, 2012

[22] Abramowitz, M., Stegun, I.A. Handbook of Mathematical Functions. New York: Dover Publications, 1964

[23] R. Huang, V. Pavlovic, and D. N. Metaxas, A tightly coupled region shape framework for 3d, in Medical Image Segmentation, IEEE International Symposium on Biomedical Imaging (ISBI06), 2006.

[24] David Martin. An Empirical Approach to Grouping and Segmentation. PhD thesis, University of California, Berkeley, 2002.

[25] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in In Proceedings of International Conference on Computer Vision, no. 2, pp. 416–432, 2001.

[26] Y. Haxhimusa, A. Ion, and W. Kropatsch, Evaluating graph-based segmentation algorithms, in Proceedings of the 18th Internation Conference on Pattern Recognition, 2006.

[27] D. Powers, Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation, Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, 2011.

[28] P. Arbelaez, C. Fowlkes, and D. Martin. The Berkeley segmentation dataset and benchmark. Computer Science Department, Berkeley University. [Online]. Available: http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/

[29] F. J. Estrada and A. D. Jepson, "Benchmarking image segmentation algorithms," International Journal of Computer Vision, vol. 85, no. 2, pp. 167–181, Nov. 2009. [Online]. Available: http://dx.doi.org/10.1007/s11263-009-025

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 669–677
DOI: 10.15439/2014F304
ACSIS, Vol. 2

3D model reconstruction and evaluation using a collection of points extracted from the series of photographs

Katarzyna Rzążewska
Faculty of Mathematics and Information Science
Warsaw University of Technology
Koszykowa 75, 00–662 Warszawa, Poland
Email: katarzyna@rzazewska.eu

Marcin Luckner
Faculty of Mathematics and Information Science
Warsaw University of Technology
Koszykowa 75, 00–662 Warszawa, Poland
Email: mluckner@mini.pw.edu.pl

***Abstract*—This work describes the whole process of 3D model reconstruction. It begins with the representation of the method that is used to find the matching between photographs and the methodology to use the data to form the initial structure of the reconstructed model, represented by a point cloud. As a next stage, a refinement process is performed, using the bundle adjustment method. A set of stereovision methods is used later on to find a more detailed solution. Those algorithms use pairs of images, therefore as a prerequisite a set of routines that aggregates those results is studied. The paper is concluded with a description of how the point cloud is processed, including the surface reconstruction, to form the result. The described methodology is illustrated with reconstructions of three series of professional photographs from a public repository and one series of amateur photographs created especially for this work. The results were evaluated by the proposed area matching and contour matching measures.** ***Index Terms*—3D Reconstruction, Image Matching, Epipolar Geometry, Features Extraction, Models Evaluation**

I. Introduction

A RECONSTRUCTION of three dimensional (3D) models is one of the areas of the Computer Vision discipline that is quickly gaining momentum c.f. [1], [2], [3]. The development of information systems and the advancement in 3D graphics in general made it possible to create models that would depict real life objects. It has become even more important to be able to create models using two–dimensional photographs, taken using regular commodity digital cameras.

As one of the contributions of the following work, a computer tool has been developed that accomplishes the whole model reconstruction process. Out of a sequence of two–dimensional photographs, it can create a three dimensional full–colour model of the photographed object. The method is a mixture of algorithms based on features and solutions used in stereovision. The following project includes a description of this method. It also presents and comments on the results of applying the theory on a set of exemplary data series of digital photographs.

What follows is the main part of the work where the very process of 3D model reconstruction is explained. It begins with the representation of the method that is used to find the matching between photographs and the methodology to use the data to form the initial structure of the reconstructed model, represented by a point cloud. As a next stage, a refinement process is performed, using the bundle adjustment method. A set of stereovision methods is used later on to find a more detailed solution. Those algorithms use pairs of images, so as a prerequisite a set of routines that aggregates those results is studied. The description is concluded with information about the cloud processing, including the surface reconstruction, to form the result.

Each step of the process is illustrated with an exemplary image that show how the process progresses. This allows the reader to observe what the reconstruction process looks like. As an additional deliverable, a set of reconstructed models is presented. By comparing those images with the original models, the reader may decide on the quality of the process.

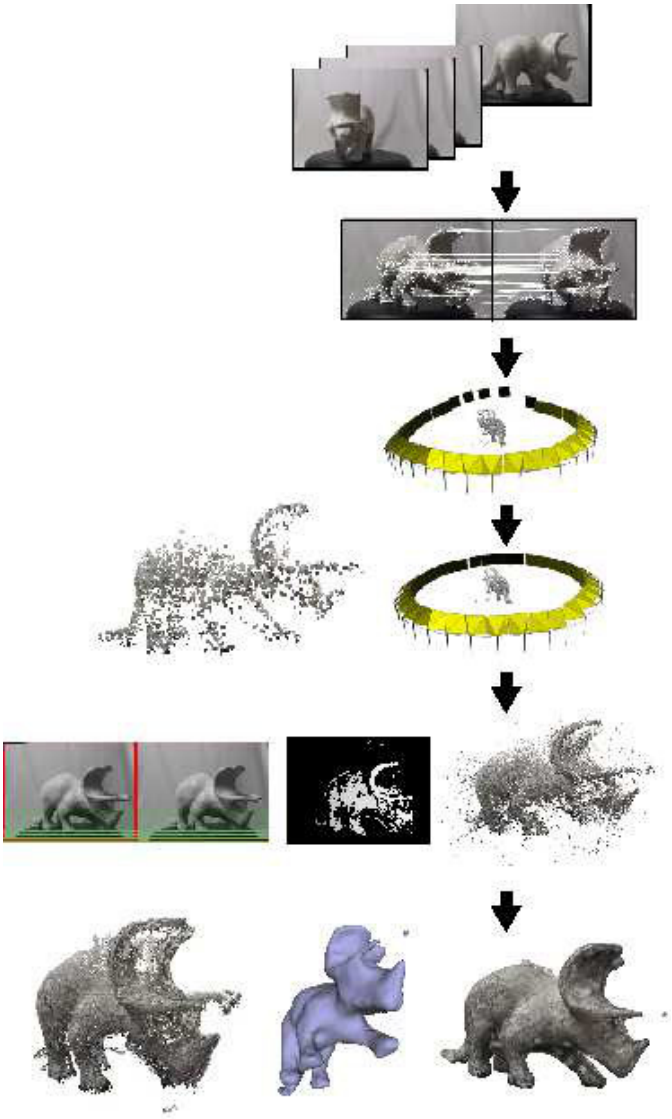
Several works present algorithms that result in high quality models. However, very often the reconstruction process bases on an expensive camera [2], specialist equipment such as a depth camera [3], or structured light [4].

The second important issue in the reconstruction based on computer vision algorithms (c.f. [5], [6], [7]) is lack of comparison methods. Very often, the result models are presented with evaluation of the quality different that visual comparison with the original object.

In this work, the proposed solution is based on low–cost algorithms and it is tested both on professional and amateur photographs. The snapshots from the four reconstructed model are presented in this work as well as the evaluation of their quality.

The paper is structured as follows. Section II presents basis of the epipolar geometry. The reconstruction process is briefly presented in Section III. The final models created from three series of photographs are described in Section IV and evaluated in Section V. Finally, the conclusions are presented in Section VI.

978-83-60810-58-3/$25.00



1) Input data (Section III-A)

2) Features detection and matching (Section III-B)

3) Structure and camera trajectory reconstruction (Section III-C)

4) Bundle adjustment (Section III-D)

5) Dense cloud creation (Section III-E)

6) Filtration and reconstruction (Section III-F)

Fig. 1. Reconstruction schema

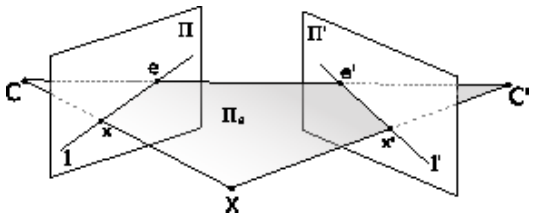


Fig. 2. Epipolar geometry defined by two cameras. $\mathbf{C}$ and $\mathbf{C}'$ – centres of cameras, Π and Π' – projection planes, $\mathbf{e}$ and $\mathbf{e}'$ – epipolar points, $\mathbf{x}$ and $\mathbf{x}'$ – projection of the point X on left and right view, $\mathbf{l}$ and $\mathbf{l}'$ – epipolar lines.

II. Preliminaries

The reconstruction bases on the epipolar geometry and essential information about its theoretical aspects are given in this section. The basis of the epipolar geometry is given in Figure 2. A 3D point X has projections $\mathbf{x}$ and $\mathbf{x}'$ on two views. The point X, both views $\mathbf{x}$ and $\mathbf{x}'$, and camera centres $\mathbf{C}$ and $\mathbf{C}'$ create the common plane. If the position of the point $\mathbf{x}$ is known it is also known that the projection $\mathbf{x}'$ lies on the line $\mathbf{l}'$. Therefore, the search for the point corresponding to $\mathbf{x}$ can be limited to the line $\mathbf{l}'$.

The algebraic representation of epipolar geometry is given by the fundamental matrix F. The matrix describes mapping between a point and its epipolar line. The special form of the fundamental matrix is the essential matrix E. The matrix E is a fundamental matrix corresponding to the pair of normalised cameras. A normalised camera describes the relation between image points expressed in normalised coordinates and 3D points. The relation is represented as the camera matrix P.

An important component of the matrix P is the calibration matrix K. The internal parameters K of the camera may be extracted from the matrix P by the decomposition. The inversion of calibration matrix creates normalised point on the basis of a point from a picture. This transformation is used to create an initial cloud of points in the presented process.

Detailed information on the epipolar geometry and relations between the matrices are given in [8].

(a) dinosaur

(b) teddy bear

(c) horse

(d) troll

Fig. 3. Photographs from tested series

III. Reconstruction

The reconstruction process described in this work consists of several stages. The schema of the process is given in Figure 1. The process starts with input photographs. Next, the matching between the photographs is detected. Two next steps create sparse clouds. Before the final reconstruction, a dense cloud is created. In the final stage, after a filtration, a surface is reconstructed. More information about the stages is given in the following sections.

A. Input data

The reconstruction procedure was tested on three examples from the public repository [9] and one created especially for this work. The data sets are series of photographs. Each professional series consists of photographs of objects placed on an automated turntable and photographed every 5 degrees. The photographs have a high 3 M–pixels resolution acquired by two Canon Powershot G1 digital cameras.

In our experiments, two series 'Teddy bear' and 'Horse' were represented by the full set of 72 photographs. In the second series 'Dinosaur', the number of photographs was reduced to 35. The resolution of photographs is 842×822 and 1600×1200 for the first and the second series respectively.

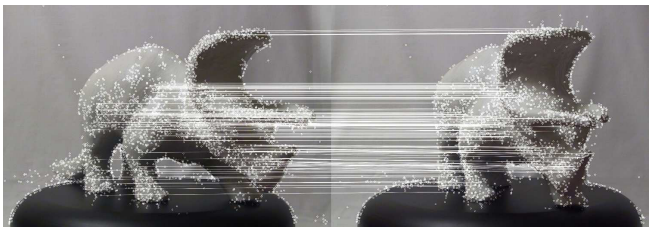


Fig. 4. Matching for two photographs with 20° rotation

The last series 'Troll' is a bit different from the others. The object was immobile and the photographer was moving. The photographs were taken without a stand with an irregular angle. The resolution of photographs is 1200×800. This series has only 33 photographs. The photographs were taken especially to test the reconstruction model presented in this work.

The three series present different approach to creation of data. The 'Teddy bear' and 'Horse' series are professional, detailed description of the objects. In the 'Dinosaur' series, photographs are still professional, but the number of images was reduced to decrease costs of documentation process. Both series were taken in a studio.

The 'Troll' series is an amateur documentation of the object created in an outdoor location.

Examples of photographs from all series are given in Figure 3.

B. Matching

In the first stage, relations between photographs are detected. The same points on multiple photographs are identified for that. In this work, the SURF method [10] was used to define characteristic points on the photographs. Other characteristic points that can be used in the matching are presented in [11].

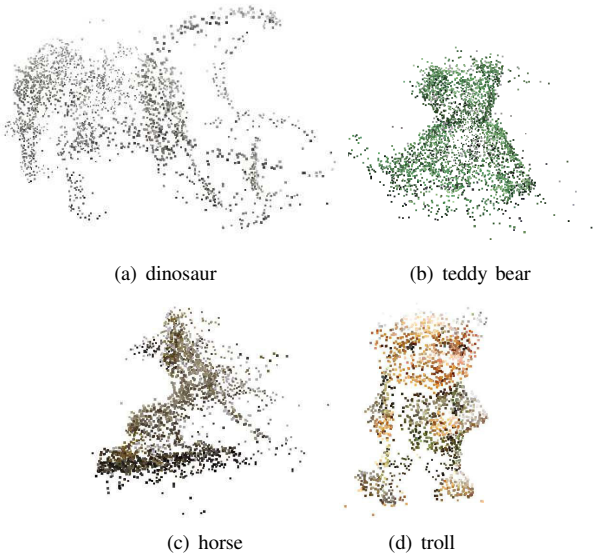
The SURF detector localises characteristic points on the basis of the maximum value of Hessian. For selected points, horizontal and vertical Haar wavelets are calculated to fix an orientation. After these operations, a description of the point is created. The description is invariant from a scale and a rotation.

The matching consists in finding the common description of two points from different images. However, to avoid a false match the following steps are added.

All matches from the background are removed. Such matches are easy to detect, because positions of characteristics points are nearly constant in all photographs in a sequence.

The second group of removed matched is established on the basis of points without a dominant match. Such points have two or more equivalents on the second photography. If any of them is not distinctly better than the rest then all matches that start from this point are eliminated.

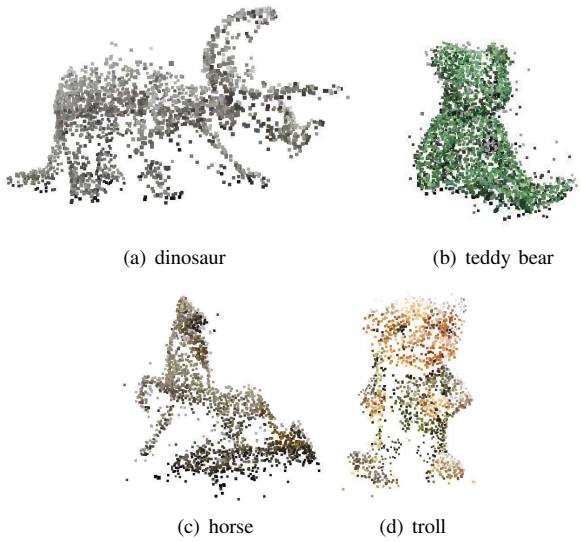
The next condition of the approval match is symmetry. The match between two points should be confirmed by two matching process. In the first process, the first photography is



(a) dinosaur (b) teddy bear

(c) horse (d) troll

Fig. 5. Sparse clouds created in the reconstruction stage



(a) dinosaur (b) teddy bear

(c) horse (d) troll

Fig. 6. Sparse clouds created in the bundle adjustment stage

taken as a source of characteristic points and the second one is area of searching for equivalents. In the second process, roles of photographs swap over.

All created matches should be confirmed by an epipolar model. The RANdom SAmple Consensus (RANSAC) approach [12] is used to detect the fundamental matrix. The matches that are inappropriate for the model are removed.

Finally, each match should be continued at least at three photographs. Each matched point from the second photograph should also be the beginning point for the subsequent math accepted in the matching process computing for the next pair of photographs.

The results of matching for a pair of photographs are presented in Figure 4. As can be seen, only a part of characteristic points have an approval match with a point from the second photograph.

C. Structure reconstruction

In the next stage, positions of detected points in 3D space are calculated as well as camera positions.

In practice, it is not enough to use the epipolar geometry to calculate positions of points separately for each pair of photographs. A calculation of the fundamental matrix is sensitive to noises and a detection of relations and translations between cameras results in errors cumulated in the final cloud.

To reduce the noises, the following solution is proposed.

A cloud of points is initiated by points localised on two first photographs. Each next photograph is used to add new points and calibrate the existing points from the cloud.

Before the calculation, coordinates of points from images are normalised. The normalised point is defined as $K^{-1}\mathbf{x}$, where $\mathbf{x}$ is the coordinate in image space and K is a calibration matrix.

The K matrix is estimated as

$$K = \begin{bmatrix} w+h & 0 & \frac{w}{2} \\ 0 & w+h & \frac{h}{2} \\ 0 & 0 & 1 \end{bmatrix}$$

where w and h are width and height of image respectively. The same estimation was used in [13].

Next, for a selected pair of photographs the eight–point algorithm is used to calculate a fundamental matrix [8]. Owning the fact that coordinates were normalised the essential matrix can be used to detect relations between cameras.

Information about sequential photographs is added iteratively. However, now 2D points from a photograph are compared with 3D points from the created cloud. If a point has an equivalent in the cloud then the 3D coordinates are recalculated on the basis of a new observation. Otherwise, a new point can be added but only if it is present on at least three following photographs.

An important aspect of the reconstruction process is that it uses neither the camera position nor the fact of using a turntable pedestal.

The stage results in a cloud. Examples for analysed objects are given in Figure 5.

D. Bundle adjustment

The created clouds show recognisable views of the modelling objects. However, the density of clouds is not good enough to reconstruct object surface. Moreover, errors from this stage may propagate on the final model. Therefore, an additional stage is necessary to improve a quality of the cloud.

Such method is the bundle adjustment [14]. The method minimises the total mean squared error between real positions of points in a photograph and a position calculated from a 3D projection and a camera position. The algorithm operates on

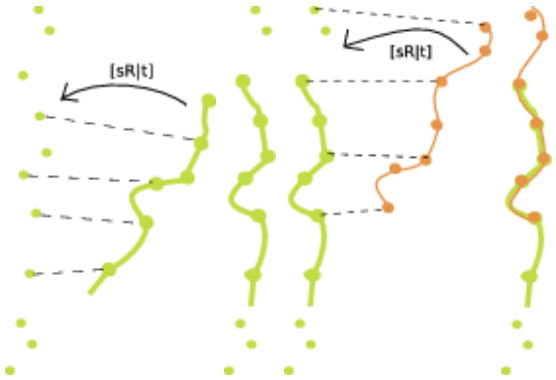


Fig. 7. Connection of the clouds:(a) the projection of the dense cloud on the sparse could, (b) the sparse cloud extended by the dense cloud, (c) the second dense cloud projected on the point of the sparse cloud matched to the first dense cloud, (d) the connected dense clouds.

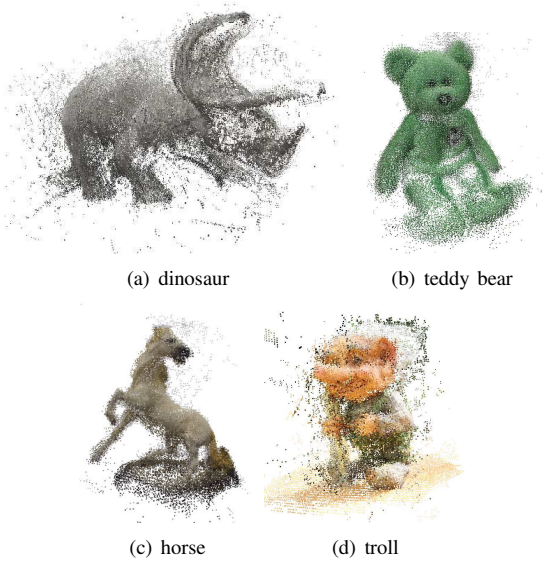


Fig. 8. Dense cloud created in the stage

the cloud calculated in the previous stage and camera parameters calculated using the Levenberg–Marquardt method [14].

The bundle adjustment is a general method. In this work, a specific implementation is used that calculates six parameters for each camera (three for a rotation and three for a translation) and three for each point from the cloud. Inner camera parameters are constant and common for all photographs.

This stage is optional, but definitely improves obtained results. In Figure 6, clouds after the bundle adjustment are presented. In the comparison with the previous clouds, objects are better visualised. However, clouds are still not dense enough to create a final reconstruction.

E. Dense cloud creation

The clouds created in the previous stages are too thin to reconstruct a high–quality surface. However, the clouds can be used to calculate a dense cloud, which will be a base for the final model.

The methodology used in this work bases on stereo block matching algorithms [15]. For a pair of photographs, equivalents of the same objects (pixels or small areas) should be localised on both photographs. A special transformation – rectification allows the algorithm to reduce a searching area to a line of even to a segment (under additional conditions) [16].

Collected data on the equivalents of points are stored as information about a distance between projections of points on a disparity map. With additional information about a camera localisation, the disparity map can be transformed into a depth map. The depth map codes information about the depth in the given point as an intensity.

In the 3D model reconstruction, many depth maps are connected and several problems arise in that process [17]. The maps are calculated for each pair of photographs. A relative small angle between the following photographs determines significant areas common for several maps.

The Iterative Closest Point technique [18] is commonly used to minimise the difference between two clouds of points. The algorithm uses the nearest neighbourhood criterion to find equivalents of an analysed point among clouds. Next, using a mean square cost function the transformation between clouds is solved. An iteration process is used to reduce the calculated cost.

When depth map are selected additional information is given and the connection process can be improved [19]. However, in this work, a new simple method that gives good results is proposed.

In the proposed method, clouds are connected together instead of disparity maps. Therefore, the filtration process plays major role in a quality of the final model. This solution is different that majority of solutions presented in other works, but a similar proposition can be found in [20].

Maps are transformed into dense clouds. Created cloud cannot be connected directly without creation of many noises. Therefore, the dense clouds are fitted in the sparse cloud created in the previous stage. Figure 7 presents the whole process.

The projection of the dense cloud to the sparse clouds is the minimalisation problem:

$$\sum_i ||b_i - \mathbf{s}\mathbf{R}a_i - \mathbf{t}||^2, \tag{1}$$

where a_i is a point from the sparse cloud, b_i an equivalent of the point in the dense cloud. The solution is the transformation that consists of the rotation R, the translation t, the scale s, and minimalises the formula (1).

The results of the stage are given in Figure 8. The created clouds are dense, but noisy. The noises will be removed in the next stage.

F. Surface reconstruction

In the first step of the final stage, the dense cloud is filtered. Several algorithms are used to improve a quality of the cloud:

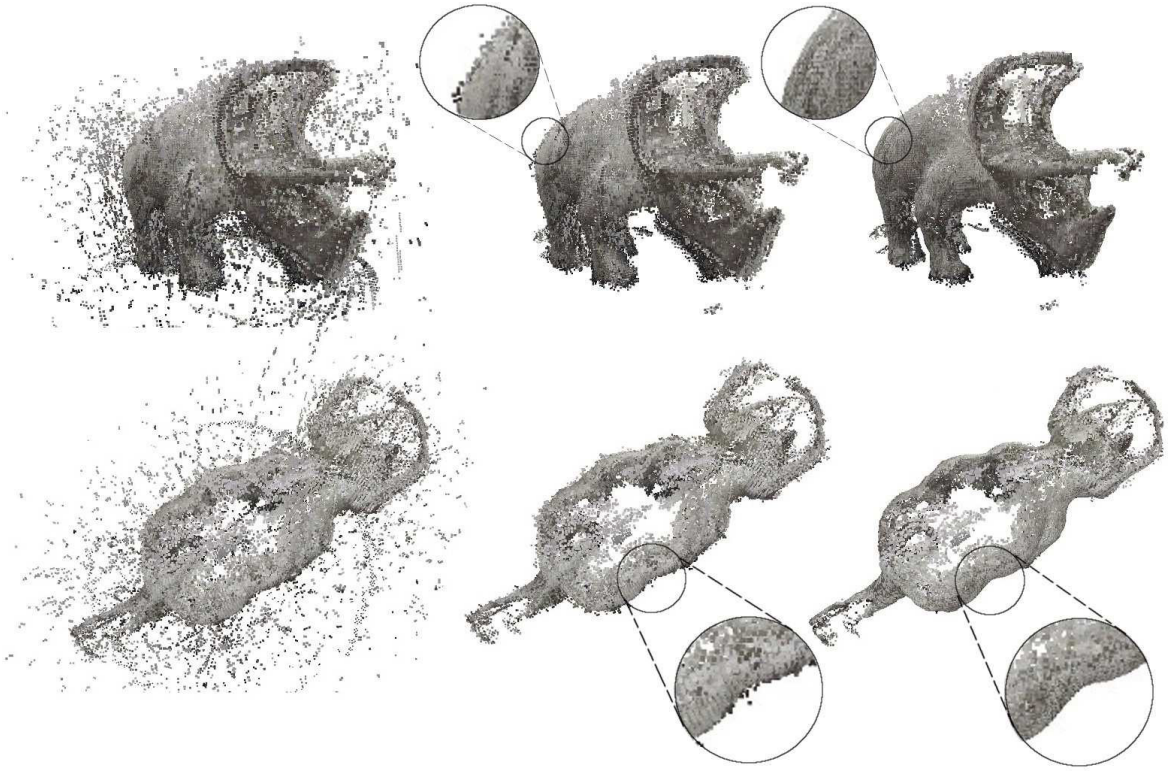


Fig. 9. Filtration. From the left: dense cloud, results of the SOR filter, and results of the MLS method.

the Statistical Outlier Removal filter, Voxel Grid filter, and the Moving Least Squares method.

The disparity maps have wide common areas. Therefore, points from an appropriate surface are located very dense. Otherwise, wrongly reconstructed points are located on random positions usually with small local density. Such points can be removed by the Statistical Outlier Removal filter [21]. For each point in the cloud, the filter calculates distances to k nearest neighbourhoods. On this base, a rejection threshold is calculated. When the average distance to the nearest points exceeds the threshold, a point is removed from the cloud.

The Voxel Grid filter reduces the number of points in the cloud. In a small neighbourhood, all points are reduced to a single point, which is the centroid.

Next, the cloud is smoothed. Noises near an appropriate surface could not be removed by the SOR filter. Therefore, the Moving Least Squares method [22] is used to remove the noises. In the method, a local surface is approximated and points are projected on the surface.

Effects of filtration are presented in Figure 9.

The last step is a surface reconstruction. The step is done by the Poisson method [23]. Before the reconstruction, normals for points should be calculated. The standard method that calculates normals [24] can be used. In the method, the Riemann graph is created with vertices defined by points and edges between nearest points. The graph forms the basis for a propagation of normals orientations.

The Poisson method creates a surface on the basis of the cloud of points with oriented normals. The method solves for an approximate indicator function of the inferred solid, whose gradient best matches the normals. The output scalar function is then iso–contoured using adaptive marching cubes.

IV. Results

The reconstruction time for a single model was about 6 minutes. The reconstructions were done on AMD Athlon 64 3000+ 2 GHz with Ubuntu system version 10.04.

The obtained results are presented in Figure 10. Reconstructed objects have good quality. Colours of triangles are interpolated from colours of corners. The colouring method is simple but brings satisfactory results. However, results show that a quality of the photography documentation influences the quality of the models.

The teddy bear (described by the full, professional documentation) is very well reconstructed including original depressions on its belly and back, while the horse model lost some details. In the dinosaur model (described by the professional, but reduced documentation), not all wrongly reconstructed points were removed. The model has a projection on its back. The troll model (described by the amateur documentation) has a distortion on the back. Moreover, a part of the stand was recognised as a part of the statue. Probably, this interpretation was caused by a shadow registered on the photographs. In a studio, this problem is eliminated.

Fig. 10. Reconstructed objects: Dinosaur, teddy bear, horse, and troll

V. Evaluation

The main issue of the evaluation of created model is a lack of digital patterns to compare with the reconstruction. Therefore, we propose the following schema of models evaluation on the base of the series of photographs.

The projection of model was projected back on the cameras. As a result, the initial two dimensions projection was reconstructed. The reconstructed projection was compared with the isolated object from the origin photograph. Figure 11 presents all elements.

Both reconstructed object and isolated object from the origin photograph was used to create masks. The areas are compared and the model is evaluated using

$$q = \frac{b + c}{s}, \tag{2}$$

where b is the number of reconstructed pixels that are not a part of the origin object, c is the number of pixels from the origin object that are not reconstructed, and s is the number of matching pixels. Figure 12 presents all elements of the area matching.

The second proposed evaluation method is a contour matching. The contour matching analyses a reconstruction of details. For the created masks, the contour is calculated as the morphological gradient with the colour structuring element and with the 11 pixels diameter. Next, two evaluation measures were calculated:

$$q' = \frac{s}{m}, \tag{3}$$

where s is the number of matching pixels for both contours and m is the number of pixels in the model contour;

$$q'' = \frac{s}{o}, \tag{4}$$

where s is the number of matching pixels for both contours and o is the number of pixels in the object contour;

The measure q' (3) defines the percent of coverage of the object contour by the model contour, and the measure q'' (4) defines the percent of coverage of the model contour by the object contour.

Figure 13 presents all elements of the contour matching.

Table I presents evaluation of the models. We calculated averages of the evaluating coefficients for whole series of photographs. For the area matching coefficient q smaller values are better, for the contour matching coefficients higher values are better.

Studios objects have the similar area matching coefficients. The contour matching coefficients show that solid objects (the horse, the dinosaur) have better reconstructed contours than fluffy (teddy bear). The worst results were obtained for the troll model. It was caused by recognising the base of the model as a part of the model. When we edited model manually, we obtained better results.

TABLE I
Evaluation of the models: q - area matching coefficient, q', q'', contour matching coefficients

model	**avg** q	**avg** q'	**avg** q''
dinosaur	0.10	0.68	0.65
horse	0.15	0.72	0.72
teddy bear	0.11	0.57	0.57
troll	0.91	0.33	0.31
edited troll	0.35	0.51	0.47

VI. Conclusions

In this work, the 3D models creating from photographs was presented. The method describes all stages of the reconstruction process from the features detection and the matching, through the creation of a sparse 3D cloud and a dense cloud, until the filtration and the surface reconstruction. All stages were illustrated with examples of their products.

The whole solution was implemented on the basis of open source libraries: OpenCV, Point Cloud Library and Sparse Bundle Adjustment. However, in several points original solutions were used. Especially original approach was used to connect the dense points clouds into one cloud using the sparse cloud.

Although some noises can be observed after a close inspection on the dinosaur model and the troll has some distortions, the obtained coloured reconstructions are good–looking and results are rewarding.

The presented process, together with the created application allows user to create complex 3D models without any expensive staff and advanced software.

References

[1] K. Kolev, T. Brox, and D. Cremers, "Fast joint estimation of silhouettes and dense 3d geometry from multiple images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 493–505, 2012. [Online]. Available: http://lmb.informatik.uni-freiburg.de//Publications/2012/Bro12

[2] D. Zawieska, "Analysis of operators for detection of corners set in automatic image matching," *Archives of Photogrammetry, Cartography and Remote Sensing*, vol. 22, pp. 423–436, 2011.

[3] I. Y. Jang, J.-H. Cho, and K. H. Lee, "3d human modeling from a single depth image dealing with self-occlusion," *Multimedia Tools Appl.*, vol. 58, no. 1, pp. 267–288, 2012.

[4] P. Lavoie, D. Ionescu, and E. Petriu, "3d object model recovery from 2d images using structured light," *Instrumentation and Measurement, IEEE Transactions on*, vol. 53, no. 2, pp. 437–443, April 2004.

[5] S. Gimjumpa, W. Narkbuekaew, M. Sangworasil, and C. Pintavirooj, "3d modeling from multiple projections with arbitrary-posed camera," in *Industrial Electronics and Applications, 2006 1ST IEEE Conference on*, May 2006, pp. 1–5.

[6] Y.-K. Zhang and Y. Xiao, "A practical method of 3d reconstruction based on uncalibrated image sequence," in *Signal Processing, 2008. ICSP 2008. 9th International Conference on*, Oct 2008, pp. 1368–1371.

[7] H. M. Nguyen, B. Wunsche, P. Delmas, C. Lutteroth, and W. van der Mark, "High resolution 3d content creation using unconstrained and uncalibrated cameras," in *Human System Interaction (HSI), 2013 The 6th International Conference on*, June 2013, pp. 637–644.

[8] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004.

[9] J. A. Pierre Moreels, "3D objects on turntable," www.vision.caltech.edu/pmoreels/Datasets/TurntableObjects/, 2006.

[10] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *In ECCV*, 2006, pp. 404–417.

Fig. 11. Evaluation preprocessing: the reconstructed projection, the original photograph, and the isolated object

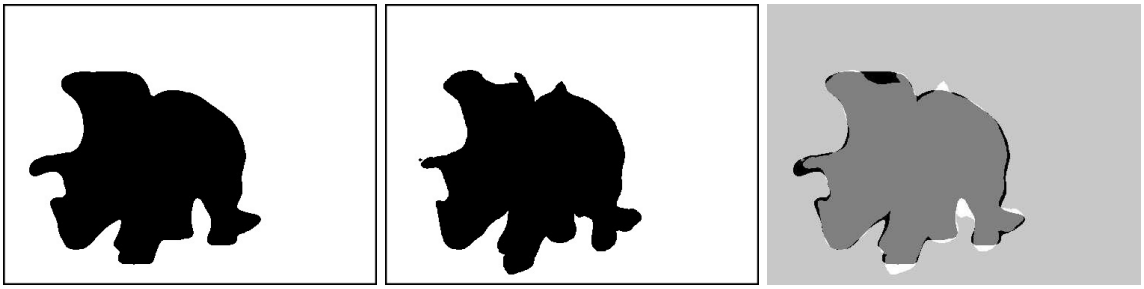


Fig. 12. Area matching: the mask from the origin photograph, the mask of the reconstructed object, and matched areas

.

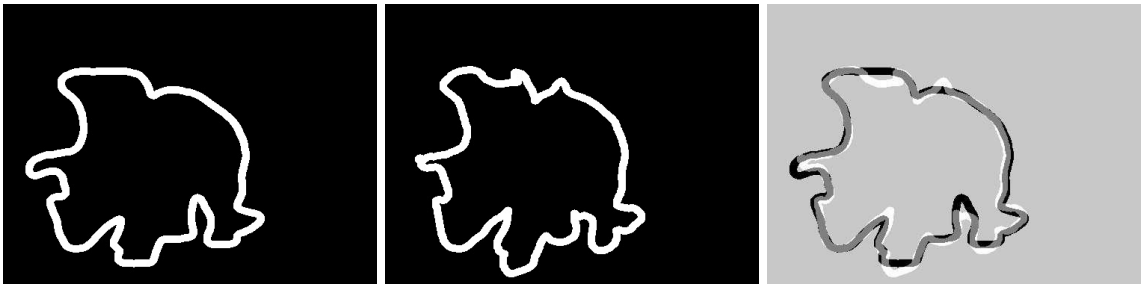


Fig. 13. Contour matching: the contour from the origin photograph, the contour of the reconstructed object, and the matched contours

[11] G. Bagrowski and M. Luckner, "Comparison of corner detectors for revolving objects matching task," in *ICAISC (1)*, ser. Lecture Notes in Computer Science, L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, Eds., vol. 7267. Springer, 2012, pp. 459–467.

[12] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.

[13] M. Pollefeys, "Visual 3d modeling from images," http://www.cs.unc.edu/~marc/tutorial/, course/Turorial notes, presented at Siggraph 2002/2001/2000, 3DIM 2001/2003, ECCV 2000.

[14] M. A. Lourakis and A. Argyros, "SBA: A Software Package for Generic Sparse Bundle Adjustment," *ACM Trans. Math. Software*, vol. 36, no. 1, pp. 1–30, 2009.

[15] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," http://vision.middlebury.edu/mview/, Washington, DC, USA, pp. 519–528, 2006.

[16] B. Cyganek and P. Siebert, *An Introduction to 3D Computer Vision Techniques and Algorithms*. John Wiley & Sons, 2009.

[17] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," http://vision.middlebury.edu/stereo/, pp. 7–42, 2001.

[18] S. Rusinkiewicz and M. Levoy, "Efficient variants of the ICP algorithm," in *Third International Conference on 3D Digital Imaging and Modeling (3DIM)*, Jun. 2001.

[19] D. Bradley, T. Boubekeur, T. Berlin, and W. Heidrich, "Accurate multi-view reconstruction using robust binocular stereo and surface meshing," in *In Proc. of CVPR*, 2008.

[20] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-d point sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, no. 5, pp. 698–700, May 1987. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.1987.4767965

[21] R. B. Rusu, Z. C. Marton, N. Blodow, M. Dolha, and M. Beetz, "Towards 3d point cloud based object maps for household environments," *Robotics and Autonomous Systems Journal (Special Issue on Semantic Knowledge)*, vol. 56, pp. 927–941, 2008. [Online]. Available: http://files.rbrusu.com/publications/Rusu08RAS-Semantic.pdf

[22] M. Alexa, J. Behr, D. Cohen-or, S. Fleishman, D. Levin, and C. T. Silva, "Computing and rendering point set surfaces," *IEEE Transactions on Visualization and Computer Graphics*, vol. 9, pp. 3–15, 2003.

[23] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proceedings of the fourth Eurographics symposium on Geometry processing*, ser. SGP '06, 2006, pp. 61–70.

[24] H. Hoppe, T. DeRose, T. Duchamp, J. Mcdonald, and W. Stuetzle, "Surface reconstruction from unorganized points," in *SIGGRAPH Comput. Graph.*, 1992, pp. 71–78.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 679–686
DOI: 10.15439/2014F195
ACSIS, Vol. 2

Indoor head detection and tracking on RGBD images

Katarzyna Niżałowska, Łukasz Burdka, Urszula Markowska-Kaczmar
Institute of Informatics
Wroclaw University of Technology
Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland
Email: urszula.markowska-kaczmar@pwr.wroc.pl

***Abstract*—A real-time human head detection and tracking method for a fall detection system is presented. It utilizes RGBD images to obtain a head position in the three-dimensional space. The proposed method is designed to be insensitive to a body orientation and requires no initial calibration for the tracked person. The evaluation was performed on the basis of annotated videos with realistic non-studio indoor everyday activities and falls. The proposed method outperforms head tracking from the Microsoft Kinect SDK skeleton tracking.**

I. Introduction

TRACKING of a human head is an important aspect of any system striving to monitor human behavior or health condition. Many conclusions can be made based on the information about a head position and orientation. A certain application that can benefit from a reliable information about a human head position is a fall detection for an elderly people monitoring system. Existing solutions focus mostly on detecting or tracking a human face instead of a head in general. Most of them also takes an assumption about constant vertical orientation of a human body. In a vast majority of situations such an approach is sufficient but in the context of a fall detection system there is no suitable existing solution.

The aim of our research was to develop a robust head detection and tracking method that is capable of tracking a human head regardless of its orientation and independently of a tracked person. The method uses joined color, motion and depth data to effectively perform this task. The introduced method maintain its performance in situations when the head position and orientation change rapidly such as during a fall.

The content of this paper is organized as follows. In section II related works in the field of head tracking are introduced. In section III we formulate the research problem, which our method is designed to solve. In section IV we describe the presented solution. Section V is dedicated to the evaluation of our method and contains the description of the experiment and the dataset followed by test results compared to the Kinect SDK head tracking [1]. In section VI we conclude our work and propose future work directions.

II. Related works

The head tracking problem has been widely studied over the past few years. In the literature, definitions of this problem describe different tasks. The majority of papers identifies the problem of head tracking with face tracking. They only tackle situations when the face is clearly visible on a video image and take the assumption that it is located near the camera, as in [2], [3], [4], [5], [6], [7]. Two most common applications of such defined head tracking are to obtain certain facial features [6], [2], [3], [4], and to approximate a spatial head orientation [5], [8]. In this paper the problem of head tracking refers to determining the position of a head regardless of its rotation around the vertical axis.

Since the information about a human head position and orientation can be utilized in a vast number of applications, there are many different approaches to solve this problem. In this paper we focus on vision systems as most versatile ones. The highest performance can be achieved using a thermal camera [9] as a data source. It is a consequence of a human head being easily distinguishable on thermal images. This solution, however, cannot be widely applied due to the high cost of thermal cameras. A common approach to this problem is using a video camera as a data source. The video camera was utilized in the methods described in [2], [3], [4], [10], [11], [6]. Recent appearance of affordable sensors containing both video and depth camera has exposed new possibilities in the field of image processing. A widely used device, integrating a depth sensor and a color camera is Microsoft Kinect. It is used for the head tracking task in [5] and [8]. Additionally Microsoft Company released SDK for Kinect [1], providing a skeleton tracking functionality. Thanks to this solution, if a skeleton is recognized properly by the Kinect sensor, information about a head position can be easily obtained, however, as shown in this paper, it lacks robustness.

Among vision systems utilizing different data sources, there are various methods solving the head detection problem. A method presented in [12] uses background subtraction to detect a moving silhouette and treats its highest point as a head. In [11] the background subtraction is also used to find interest points. Then, a classifier is applied. In [10] each tracked head must be initially introduced to the tracking system from four directions. In [3] and [7] only a face is detected using a generic Haar cascade face detector [13]. In this case, a face needs to be visible in satisfactory resolution.

After the head is detected, the tracking process can be initiated. Most methods assume an invariant orientation of a head during tracking. Therefore, a template is captured

978-83-60810-58-3/$25.00

once and subsequently SURF [7] or a Template Matching algorithm is used to track the template. Other methods use CAMSHIFT [2] algorithm or intensity gradient and color histograms analysis [10] to perform tracking.

Existing methods fail, when the head is seen from different perspectives or a human body is not arranged vertically, for example during a fall. They also focus only on using a single data source, while composition of color and depth information allows greater versatility of a tracking system.

III. Research problem

The method described in this paper is designed to solve the head detection and tracking problem with an assumption to apply it to an elderly people monitoring system. The detection task is defined as locating a single human head on an image regardless of its rotation around the vertical axis. The tracking task is interpreted as providing consecutive information about the location of the initially detected head in each frame. Given a streaming sequence of color and depth images obtained from a RGBD sensor, the method returns a spatial position of a tracked head or information that there is no head detected.

The returned position is defined in the 3-dimensional coordinate system described in section IV. The detection and tracking is performed in the real time, thus providing the position of a head in the last processed frame. The input stream is analyzed with the speed of fifteen frames per second. The following assumptions are made. The method tracks a single person in an indoor environment. Neither person-specific nor room-specific calibration needs to be performed. The method is sensor-independent, however, for a head to be detected, it should be located within the sensor depth range, which is from 0.8m to 4m for the Kinect. In order to perform tracking successfully, the initially detected head needs to be visible on a color image. It is not required to stay in the depth range. The method should be robust to rapid changes in a position and orientation of a head in situations such as a fall.

IV. Method description

Our method uses image processing techniques and simple decision rules. It utilizes a combined information, extracted from color and depth images, obtained from a RGBD sensor. The method consists of four modules. The interaction between them for the *n*-th frame is presented in the Fig. 1.

The first module is capable of creating a motion image, based on three consecutive color frames. The second module, referred to as Detector, detects a head based on current color, motion and depth frames. The Tracker module also uses current color, motion and depth frames to track the head, detected in the previous frame. In the fourth module, referred to as Integrator, the information about the head position provided by Detector and Tracker is integrated and the spatial position of the head is returned.

An interest point for the head detection is the top of every vertical silhouette, segmented from the depth image. The interest region is marked as a *head* if it is recognized as a face or a movement in this area is detected on the color image.

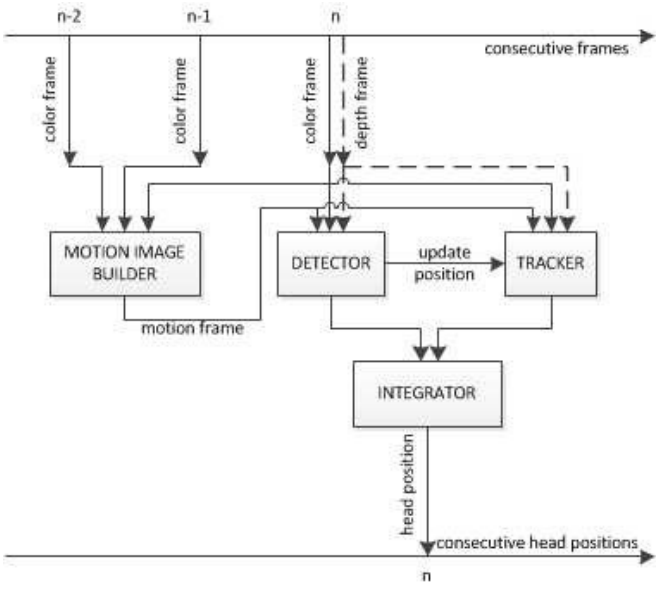


Fig. 1. The interaction between method modules for the *n*-th frame

TABLE I
Sensor-specific parameters

Symbol	Description	Kinect value
c_w, c_h	color image width and height [pixels]	640, 480
d_w, d_h	depth image width and height [pixels]	640, 480
sf	scaling factor	1.06
hd	horizontal displacement of the depth image on scaled color image [pixels]	6
hda	horizontal view angle of the depth camera [degrees]	58.5
vda	vertical view angle of the depth camera [degrees]	45.6
hca	horizontal view angle of the color camera [degrees]	62.0
vca	vertical view angle of the color camera [degrees]	48.6

The tracking is performed independently on the color image and on an image created as an absolute difference between consecutive color frames. This image will be referred to as a motion image. To increase the robustness of the tracking algorithm, both images are multiplied by the depth mask which cuts out the regions where the depth value differs significantly from the last known depth of the tracked head. The results returned by tracking on color and motion images are integrated as a weighted average based on the certainty of each of them.

A. Sensor-specific parameters

Due to a variety of RGBD sensors with different characteristics such as focal parameters, resolution and displacement of color and depth cameras, the presented method is parameterized to make it sensor-independent. Although the method was implemented and tested with the Kinect sensor, no additional Kinect SDK functionalities, such as skeleton tracking, were used to assure portability to other sensors. Defined parameters and their values for the Kinect sensor, described in the Kinect specification, are listed in the Tab. I

Color and depth image resolutions were selected from several options available for the Kinect sensor. The scaling factor was calculated with respect to the ratio of view angles of depth and color camera. The horizontal displacement is the consequence of displacement of cameras in the sensor. View angles are characteristics of the camera specified by a manufacturer.

To enable locating various points on the image in a real-world coordinate system, the information about the view angle of the depth camera and the depth value of the given point are used. The world coordinates are expressed in the right-handed coordinate system consistent with the one specified by Kinect SDK [1]. The origin is located at the center of the sensor, the Z-axis is pointing toward the direction of view and Y-axis points upwards. Whereas the image coordinate system has its origin in the top left corner of the image with the X-axis pointing to the right and the Y-axis pointing downwards. Image coordinate values are expressed in pixels. The real-world spatial coordinate (s_x, s_y, s_z) of a point (x, y) on the image, located at the distance d, measured in meters, is given by equations (1)

$$\begin{aligned} s_x &= 2d(\frac{x}{d_h} - 0.5)tg(\frac{hda}{2}), \\ s_y &= 2d(0.5 - \frac{y}{d_w})tg(\frac{vda}{2}), \\ s_z &= d, \end{aligned} \tag{1}$$

where d_w, d_h, hda, and vda are defined in the Tab. I.

The calculation allows a fast and accurate conversion from the depth image space to real-world coordinates. The choice of such coordinate system is justified by the possibility of comparing the method output with the Kinect skeleton tracking.

B. Preprocessing

The main problem, which needs to be solved in order to allow combining the depth and the color images is mapping between pixels of both images. The displacement of cameras and their different focal characteristics cause the difference between areas visible on both images. Kinect SDK provides an accurate conversion from the depth to color space. This conversion, however, only works in one direction and is only applicable to Kinect sensor. Additionally, it can only map a single depth pixel to color pixel which is computationally ineffective. A method similar to the Kinect solution was proposed in [14]. It is also unidirectional but in a contrast to the previous approach, it does not need Kinect sensor to be plugged in during its usage which allows wider application of this solution. It is only unidirectional since the precise inverse operation is not possible. It would require the analysis of depth pixels to find the one that matches best to the given color pixel. Such a solution would be very computationally complex. Since the existing methods do not use any fast bidirectional space mapping between color and depth images, a fast method of solving this problem is proposed. The color image is scaled and shifted to match the depth image. Given the resolution of the depth image d_w, d_h, the new size of the color image c_w', c_h' is calculated as in formula (2)

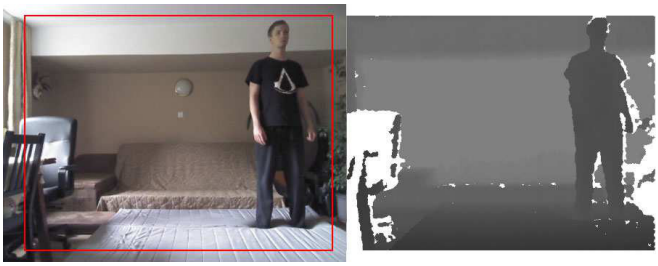


Fig. 2. Color image cropp rectangle

$$\begin{aligned} c'_w &= d_w sf, \\ c'_h &= d_h sf, \end{aligned} \tag{2}$$

where sf is the scaling factor defined in the Tab. I.

Subsequently, the color image is cropped to match the size of the depth image by extracting a sub-image of the size d_w, d_h, which top left corner is located in the point p_x, p_y calculated as in formula (3)

$$\begin{aligned} p_x &= \frac{c'_w - d_w}{2} + hd, \\ p_y &= \frac{c'_h - d_h}{2}, \end{aligned} \tag{3}$$

where d_w, d_h, and hd are defined in the Tab. I. The cropped subimage is presented in the Fig. 2.

Once the transformation is completed, both images are aligned and no further calculations are necessary. The transformation is designed to give the best projection between images. While the method is not as accurate as the one in [14] or Kinect SDK mapping, it is bidirectional and much faster. The error of this method is greater for pixels located near the camera but it is tolerable at the range where the full body is visible.

The motion image is created based on the absolute difference between three consecutive color frames. It is a simplified version of a Motion History Image, described in [21]. A value of the pixel (x, y) is given by the equation (4)

$$\begin{aligned} M_n(x,y) = \ & |2(K_n(x,y) - K_{n-1}(x,y))| \\ & + |(K_{n-1}(x,y) - K_{n-2}(x,y))|, \end{aligned} \tag{4}$$

where $M_n(x, y)$ is the value of pixel (x, y) on the motion image, $K_{n-2}(x, y)$, $K_{n-1}(x, y)$ and $K_n(x, y)$ are values of corresponding pixels on three consecutive color frames. This approach is used to accurately highlight the area where the silhouette is currently located.

C. Head detection

The Detector module utilizes a current color, depth and motion frame and returns lists of *heads* and *uncertain heads* detected as a result for an input frame.

Algorithm 1 Head detection - finding regions of interest

```
image ← Dilatate(image);
image ← Canny(image);
contourList ← Group(image);
for all contour in contourList do
  d ← AvgDepth(contour);
  h ← Height(contour);
  if h ≥ hTreshold then
    rect ← CropV(contour, hHeight);
    median ← Median(rect, contour);
    if minHBreadth ≤ median ≤ maxHBreadth
    then
      regionOfInterest ← CropH(rect, median);
      interestRegions.Add(regionOfInterest);
    end if
  end if
end for
return interestRegions;
```

The head detection is performed in two steps. In the first one, regions of interest are found. An interest region is a sub-image that may contain a head. In the second step, each interest region is labeled either as a *head* or as an *uncertain head*.

Since our methods measures detected objects to rule out those, that cannot be human silhouettes, we define size constraints that need to be satisfied. Corresponding to the anthropometric studies of a human body [15], [16], the average breadth of a human head is 13.9 cm for men and 13.3 cm for women, while its length is 18.0 cm for men and 17.2 cm for women and its height is 21.2 cm for men and 19.8 cm for women. Additionally, according to [17], the height of a human body exceeds 1m after being four years old. Based on the introduced measurements and taking into account various transformations applied to the processed image, we define the following parameters:

- *hHeight* - the height of a head (26.4 cm),
- *minHBreadth* - the minimum width of a head (12 cm),
- *maxHBreadth* - the maximum width of a head (24 cm),
- *hTreshold* - the minimum height of a silhouette (1 m).

The proposed values exceed the top and bottom limit of a size of an adult human in order to avoid an omission of any real head detection.

During the first stage of the head detection, the depth image is analyzed. We use it to detect regions of interest. The aim of its analysis is to find silhouettes of a human-like shape. The process of finding interest regions is presented in the Algorithm 1.

Firstly, the image is repeatedly dilated to rule out the noise and erroneous pixels, which are white pixels, indicating no depth data. It is then binarized by Canny edge detector. Edge points are grouped into contours using a method described in [18]. As a consequence of the dilatation, the average size of a head on the image is increased by approximately 20%. This information is vital since in our method, boundaries for various measurements of a human body are defined in the real-world coordinate system. For each silhouette, its average depth is calculated to determine its distance from the camera. Having the distance, the height of the human silhouette is calculated using formulas (1). Only silhouettes that are at least the height of *hTreshold* are considered as possible human silhouettes, others are instantly eliminated. Subsequently, the bounding box of each silhouette is cropped leaving only its top *hHeight*. The median of the width of the silhouette's part located within a cropped rectangle is calculated. The rectangle is rejected if the calculated median is outside of the range from *minHBreadth* to *maxHBreadth*. Otherwise, the rectangle is cropped horizontally to obtain the interest region of a width equal to the introduced median.

Algorithm 2 Head detection - head classification

```
for all region in interestRegions do
  if FaceDetected(region) then
    heads.Add(region);
  else
    if MovementDetected(region) then
      heads.Add(region);
    else
      uncertainHeads.Add(region);
    end if
  end if
end for
```

After finding regions of interest, each of them is labeled as a *head* or an *uncertain head*. The classification process is shown in the Algorithm 2.

A simple decision sequence is used for this labeling task. Initially a Haar cascade is used to detect a face in the given region. We use the cascade model for frontal face recognition provided by OpenCV [19]. If the face is detected, the region is labeled as a *head*. Otherwise, the occurrence of the movement is checked in the interest region. If the movement is detected, the region is labeled as a *head*. Otherwise it is labeled as an *uncertain head*. To perform the movement detection, the motion image is thresholded to filter out the noise and the erosion is performed to clear the stronger noise. Subsequently the image is transformed by a multiple dilatation to highlight the movement region. The process is illustrated in the Fig. 3. If the region being an *uncertain head* contains at least 20% of white pixels we treat it as a *head*.

This approach has the following justification. The situation when there is a vertical silhouette detected, which contains an object of a size matching a human head in its top part, is not sufficient to classify this object as a head. However, because a human is not able to stand still without any movement, the movement is a reasonable indicator of a human presence.

D. Head tracking

The Tracker module takes a current color, depth and motion frame as an argument and, using the information about the previously detected head, performs tracking and returns up to

Fig. 3. Color image (left), motion image (center), and post-processed motion image (right)

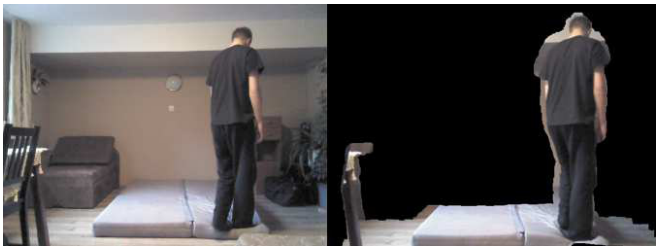


Fig. 4. Color image (left) and masked color image (right)

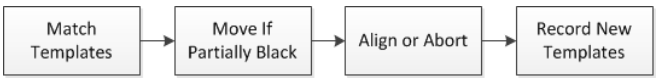


Fig. 5. Single tracking step for one image

4 possible head positions together with their *certainty factors*.

Tracking is initiated when an area labeled as a *head* is detected. It is performed independently on the color image and on the motion image introduced in section IV. Both images are multiplied by a binary mask cutting off areas where the depth value is significantly different from the depth value of the head. Such a mask is referred to as a depth mask.

In order to obtain the depth mask for each frame, auxiliary tracking is performed on the depth image. When the head is detected, a sub-image containing the head is recorded as a template. In the subsequent frame, an attempt is made to match the recorded template on the image. The search is performed in the region of interest, specified as the area containing the head in the previous frame, extended in each direction by a certain margin. The margin is designed to cover the distance that a head can traverse during the time of one frame. It is expressed in meters and transformed into pixels using the distance from the head to the sensor. The width of the margin is a parameter of the method and should be adjusted to match needs of the application. In our case, the video's frame rate is equal to 15 frames per second, thus the time span between consecutive frames is 66ms. For the fall detection system, where the speed of the head can be high, the reasonable margin value is 0.5m.

The template matching within the region of interest is performed using the Normalized Cross Correlation method (NCC), described in [20]. Once the best match is found, the template is updated with the newly found region, keeping its original size. The method returns a certainty measure to describe the quality of the match. This value is recorded together with a template position and will be referred to as a *certainty factor*. In the next step, the depth of the head is calculated and the depth mask is created as a binary image. The image has a black background and contains white pixels only in the regions, where the difference of the depth value and the depth of the head is not greater than the margin value.

Subsequently, the color and the motion images are multiplied by the depth mask. As a result of the multiplication, the areas where the depth value differs from the depth of the head are black, and those with a similar depth value are left unchanged. The example of a depth masking is shown in the Fig. 4.

Once masking is done, tracking on color and motion images is performed. For both images the same technique is used. The single step of tracking on one image is presented in Fig. 5. For each image, two templates are recorded. One having the size of the detected head and the other, located centrally within the first one and having a half of its size. It has been noticed that different sizes of the template perform better under different conditions, therefore tracking is executed independently for two templates. NCC is used to find the best template match. Next, the correction of the determined template location is done. In case a part of the template contains a large number of black pixels, its location is moved in the opposite direction.

If the best match of the template is located entirely in the black area, its position is aligned to the second tracked template. If this situation occurs to both templates simultaneously, tracking on the image is terminated.

Tracking is executed in parallel to the head detection process for each frame. If the *head* is detected within the range of the margin value of any of tracked templates, the size and position of each tracked template, including the auxiliary depth template, is updated with the area of the newly detected head. If no *head* is detected but there is an *uncertain head* in the rectangle overlapping one of the tracked templates, it is considered a *head* and all templates are updated as stated above.

After tracking is ceased, the positions of the tracked templates are returned together with their *certainty factors*.

E. Result integration

The Integrator module is capable of integrating the results returned by the Detector and the Tracker. The output of this module is a 3-dimensional position of a single head or information, that there was no head detected.

If any *head* was detected by the Detector, the results of the Tracker are ignored and the detected head is treated as a final result. Otherwise, the results of tracking are integrated in the following manner. In each frame there are two templates tracked on the color image and additional two on the motion image, therefore tracking results need to be integrated in order to provide a single location of the head. The location of the head is determined as a weighted average of all template positions. A *certainty factor* is used as a weight. This approach

Fig. 6. The example of a color frame (left) and a depth frame (right) from the dataset

causes inaccurate matches having less impact on the final outcome. Only templates matched on the color and the motion images are used in the integration process, because tracking on the depth image tends to be less accurate.

In order to obtain the spatial location of the tracked head, the formula (1) is used with the head position on the image and its depth value. It is then returned as a result of this module and the whole method.

V. Evaluation

The evaluation of the method was performed to assess its effectiveness and to compare it to the existing solution, implemented in the Kinect SDK. Since there is no RGBD benchmark dataset for the head tracking problem, the proper dataset was prepared. The dataset is described in details in the next subsection. Subsequently, the evaluation procedure, including experimental method and description of used measures, is presented. Finally, the results for our method and the Kinect SDK are shown and commented.

A. Dataset description

The dataset consists of 480 short films, recorded at 15 fps and containing video and depth images for each frame. All scenes are recorded in the indoor scenery and last from 6s to 24s, averagely 13s. Each scene shows one of two actors: a man or a woman. Each film presents either a daily action or a fall. The following actions were recorded: walking, standing, sitting on a chair, sitting on the ground, bending down, lying on a bed, lying on the ground, standing on a chair, cleaning, falling forward, falling backward and falling sideward.

Each action was recorded 20 times per actor and, excluding lying on a bed, contains records, where the actor was viewed from the front, back, left and right. Each film begins showing the empty room and captures the moment, when the person enters the frame and, except falls, finishes after the person leaves the frame. Films showing falls end when a person is lying on the ground. The example frame is shown in Fig. 6.

The dataset was annotated with a current head position to enable the evaluation of a head detection method, however, due to laboriousness of the frame annotation process, for each film, only 3 frames were annotated. They were located at 1/4, 1/2 and 3/4 of the film duration. The first and the last frame was not taken into account, because a vast majority of them was showing only the empty room. The annotation was performed

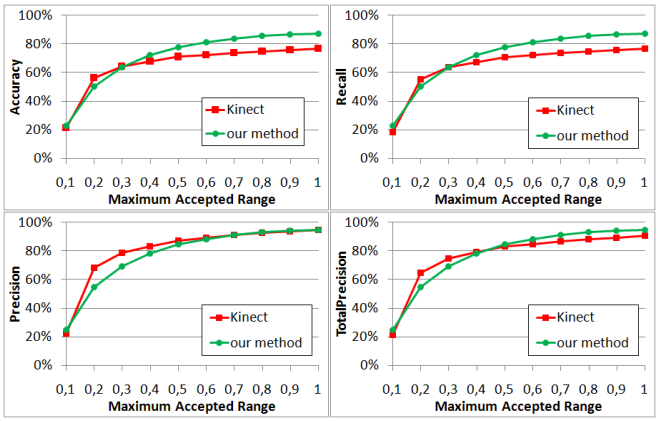


Fig. 7. Accuracy, Recall, Precision and TotalPrecision as a function of accepted range

by one person and validated by another one. On each annotated frame, the head position is specified or the frame was labeled as containing no head. The head position was labeled as a single point on the color image, located in the center of a head and transformed, using the depth information, into the real-world coordinate system. Together the 1440 frames are annotated in 480 films. In 1350 of them, there was a head visible in the frame, and in 90 of them, there was no head visible.

To compare our method to the method from the Kinect SDK, the skeleton information obtained from the Kinect sensor is also recorded for each frame.

B. Test method

Both methods were evaluated using various measures to provide a more comprehensive analysis of their performance. They were then compared to recognize differences in their functioning.

In the first step of the evaluation process, each film was replayed and analyzed by our method to detect and track the head in the real time. In each film, 3 frames, that are annotated with the head position, were evaluated. The number of heads annotated, tracked by our method and tracked by the Kinect SDK was stored for each frame. Additionally, if both numbers of annotated and tracked heads were greater or equal to 1, the following distances were calculated:

- the 3-dimensional Euclidean distance between tracked and annotated head, measured in meters, presented in formula (5),
- the distance between vertical components of positions of both heads, measured in meters, calculated as in (6).

$$d = \sqrt{(t_x - a_x)^2 + (t_y - a_y)^2 + (t_z - a_z)^2}, \quad (5)$$

$$d_v = \sqrt{(t_y - a_y)^2}, \quad (6)$$

where d is the 3-dimensional distance, (t_x, t_y, t_z) is the position of a tracked head in the real-world coordinate system, (a_x, a_y, a_z) is the position of an annotated head in the real-world coordinate system and d_v is a distance between vertical components of those points.

In order to consider different precision requirements, various measures were calculated as functions of the maximal distance between the annotated and the tracked head to treat it as detected correctly. The maximal distance is referred to as a range and extends from 10 cm to 1 m with 10 cm intervals.

The measures presented in the formula (7) were calculated for both our method and the Kinect SDK method

$$\begin{aligned} Accuracy(r) &= \frac{TP_r + TN}{AF}, \\ Recall(r) &= \frac{TP_r}{AH}, \\ Precision(r) &= \frac{TP_r}{DH}, \\ TotalPrecision(r) &= \frac{ATP_r}{ADH}, \end{aligned} \tag{7}$$

where: r is the range, the measure is calculated for, TP_r is a number of frames, in which the head was detected within a range r, TN is the number of frames, correctly classified as containing no head, AF is the number of annotated frames (1440), AH is the number of frames, annotated as containing a head (1350), DH is thenumber of frames, in which at least one head was detected, ATP_r is the number of heads, detected within a range r, including multiple heads detected in one frame, and ADH is the number of all detected heads, including multiple heads detected in one frame.

Due to the fact, that our method is designed to track one head for each frame and Kinect SDK can track higher number of heads, to calculate *Accuracy*, *Recall* and *Precision* we choose one head tracked by the Kinect, closest to the head annotated in this frame, and compare it to the result of our method. Only while calculating the *TotalPrecision*, we take into account all the heads tracked by the Kinect independently, even if there were more than one head in a given frame.

C. Experimental results

In this subsection, results of the experiment described previously are presented and commented. Fig. 7. shows the comparison of *Accuracy*, *Recall*, *Precision* and *TotalPrecision* calculated for our method and for the Kinect SDK in the function of the range.

As it can be seen, the *Accuracy* and *Recall* of our method is mostly higher than of Kinect, except for the range of 20 cm while the *Precision* of Kinect is greater than ours considering the range from 20 cm to 60 cm. For the rest of ranges, *Precision* of both methods is comparable. However, the *TotalPrecision* calculated for all heads detected by the Kinect is only greater in the range of 20 cm to 40 cm, while in the range wider than 50 cm, our method outperforms the Kinect. It is also notable, that considering the range lower than 10 cm, every measure is slightly higher for our method. This can be caused by the fact, that Kinect tends to detect the head on the chin level or even on the neck, while the dataset was annotated with points located in the center of the head, approximately on the line of ears and eyes. The significant difference between the *Precision* and *TotalPrecision* of Kinect indicates the high number of heads detected wrong while there was another head detected correctly. It is important to highlight that in situations where the Kinect detected more than one head, only the best one was used to calculate *Accuracy*, *Precision* and *Recall* while the rest was ignored. The higher *Accuracy* and *Recall* of our method indicate that Kinect detects less heads but when it does, it tends to be more precise.

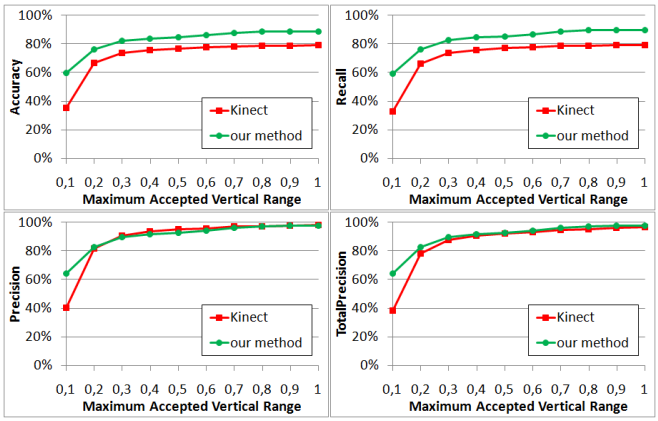


Fig. 8. Accuracy, Recall, Precision and TotalPrecision as a function of accepted vertical range

Regarding the fact, that our method is designed to match needs of a fall detection system, the correct vertical coordinate of a detected head is the one crucial for the proper functioning of such system. To asses both methods in terms of application in a fall detection system, the measures described by formulas (7) were calculated considering only the vertical component of the distance between the annotated and the tracked head. Results are presented in Fig. 8.

When only the distance between Y-coordinates is taken into account, *Accuracy* and *Recall* of our method is significantly greater than *Accuracy* and *Recall* of the Kinect. The *Precision* and *TotalPrecision* are also much greater considering the range of 10 cm and comparable for the rest of ranges.

VI. Conclusions

The above results lead to the conclusion that our method is more suitable for the application in a fall detection system than the method from the Kinect SDK. Among other possible applications, our method can be recommended for those, where the *Accuracy* and *Recall* are more important than the high *Precision*.

A. Summary

The method of human head detection and tracking on RGBD images was presented. As the evaluation shows, it outperforms the Kinect SDK skeleton tracking. Furthermore the method is independent of a used sensor and therefore its usage is not limited to the Microsoft Kinect. Promising

evaluation results indicate that it is a valuable head tracking method that can be successfully applied to tracking a single person in the indoor conditions. Our solution is particularly useful when there is no annotated dataset that could be used for any other machine learning approach since it requires no initial training and can be used to track any person without the necessity of previous calibration. Furthermore it is suitable for the fall detection task as it maintains its effectiveness during a fall. Therefore it can be used either as a primary data source for a newly developed fall detection system or as an auxiliary tracking method to boost the robustness of an existing fall detection system.

B. Future works

During the development and evaluation of our method, we identified various improvements that could potentially increase its effectiveness. During the classification of interest regions recognized as *uncertain heads*, a machine learning approach can be used to decide if the region should be classified as a *head*. Such a solution would require annotated objects, which are not heads, as negative examples and use them together with positively annotated heads to train a classifier. For that approach to be effective, the size of the dataset should be greater than the one used during the development of our method. The necessity to train a classifier would decrease the assumed versatility of our method but could improve its effectiveness in a specific target scenario. Potentially useful features for the classification task would be shape descriptors and color histogram components.

Another promising improvement is the dynamic adjustment of the margin value used during the tracking process to define the area where the search for a tracked head is performed. At present, the size of the margin is fixed to a reasonable value of 0.5m, which is suitable for the fall detection task. This value however could be adjusted based on the current speed of the tracked head.

Our method was designed to solve the head detection and tracking problem for a single person in the room, however it can easily be extended to track any number of heads. Such a modification would require developing a method of matching tracked and detected heads to allow integrating positions of corresponding heads.

Even though the presented configuration of our method has proven to be effective, its further improvements are still a subject of research. Since the method is flexible and divided into modules, each module can be used, modified and adjusted to fit special requirements independently.

REFERENCES

[1] Microsoft. Kinect SDK reference. [Online]. Available: http://msdn.microsoft.com/en-us/library/hh855347.aspx

[2] G. Bradski, "Real time face and object tracking as a component of a perceptual user interface," in *Proceedings Fourth IEEE Workshop on Applications of Computer Vision. WACV '98*, Oct 1998, pp. 214–219. [Online]. Available: http://dx.doi.org/10.1109/ACV.1998.732882

[3] Y. Li, H. Ai, C. Huang, and S. Lao, "Robust head tracking based on a multi-state particle filter," in *7th International Conference on Automatic Face and Gesture Recognition. FGR 2006*, April 2006, pp. 335–340. [Online]. Available: http://dx.doi.org/10.1109/FGR.2006.96

[4] P. Fieguth and D. Terzopoulos, "Color-based tracking of heads and other mobile objects at video frame rates," in *Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun 1997. ISSN 1063-6919 pp. 21–27. [Online]. Available: http://dx.doi.org/10.1109/CVPR.1997.609292

[5] S. Li, K. N. Ngan, and L. Sheng, "A head pose tracking system using RGB-D camera," in *Proceedings of the 9th International Conference on Computer Vision Systems*, ser. ICVS'13. Berlin, Heidelberg: Springer-Verlag, 2013. ISBN 978-3-642-39401-0 pp. 153–162. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-39402-7_16

[6] M. Weber, W. Einhauser, M. Welling, and P. Perona, "Viewpoint-invariant learning and detection of human heads," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 20–27. [Online]. Available: http://dx.doi.org/10.1109/AFGR.2000.840607

[7] J. Choi, Y. Dumortier, S.-I. Choi, M. Ahmad, and G. Medioni, "Real-time 3-D face tracking and modeling from awebcam," in *IEEE Workshop on Applications of Computer Vision (WACV)*, Jan 2012. ISSN 1550-5790 pp. 33–40. [Online]. Available: http://dx.doi.org/10.1109/WACV.2012.6163031

[8] F. Kondori, S. Yousefi, H. Li, S. Sonning, and S. Sonning, "3D head pose estimation using the kinect," in *International Conference on Wireless Communications and Signal Processing (WCSP)*, Nov 2011, pp. 1–4. [Online]. Available: http://dx.doi.org/10.1109/WCSP.2011.6096866

[9] W. K. Wong, Z. Y. Chew, C. K. Loo, and W. S. Lim, "An effective trespasser detection system using thermal camera," in *Second International Conference on Computer Research and Development*, May 2010, pp. 702–706. [Online]. Available: http://dx.doi.org/10.1109/ICCRD.2010.161

[10] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," in *Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun 1998. ISSN 1063-6919 pp. 232–237. [Online]. Available: http://dx.doi.org/10.1109/CVPR.1998.698614

[11] V. Subburaman, A. Descamps, and C. Carincotte, "Counting people in the crowd using a generic head detector," in *IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, Sept 2012, pp. 470–475. [Online]. Available: http://dx.doi.org/10.1109/AVSS.2012.87

[12] M. Munaro, F. Basso, and E. Menegatti, "Tracking people within groups with RGB-D data," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2012. ISSN 2153-0858 pp. 2101–2107. [Online]. Available: http://dx.doi.org/10.1109/IROS.2012.6385772

[13] P. I. Wilson and J. Fernandez, "Facial feature detection using Haar classifiers," *J. Comput. Sci. Coll.*, vol. 21, no. 4, pp. 127–133, Apr. 2006. [Online]. Available: http://dl.acm.org/citation.cfm?id=1127389.1127416

[14] N. Burrus. (2011, Aug.) Kinect calibration. [Online]. Available: http://burrus.name/index.php/Research/KinectCalibration

[15] A. Poston, *Static adult human physical characteristics of the adult head*. Washington DC, USA: Department of Defense Human Factors Engineering Technical Advisory Group, 2000. [Online]. Available: http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA467401

[16] Z. Zhuang and B. Bradtmiller, "Head-and-face anthropometric survey of U.S. respirator users," in *Journal of Occupational and Environmental Hygiene*, vol. 2. ACGIH, 2005, pp. 567–576. [Online]. Available: http://www.nap.edu/html/11815/Anthrotech_report.pdf

[17] C. for Disease Control and Prevention, *Anthropometric reference data for children and adults: United States, 2003-2006*. USA: National Center for Health Statistics US Department of Health and Human Services, 2008. [Online]. Available: http://www.cdc.gov/nchs/data/nhsr/nhsr010.pdf

[18] S. Suzuki and K. be, "Topological structural analysis of digitized binary images by border following," *Computer Vision, Graphics, and Image Processing*, vol. 30, no. 1, pp. 32 – 46, 1985. [Online]. Available: http://dx.doi.org/10.1016/0734-189X(85)90016-7

[19] A. F. Reimondo. Haar cascades. [Online]. Available: http://alereimondo.no-ip.org/OpenCV/34/

[20] K. Briechle and U. D. Hanebeck, "Template matching using fast normalized cross correlation," in *Proc. SPIE*, vol. 4387, 2001, pp. 95–102. [Online]. Available: http://dx.doi.org/10.1117/12.421129

[21] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, Mar 2001. [Online]. Available: http://dx.doi.org/10.1109/34.910878

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 687–694

DOI: 10.15439/2014F42
ACSIS, Vol. 2

High quality, low latency in-home streaming of multimedia applications for mobile devices

Daniel Pohl*, Stefan Nickels†, Ram Nalla*, Oliver Grau*
* Intel Labs
Intel Corporation
Email: {Daniel.Pohl|Ram.Nalla|Oliver.Grau}@intel.com
† Intel Visual Computing Institute
Saarland University
Saarbrücken
Email: nickels@intel-vci.uni-saarland.de

***Abstract*—Today, mobile devices like smartphones and tablets are becoming more powerful and exhibit enhanced 3D graphics performance. However, the overall computing power of these devices is still limited regarding usage scenarios like photo-realistic gaming, enabling an immersive virtual reality experience or real-time processing and visualization of big data. To overcome these limitations application streaming solutions are constantly gaining focus. The idea is to transfer the graphics output of an application running on a server or even a cluster to a mobile device, conveying the impression that the application is running locally. User inputs on the mobile client side are processed and sent back to the server. The main criteria for successful application streaming are low latency, since users want to interact with the scene in near real-time, as well as high image quality. Here, we present a novel application framework suitable for streaming applications from high-end machines to mobile devices. Using real-time ETC1 compression in combination with a distributed rendering architecture we fully leverage recent progress in wireless computer networking standards (IEEE 802.11ac) for mobile devices, achieving much higher image quality at half the latency compared to other in-home streaming solutions.**

I. Introduction

THE ADVENT of powerful handheld devices like smartphones and tablets offers the ability for users to access and consume media content almost everywhere without the need for wired connections. Video and audio streaming technologies have dramatically evolved and have become common technologies. However, in scenarios where users need to be able to interact with the displayed content and where high image quality is desired, video streaming derived technologies are typically not suitable since they introduce latency and image artifacts due to high video compression. Remote desktop applications fail when it comes to using 3D graphics applications like computer games or real-time visualization of big data in scientific HPC applications. Enabling these applications over Internet connections suffers significantly due to restrictions induced by the limits of today's Internet bandwidth and latency. Streaming those in local networks, commonly referred to as in-home streaming, still remains a very challenging task in particular when targeted at small devices like tablets or smartphones that rely on Wi-Fi connections.

In this paper, we present a novel lightweight framework for in-home streaming of interactive applications to small devices, utilizing the latest developments in wireless computer networking standards (IEEE 802.11ac [1]) for mobile devices. Further, we use a distributed rendering architecture [2] in combination with a high-quality, hardware-accelerated decompression scheme utilizing the capabilities of modern handheld devices, resulting in much higher image quality and half the latency compared to other streaming solutions.

The setup is shown in Figure 1. The main application is running on a server or even a group of servers. Via network connection, the graphical output of the server application is streamed to a client application running on a mobile device. In addition, a back channel connection is present that collects user input events on the client and sends it back to the server. The server reacts to this input and produces an updated image, which is then transferred back and displayed at the client. The Quality of Experience is determined mainly by two factors: firstly, the delay between a user input issued on the client and the server-provided graphics refresh displayed at the client should be as low as possible. Secondly, the graphics quality of the streamed application on the client side should be as high as possible, even during scenarios with high motion.

II. Related Work

In this section we give an overview of known streaming technologies and applications, which we separate into three classes.

Classical desktop sharing and terminal applications: Examples are Microsoft's Remote Desktop Connection [3] or VNC (Virtual Network Computing) [4]. These are optimized for typical 2D applications like text processing or spreadsheet

978-83-60810-58-3/$25.00

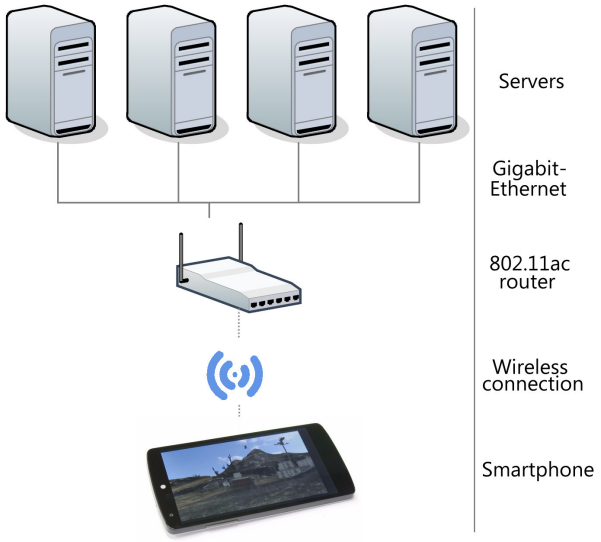


Fig. 1. *Distributed rendering, in-home streaming setup targeted at mobile devices. The four servers are rendering an image. Over Gigabit-Ethernet they transport it to a router with support for IEEE 802.11ac. The router sends the image data wirelessly to the client device (smartphone) which displays it.*

calculations. 3D support is typically very limited and, if supported, not capable to cope with the demands of real-time 3D games and visualizations.

Cloud gaming: The second class of streaming technologies has emerged from the field of computer gaming. Popular commercial solutions like Gaikai [5] or OnLive [6] aim at streaming applications, mainly games, via Internet connection from a cloud gaming server to a user's desktop. There are also open source approaches like Gaming Anywhere [7]. All of these are specifically optimized for usage with Internet connections and typically rely on the H.264/MPEG-4 AVC [8] video codec for streaming graphics. Gaikai and OnLive require a 3-5 Mbps internet connection at minimum and end-to-end latency values are at best 150 ms under optimal conditions (c.f. [7]). OnLive uses a proprietary hardware compression chip in dedicated gaming servers hosted by OnLive. Gaikai's approach has meanwhile been integrated into the PlayStation 4 console [9] after Sony aquired the company and has been renamed to "PlayStation Now" [10]. The service is limited on both the client and server side to dedicated hardware. In general, cloud gaming approaches are not optimized for in-home streaming, sacrificing image quality for lowering network traffic and reducing processing latency.

Dedicated in-home streaming: The third category are approaches designed for delivering applications to other devices in a local network using wired or wireless connections. Recently, Valve's game distribution platform Steam [11] introduced in-home streaming support in a beta version. Nvidia released a portable game console named Shield [12] in 2013, based on current mobile device hardware running an Android operating system. From a PC a game can be streamed to the console. Both approaches rely again on the H.264 codec and are exclusively capable of streaming games. In addition, Nvidia Shield is bound to Nvidia graphics cards and mobile platform architectures. A further streaming approach, which also handles in-home usage is Splashtop [13]. It can stream any application, game or the complete desktop from a single machine using H.264 compression.

We are not considering solutions like Miracast [14], Intel's WiDi [15] or Display as a Service [16] as these are aimed at replicating pixels to another display, not at streaming interactive applications. Further we exclude approaches like Games@Large [17] that stream hardware-specific 3D function calls that are usually not compatible with mobile devices.

III. System

In the following, we will propose a new framework for streaming applications from one or many high-end machines to a mobile device. Using a hardware-enabled decompression scheme in combination with a distributed rendering approach, we fully utilize the potential of recent progress in wireless computer networking standards on mobile devices. With this setup we achieve higher image quality and significantly lower latency than other established in-home streaming approaches. We first give an overview of the hardware setup used in our approach. Then we explain our decision on the compression scheme we used and after that we talk about the details of our software framework and application setup.

Our general system setup is depicted in Figure 1. The graphical output of a server application is streamed to a mobile device, in this case a smartphone, running the client mobile app. The server side consists of four machines in a distributed rendering setup. All devices operate in the same LAN. The server machines are connected by wire to a router which additionally spans a Wi-Fi cell to which the smartphone is connected.

A. Hardware Setup

Here, we describe the hardware specifics of our servers, the client and the network devices.

Our distributed rendering setup consists of four dual-socket workstations using the Intel Xeon X5690 CPUs (6 cores, 12 threads, 3.46 GHz) and the Intel 82575EB Gigabit Ethernet NIC. The client is a LG Nexus 5 smartphone which uses the Snapdragon 800 CPU (4 cores, 2.3 GHz) and the Broadcom BCM4339 802.11ac wireless chip. The devices are connected together through a Netgear R6300 WLAN Gigabit Router. The servers use wired Gigabit Ethernet to connect to the four Ethernet ports of the router. The smartphone connects wirelessly over 802.11ac (1-antenna setup).

B. Compression Setup

In this section we explain why we have chosen the Ericsson Texture Compression format (ETC1) [18] over other commonly used methods.

First we have a look at how displaying of streamed content is usually handled on the client side. Using the popular video library FFmpeg [19] and the H.264 codec an arriving stream at the client needs to be decoded. Using a CPU-based pipeline the decoding result is an image in the YUV420 [20] color space. As this format is usually not natively supported for displaying, the data is converted into the RGB or RGBA format. From there, the uncompressed image data will be uploaded to the graphics chip to be displayed.

If a hardware H.264 decoder is available then the arriving stream needs to be converted into packets, suited for that hardware unit and uploaded to it. The decoding process is started over a proprietary API and usually acts as a black box. Some decoders only handle parts of the decompression procedure; others do the full work and offer an option for either directly displaying the content or sending it back into CPU memory. Hardware H.264 decoders are usually optimized to enable good video playback, but not specifically for low-latency.

Next we have a look at our approach on displaying streamed content. An important feature of modern mobile device GPUs (supporting OpenGL ES 1.0 or higher) is that they have native support for displaying ETC1 textures. Therefore once an ETC1 compressed image arrives at the client we can directly upload it to the graphics chip where decoding to RGB values happens. Given the fixed compression ratio of ETC1 of 1:6 for RGB data the required transfer to the graphics chip is lower compared to uploading uncompressed RGB or RGBA data as described in the CPU-based pipeline for H.264.

ETC1 does an image by image (intra-frame) compression instead of using information across multiple frames (inter-frame). Therefore even if there is a lot of motion between frames a robust image quality is guaranteed. The video codec MJPEG [21] also has this characteristic, but as it lacks hardware decompression support on mobile devices it is not suited as it still requires non-accelerated decompression and the more bandwidth-intensive upload of uncompressed RGB/RGBA pixels to the GPU. Nevertheless, when comparing the image quality of an intra-frame with an inter-frame approach (like H.264) at the same bit rate, the latter will usually be of higher quality. However, codecs with inter-frame compression usually have higher latency.

C. Software Setup

Here, we describe the software setup and the communication between the client and server to enable streamed, distributed rendering.

The Microsoft Windows 7, 64-bit servers are running our custom written HPC ray tracing software, partly accelerated by Intel Embree [22] and multi-threaded through Intel Cilk Plus [23]. The ray tracer can be given the task to only render certain regions of the complete image. The ray tracer hands over the image section to the streaming module of our framework. This module can compress image data into the ETC1 format by using the etcpak library [24] which we multi-threaded using Intel Cilk Plus. Compressed data can be sent to the client using TCP/IP, supported by libSDL_net 2.0 [25]. Further, the server listens on a socket for updates that the client sends.

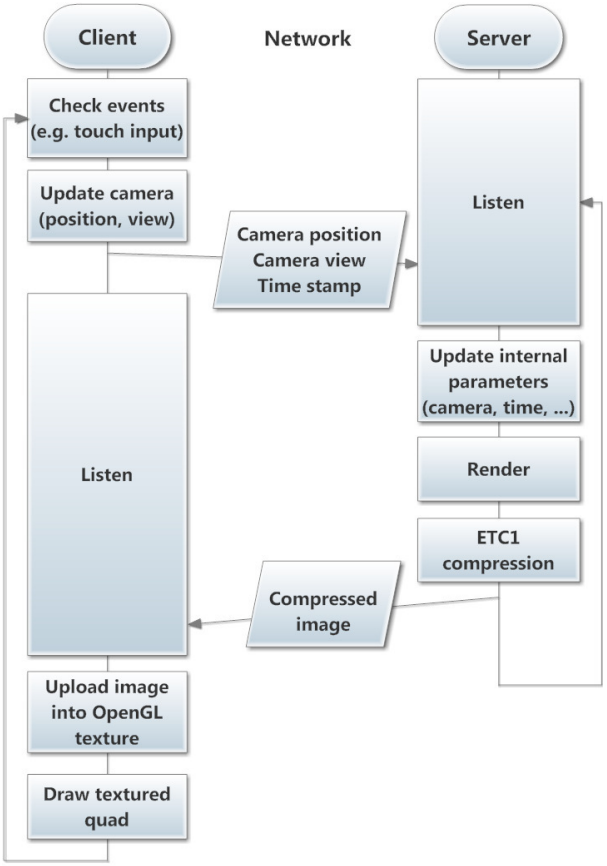


Fig. 2. *Tasks of the client and server architecture.*

The client runs Android 4.4.2 and executes an app that we wrote using libSDL 2.0 [25], libSDL_net 2.0 and OpenGL ES 2.0. All relevant logic has been implemented using the Android Native SDK (NDK r9b).

In the initialization phase the client informs the servers about the rendering resolution and which parts the render server should handle, parameters for loading content, and initial camera settings for rendering. Then the client will receive an image (or part of it) from the server. After the initialization steps the following procedure as described in Figure 2 will be processed every frame. The client checks for user input (like touch) and interprets this into changes in the camera setup (this step can also be done on the server side to stay application-independent). Those new settings, an unique time stamp and other application relevant data (total of 192 bytes) are then sent to the server. The server receives this

over the network and updates its internal states. An image (or part of it) is rendered, then compressed to ETC1 and sent to the client. There, the compressed image data is uploaded as an OpenGL ES texture. Next, a quad is drawn on the screen using that texture at the representing areas that have been assigned earlier to the rendering server.

Our used rendering algorithm, ray tracing, is known as an "embarrassingly parallel" problem [26] with very high scalability across the number of cores, CPUs and servers, because rendering the image can be split into smaller, independent tasks without extra effort. Therefore for the multi-server setup we naïvely split the image into four parts (one for each server), dividing the horizontal resolution by 4 and keeping the full vertical resolution. In order to achieve good scaling for a high number of servers we recommend instead using smaller tiles and to smartly schedule them over e.g. task stealing [27]. In addition to Figure 2 the client will now send one packet (192 bytes) to each server regarding the updates. The client will await the first part of the image from the first server, then upload it as OpenGL ES texture, then receive the second part of the image and so on.

As exemplary rendering content the "island" map from the game *Enemy Territory: Quake Wars*[1] is used.

It is our goal to make a very smooth experience by fully utilizing the display refresh rate of the smartphone (60 Hz). Given the properties of our hardware setup we choose to render at 1280×720 pixels instead of the full physical display resolution of the Nexus 5 (1920×1080 pixels) as it was in the current set-up not possible to transit higher resolutions with 60 Hz due to limitations of the data rate of a 1-antenna 802.11ac setup in current smartphones.

IV. Evaluation

Here, we evaluate our and other in-home streaming approaches in terms of data rate, latency, image quality and power consumption. We compare our implementation with Nvidia Shield[2] and Splashtop[3]. Steam's in-home streaming is currently still in beta and we encountered a high amount of frame drops during our tests, so we will not further compare it here.

A. Data Rate

We discuss the available data rate and the implications of it. The wireless networking standard 802.11ac allows a maximum data rate of 433 Mbit/s (for a 1-antenna setup). Hardware tests show an effective throughput of 310 Mbit/s (38.75 MB/s) for our router [28]. As our rendering resolution is 1280×720 pixels and has 8 bit per color channel this makes about 2.64 MB per image for uncompressed RGB data. Using ETC1 with the fixed compression ratio of 1:6 leads to 0.44 MB per image. Assuming the effective throughput this results in a maximum of about 88 frames per second (fps).

[1] id Software and Splash Damage

[2] Android 4.3, System Update 68

[3] Splashtop Personal 2.4.5.8 and Splashtop Streamer 2.5.0.1 on the Nexus 5

B. Latency

We have a detailed look at the latency of our and other streaming approaches.

The total latency from an user input to an update on the screen (motion-to-photons time) can have various causes of lag in an interactive streaming setup. The user input takes time to get recognized by the operating system of the client. Next, the client application needs to react on it. However, especially in a single-threaded application, the program might be busy doing other tasks like receiving image data. Afterwards, it takes time to transfer that input (or its interpretation) to the rendering server over network. There, the server can process the new data and start calculating the new image. Then compression takes place and the data is sent to the client, which might be delayed if the client is still busy drawing the previous frame. Once the image data is received, it will be uploaded to the GPU for displaying. Fixed refresh rates through VSync might add another delay before the frame can be shown. Some displays have input lag, which describes the time difference between sending the signal to the screen and seeing the actual content there. Additional delay happens when the client or server use multiple buffers for graphics. 3D application use double buffering, sometimes even triple buffering, to smooth the average frame rate. In our setup we are keeping the number of buffers as low as possible.

For the following measurements of our implementation we took the setup with four servers. As the distributed rendering approach does not work with the solutions we are comparing to, we modified the setup to use only one server and a very simple scene that achieves the same frame rate on a single machine as our four servers in the more complex rendering scenario. That way we have a fair comparison of the latency across the approaches. To get accurate motion-to-photons time we captured videos of user input and waiting for the update on the screen. Those videos are sampled at 480 frames per second using the Casio Exilim EX-ZR100 camera. In a video editing tool we analyzed the sequence of frames to calculate the total latency. Using our approach led to a motion-to-photons latency of 60 to 80 ms. On Nvidia Shield, which uses H.264 video streaming, we measured 120 to 140 ms. The Splashtop streaming solution, also relying on H.264, shows 330 to 360 ms of lag.

C. Image Quality

In this section we compare the image quality of our method with Splashtop, Nvidia Shield and with creating our own H.264 stream with different bit rate settings.

To analyze the difference in image quality we chose one image of a sequence in which a lot of camera movement is happening as shown in Figure 3. We quantify the image quality using the metrics of Peak Signal-to-Noise Ratio (PSNR) [29]

Fig. 3. *Left: Previous frame. Right: Frame for analysis with marked red area.*

and Structural Similarity (SSIM) [30] index, which takes human visual perception into account. For Nvidia Shield and Splashtop we were not able to test the distributed rendering setup, so we precalculated the ray traced frames offline and played them back from a single machine at the same speed that they would have been generated using four servers. That way a fair comparison of the image quality happens across all approaches. In Table I one can see that our approach has better image quality compared to Nvidia Shield, Splashtop and H.264 encoding at 5 Mbit/s. As expected, higher bit rate inter-frame encoding offers even higher image quality: going to 50 MBit/s using H.264 succeeds the quality delivered by ETC1. Using an even higher bit rate than 50 MBit/s during H.264 encoding does practically result in the same image quality for our setup.

D. Performance

Here, we report the rendering performance of our approach, the effective throughput and which tasks consume how much time.

We are able to achieve 60 frames per second at 1280×720 pixels. Given the ETC1 compression ratio of 1:6, this corresponds to an effective throughput of 210.93 Mbit/s (26.36 MB/s). The relevant components on the server side are rendering of the image region ($\sim$7 ms), network transfer ($\sim$3 ms) and ETC1 compression ($\sim$2 ms). On the client side network transfer ($\sim$12 ms) and OpenGL ES commands including swapping the display buffer and waiting on VSync ($\sim$4 ms) are the most time consuming tasks.

E. Battery Drain

Here we compare the battery drain of our approach, Splashtop and a locally rendered 3D application on the same smartphone and Nvidia Shield.

For the Nexus 5 we use the "CurrentWidget" app [31]. In our approach we observed an average battery usage of 850 mA, 605 mA for Splashtop and 874 mA for the locally rendered 3D game "Dead Trigger 2" [32] . The higher battery usage of our approach compared to Splashtop can be explained by the fact that we are handling much more data. The Nvidia Shield console uses different hardware; therefore the architectural difference has impact on the result and cannot be compared directly. Nevertheless, we report the number for completeness. The "CurrentWidget" app does not work on Nvidia Shield, so we measured the drop in percentage of the available battery power for an hour and by knowing the total battery capacity we got a value of 880 mA.

V. Enhanced Applications

In this section we have a look on various application scenarios that are enhanced by our high-quality, low latency in-home streaming solution. These can vary widely as the content of the displayed image is independent of the internals of the used compression, transportation and displaying method.

High-quality Gaming As there are already products evolving for in-home streaming for games, like Nvidia Shield and Steam's in-home streaming, this could potentially be an area of growth. The benefits are e.g. to play on the couch instead of sitting in front of a monitor or to play in another room where an older device is located that would not be able to render the game in the desired high quality. In fast-paced action games like first person shooters it is important to be able to react as fast as possible, therefore our reduced latency setup enriches the gaming experience. The commercially available solutions for in-home streaming of games are typically limited to using the rendering power of only one machine. Through the distributed rendering approach we potentially enable closer to photo-realism games by combining the rendering power of multiple machines.

Virtual Reality for Smartphones For virtual reality there are projects like FOV2GO [33] and Durovis Dive [34] (see Figure 4) that developed cases for smartphones with wide-angle lenses attached to it. Once this is strapped on the head of a user, mobile virtual reality can be experienced. For a good Quality of Experience high-quality stereo images need to be rendered that have pre-warped optical distortion compensation to cancel out spatial and chromatic distortions of the lenses. While this works good on desktop PCs [35], the performance

	Original	H.264 211 Mbit/s	H.264 50 Mbit/s	ETC1	H.264 5 Mbit/s	Splashtop	Nvidia Shield
PSNR	-	47.0	46.9	37.3	32.9	31.0	29.8
SSIM	1.0	0.997	0.997	0.978	0.877	0.861	0.779

TABLE I

PSNR and SSIM values for different codecs and platforms, exemplified by respective image sections as marked in Figure 3. Higher values are better. While with a high bit rate the image quality of H.264 surpasses ETC1, the later approach has significantly lower latency, which we show in section IV-B.

and quality that smartphones can achieve today is not very compelling for virtual reality. To achieve higher image quality, these applications have to switch from a local to a server-based rendering approach. As latency is an even more important issue in virtual reality, our latency-optimized approach is in particular suitable for this scenario. Solutions with a latency of 120 to 140 ms (Nvidia Shield) would lead to much more motion sickness compared to a latency of 60 to 80 ms. Nevertheless, optimizing virtual reality streaming applications for even lower latency might become more relevant in the future.

Fig. 4. *A mobile virtual reality platform that can be strapped on the head. In front of the case a smartphone is plugged in. Lenses bring the image into focus for the viewer. To compensate for optical distortions a high-quality, pre-warped stereoscopic image is used and streamed with low latency using our framework.*

Wearables: One of the emerging trends in the space of wearables are smartwatches. E.g. the Neptune Pine [36] is a fully independent Android device, equipped with its own CPU and GPU, 802.11n Wi-Fi and a 320x240 resolution display. Size, battery life and cost are limiting factors, so these devices are usually equipped with less capable processing units compared to other mobile devices. There is a chance that ETC1 streaming could unlock the full power of smartwatches - independent of their weak internal components. Further one could consider having the more powerful smartphone or tablet acting as a rendering server to feed the low-resolution smartwatch.

HPC and Big Data: A scenario where our streaming solution is also well suited for is real-time visualization of data-intensive computations like in the big data and HPC domain. Here, specialized applications either run analyses on huge amounts of data or computationally intensive calculations, typically relying on powerful back ends with a high amount of system memory. Typical application domains are health sciences, simulations in engineering, geographic information systems or marketing and business research. Being able to control, monitor and visualize these computations running on big server farms from small handheld devices is a very convenient benefit. Our solution, in comparison to other in-home streaming approaches, enhances graphics streaming for HPC applications as it supports a distributed scheme for rendering natively at high-quality and low latency. A testbed where we are currently integrating our streaming solution into is the molecular modeling and visualization toolkit BALL/BALLView [37], [38], see Figure 5. In BALLView,

running computationally demanding molecular dynamics simulations in combination with real-time ray tracing visualization on complex molecular data sets [39] requires a powerful compute server. Operating these experiments from a handheld device like a tablet is considered highly preferable.

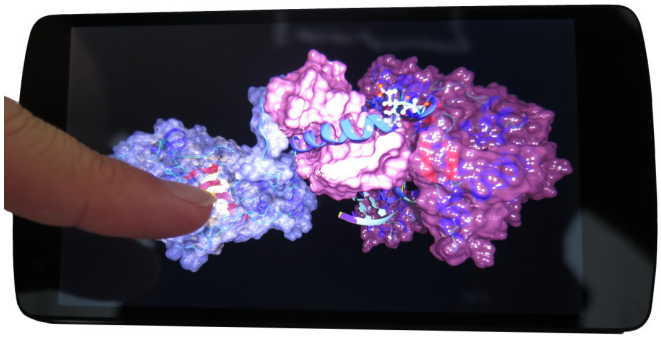


Fig. 5. *Molecular visualization from BALLView displayed on a mobile device.*

VI. Conclusion and Outlook

We conclude what we have demonstrated in this paper and give an outlook on future work.

We have shown a new approach to in-home streaming that fully leverages the latest development in wireless network standards and utilizes a hardware-accelerated intra-frame decompression scheme supported by modern mobile devices. The approach is well suited for streaming of interactive real-time applications and offers significantly higher image quality and half the latency in comparison to other recent solutions targeting this space.

Further optimizations like hardware encoders for ETC, switching to ETC2 [40] compression and using a 2×2 antenna network connection setup, as supported by IEEE 802.11ac, will lead to even higher image quality, faster performance and will enable 1080p at 60 fps.

Acknowledgements

Thanks to Timo Bolkart (MMCI, Saarland University) for his continued support during the writing of this paper. Thanks to Sven Woop for his support with Intel Embree. We thank Bradley Jackson for his feedback. Furthermore, the authors acknowledge financial support through the Intel Visual Computing Institute (IVCI) of Saarland University.

References

[1] "IEEE Standard 802.11ac-2013 (Amendment to IEEE Std 802.11-2012)," 12 2013.

[2] A. Chalmers and E. Reinhard, "Parallel and distributed photo-realistic rendering," in *Philosophy of Mind: Classical and Contemporary Readings. Oxford and*. University Press, 1998, pp. 608–633.

[3] Microsoft Corporation, "Remote Desktop Connection," http://windows.microsoft.com/en-us/windows/connect-using-remote-desktop-connection.

[4] T. Richardson, Q. Stafford-Fraser, K. Wood, and A. Hopper, "Virtual network computing," *IEEE Internet Computing*, vol. 2, no. 1, pp. 33–38, 1998. doi: 10.1109/4236.656066

[5] Gaikai Inc., "Gaikai," http://www.gaikai.com.

[6] OnLive Inc., "Onlive," http://www.onlive.com.

[7] C.-Y. Huang, C.-H. Hsu, Y.-C. Chang, and K.-T. Chen, "GamingAnywhere: An open cloud gaming system," 2013. doi: 10.1145/2483977.2483981 pp. 36–47.

[8] T. Wiegand, G. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003. doi: 10.1109/TCSVT.2003.815165

[9] Sony, "PlayStation 4," http://us.playstation.com.

[10] ——, "PlayStation Now," http://us.playstation.com/playstationnow.

[11] Valve Corporation, "Steam," http://store.steampowered.com.

[12] Nvidia Corporation, "Nvidia Shield," http://shield.nvidia.com.

[13] Splashtop Inc., "Splashtop personal," http://www.splashtop.com.

[14] Wi-Fi Alliance, "Miracast," http://www.wi-fi.org/wi-fi-certified-miracast\%E2\%84\%A2.

[15] Intel Corporation, "Intel wireless display," https://www-ssl.intel.com/content/www/us/en/architecture-and-technology/intel-wireless-display.html.

[16] A. Löffler, L. Pica, H. Hoffmann, and P. Slusallek, "Networked displays for VR applications: Display as a Service (DaaS)," in *Virtual Environments 2012: Proceedings of Joint Virtual Reality Conference of ICAT, EuroVR and EGVE (JVRC)*, 10 2012. doi: 10.2312/EGVE/JVRC12/037-044

[17] I. Nave, H. David, A. Shani, Y. Tzruya, A. Laikari, P. Eisen, and P. Fechteler, "Games@Large graphics streaming architecture," 2008. doi: 10.1109/ISCE.2008.4559473

[18] J. Ström and T. Akenine-Möller, "ipackman: High-quality, low-complexity texture compression for mobile phones," vol. 2005, 2005. doi: 10.1145/1071866.1071877 pp. 63–70.

[19] FFmpeg project, "FFmpeg," http://www.ffmpeg.org.

[20] "YUV420," http://www.fourcc.org/yuv.php\%23IYUV.

[21] D. Vo and T. Nguyen, "Quality enhancement for motion JPEG using temporal redundancies," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 5, pp. 609–619, 2008. doi: 10.1109/TCSVT.2008.918807

[22] Woop, S. and Benthin, C. and Wald, I., "Embree ray tracing kernels for the Intel Xeon and Intel Xeon Phi architectures," http://embree.github.io/data/embree-siggraph-2013-final.pdf.

[23] Madhusood, A., "Best practices for using Intel Cilk Plus," Intel Corporation, White Paper, 7 2013, http://software.intel.com/sites/default/files/article/402486/intel-cilk-plus-white-paper.pdf.

[24] Taudul, B., "etcpak 0.2.1: The fastest ETC compressor on the planet," https://bitbucket.org/wolfpld/etcpak.

[25] "Simple DirectMedia Layer," http://libsdl.org.

[26] G. Fox, R. Williams, and G. Messina, *Parallel Computing works!*, 1st ed. Morgan Kaufmann, 1994.

[27] J. Singh, A. Gupta, and M. Levoy, "Parallel visualization algorithms: performance and architectural implications," *Computer*, vol. 27, no. 7, pp. 45–55, 1994. doi: 10.1109/2.299410

[28] E. Ahlers, "Rasante Datenjongleure," *c't Magazin*, vol. 1, pp. 80–89, 2014.

[29] Y. Wang, J. Ostermann, and Y. Zhang, *Video Processing and Communications*. Prentice Hall, 2002, p. 29.

[30] Z. Wang, L. Lu, and A. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 121–132, 2004. doi: 10.1016/S0923-5965(03)00076-6

[31] RmDroider, "CurrentWidget: Battery Monitor," http://code.google.com/p/currentwidget/.

[32] MADFINGER Games, "Dead Trigger 2," https://play.google.com/store/apps/details?id=com.madfingergames.deadtrigger2.

[33] "FOV2GO," http://projects.ict.usc.edu/mxr/diy/fov2go/.

[34] Shoogee GmbH & Co KG, "Durovis Dive," http://www.durovis.com/index.html.

[35] D. Pohl, G. Johnson, and T. Bolkart, "Improved Pre-Warping for Wide Angle, Head Mounted Displays," 2013. doi: 10.1145/2503713.2503752 pp. 259–262.

[36] Neptune Computer Inc., "Neptune Pine," http://www.neptunepine.com.

[37] A. Hildebrandt, A. Dehof, A. Rurainski, A. Bertsch, M. Schumann, N. Toussaint, A. Moll, D. Stockel, S. Nickels, S. Mueller, and O. Lenhof, H.-P.and Kohlbacher, "BALL - Biochemical Algorithms Library 1.3," *BMC Bioinformatics*, vol. 11, no. 1, p. 531, 2010. doi: 10.1186/1471-2105-11-531

[38] A. Moll, A. Hildebrandt, H.-P. Lenhof, and K. O., "BALLView: a tool for research and education in molecular modeling." *Bioinformatics*, vol. 22, no. 3, pp. 365–366, 2006. doi: 10.1093/bioinformatics/bti818

[39] L. Marsalek, A. Dehof, I. Georgiev, H.-P. Lenhof, P. Slusallek, and A. Hildebrandt, "Real-time ray tracing of complex molecular scenes," in *Information Visualisation (IV), 2010 14th International Conference*. IEEE, 2010. doi: 10.1109/IV.2010.43 pp. 239–245.

[40] Ericsson Labs, "Ericsson texture compression tool etcpack v2.60: ETC2," https://labs.ericsson.com/research-topics/media-coding.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 695–699
DOI: 10.15439/2014F258
ACSIS, Vol. 2

An Optimized Version of the K-Means Clustering Algorithm

Cosmin Marian Poteraș
University of Craiova
Faculty of Automation, Computers and Electronics
Blvd. Decebal nr. 107,
Craiova, Romania
Email: cpoteras@software.ucv.ro

Marian Cristian Mihăescu
University of Craiova
Faculty of Automation, Computers and Electronics
Blvd. Decebal nr. 107,
Craiova, Romania
Email: mihaescu@software.ucv.ro

Mihai Mocanu
University of Craiova
Faculty of Automation, Computers and Electronics
Blvd. Decebal nr. 107,
Craiova, Romania
Email: mmocanu@software.ucv.ro

***Abstract*—This paper introduces an optimized version of the standard K-Means algorithm. The optimization refers to the running time and it comes from the observation that after a certain number of iterations, only a small part of the data elements change their cluster, so there is no need to re-distribute all data elements. Therefore the implementation proposed in this paper puts an edge between those data elements which won't change their cluster during the next iteration and those who might change it, reducing significantly the workload in case of very big data sets. The prototype implementation showed up to 70% reduction of the running time.**

I. Introduction

THE MORE data volumes continue to grow in complexity and diversity, the harder it is to structure and manipulate them. Finding data with similar characteristics and labeling it accordingly has become one of the greatest challenges in nowadays data analyzing applications. Grouping data based on common characteristics is what we call clustering. Clustering algorithms fall into the unsupervised classification techniques category. They classify a set of objects into a subset of clusters based on similarities between them. Differences between clusters have to be obvious and clearly expressed.

Clustering can be applied to a wide range of domains like: marketing [1] (market analysis and recommendations, methodological weaknesses), medicine [2] (medical image segmentation), e-business [3] (comments analysis on news portal) or e-learning [4][5] (prediction of students' academic performance).

There is no secret recipe for choosing the best clustering algorithm. The choice should be based on experimental studies and data description possibly mixed with some human intuition unless there is no obvious mathematical model.

Some of the problems raised by clustering algorithms which worth investing research efforts are: scalability, handling heterogeneous data, execution time complexity when dealing with very large data sets, multi-dimensional data, etc.

The standard K-Means algorithm represents one of the most popular unsupervised exclusive clustering algorithms. It has been successfully applied to medical image segmentation as shown in [6] where the authors propose an algorithm for the segmentation of three-dimensional (3-D) image data based on a combination of adaptive K-Means clustering and knowledgebased morphological operations.

K-Means is based on the minimization of the average squared Euclidean distance between the data items and the cluster's center (called centroid). The results of the algorithm are influenced by the initial centroids. Different initial configurations might lead to different final clusters. The cluster's center is defined as the mean of the items in a cluster.

This paper focuses on the execution of the K-Means algorithm, namely it tries to improve the running time when dealing with high volumes of data. The standard implementation of K-Means consists of successive iterations. Each iteration requires visiting the entire data set in order to assign data objects to their corresponding cluster. At the end of each iteration, new centroids are being computed so that the next iteration will employ the new centroids. After a certain number of such iterations, the centroids will keep the same and the algorithm stops.

The optimization proposed by this paper relies on the observation that after performing a number of iterations, just a small part of the data set might change the cluster it belongs to. Our implementation traces a border between that part of the data set which could possibly switch to another cluster and the data that will hold the cluster it belongs to, during the next iteration. As K-Means algorithm's execution advances, the centroids come closer to their final position. The more iterations are performed, the less the centroids deviate from their current position, resulting in less data objects to be checked against. Similar to the classical implementation, the final clusters are sensitive to the initial configuration (initial centroids).

The rest of the paper is structured as follows: section II presents previous results in speeding up the K-Means algorithm, section III describes the proposed optimization for the K-Means algorithm, section IV experimentally evaluates the potential of the proposed optimization, while section V concludes the paper and presents our future research intentions.

978-83-60810-58-3/$25.00

Algorithm 1 Standard K-Means algorithm

1. Choose k data objects representing the cluster centroids;
2. Assign each data object of the entire data set to the cluster having the closest centroid
3. Compute new centroid for each cluster, by averaging the data objects belonging to the cluster
4. If at least one of the centroids has changed, go to step 2, otherwise go to step 5
5. Output the clusters.

II. Related Work

Researches have shown special interest for speeding up the K-Means algorithm, by either reducing the computation complexity or by adopting K-Means implementations for parallel and distributed platforms.

In [7] the authors propose an efficient implementation of the Lloyd's (K-Means) algorithm called the filtering algorithm which employs kd-trees for storing the data elements.

In [8], the authors propose an algorithm which reduces the computations for determining the closest centroid of a data element, by making use of the observation that if a data element gets closer to the centroid defined at the previous iteration, it won't switch the cluster it belongs to.

Other strategies [9][10][11][12][13] focused on parallelizing the K-Means algorithm and take advantage of powerful parallel and distributed environments, addressing issues specific to those environments, like data availability, synchronization, etc. and adapting the K-Means algorithm to different distributed architectures (client-server, peer-to-peer, etc). The results were satisfactory for very big data sets.

The highly-parallel GPUs haven't been ignored either. Papers [14][15] propose parallel implementations of K-Means to be run on GPUs.

III. Optimized K-Means Method

Before proceeding with our optimized K-Means, let us examine first the standard K-Means algorithm. It consists of repetitive steps, as presented in algorithm 1.

Let us have a look at the algorithm and try to identify what step causes the most computations. Obviously in case of very large data sets, step number 2 would require the biggest time frame in the algorithm's execution. The bigger the data set, the wider the time frame of step 2's execution as it visits each data object and performs some computations on it. The question that arises here is: do we need to visit the entire data space? Figure 1 illustrates the centroids' evolution in a standard K-Means execution.

The data objects are represented by 2D points. The algorithm starts with centroids A1, B1 and C1, which change their position with each iteration, successively to A_i, B_i and C_i, where i=1..6 until they no longer change. If we take a closer look, we can easily see that as the execution progresses, the centroids get very close to their final position. Actually, it happens very often that after only few iterations, the centroids undergo their trip to the very close neighborhood of their final position. This observation leads us to the conclusion that most of the data objects belonging to a cluster whose centroid slightly moves, should not be affected by the move; they will remain part of the same cluster during the next iteration. The less the centroid moves, the less points get affected by the move.

Being able to determine which of the data objects could be affected by a move, could lead us to a very important improvement on step number 2 as we no longer need to visit the entire data set, but just a small list of data objects (let us call that the 'border' list). Before deciding which data objects should be placed into the 'border' list we need to establish the criteria that need to be fulfilled by a data element so that it can be considered a 'border' element. Let us consider Figure 2 which assumes the iteration i is to be computed.

Let point P be part of cluster C. All other points have been omitted on purpose for the ease of presentation. Point P is part of cluster C as the distance from P to C (d_{PC})is less than the distance to A (d_{PA}) and less than the distance to B (d_{PB}). We want to know, how far away is point P from jumping to another cluster. That would obviously be:

$$e_P = min(d_{PA}\text{-}d_{PC},\ d_{PB} - d_{PC}) \quad (1)$$

We've labeled as e_P the distance from P to the closest edge. We can say that point P is e_P-away from switching the cluster.

At the end of iteration i, centroids need to be updated based on the new clusters' configuration. Let us assume that centroid A moved to A', centroid B moved to B' and centroid C moved to C'.

In the context shown in figure 2, the worst case scenario for point$P$ would be: point C got farther away from P by $|CC'|$ while point A got closer by $|AA'|$ and point B got closer by $|BB'|$. What would be the condition for P to stay in cluster C? Obviously that would be

$$e_P > |CC'| + |AA'| \quad (2)$$

and

$$e_P > |CC'| + |BB'| \quad (3).$$

To simplify a little the algorithm and reduce the computations, we can blend conditions (2) and (3):

$$e_P > 2 * max(|AA'|, |BB'|, |CC'|). \quad (4)$$

That being said, we've just found a way of determining whether a point is part of the 'border' list or not.

But, we're still not ok because checking the inequality for each of our data elements at each iteration gets us back to where we started. To avoid such computations, we can map all of our data elements into wider intervals for the value of e, as shown in algorithm 2:

The algorithm 2 groups points with close values of e so that instead of visiting each data element and checking against their close-to-the-edge distance, we can do that that for the entire group. This compromise is the key of the entire optimization. The *WIDTH* constant has a big influence on the optimization. If the value of *WIDTH* is to small, then the

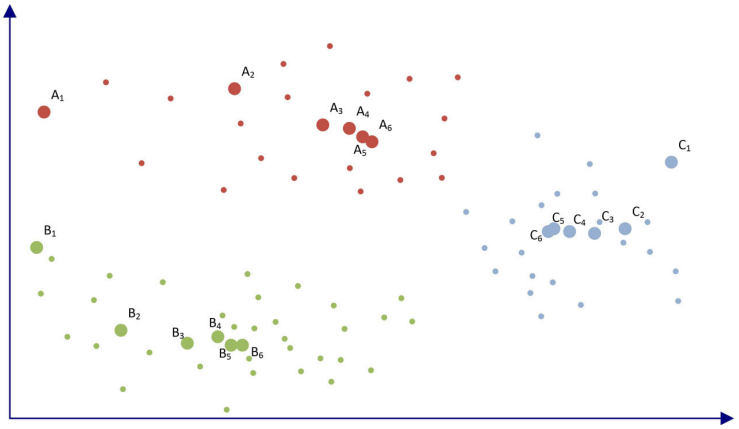


Fig. 1. Example of centroids evolution

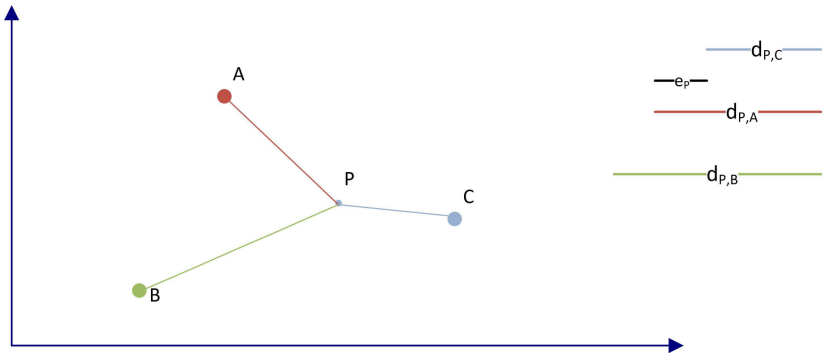


Fig. 2. Edge condition for data elements

Algorithm 2 Optimized K-Means algorithm

1. Define constant *WIDTH*
2. Define intervals I_i = *[i * WIDTH, (i+1) * WIDTH)* and tag them with value *i * WIDTH*
3. Mark the entire data set to be visited
4. For each point to be visited
5. Compute $e = min(d_{PC_l} - d_{PC_w})$ where C_w is the center of the winner (closest) cluster and C_l, $l=1..k$, $l \neq w$ stands for all other centroids
6. Map all points with $i * WIDTH < e < (i+1) * WIDTH$ to interval $i * WIDTH$ where i is a positive integer
7. Compute new centroids C_j, where $j=1..k$ and their maximum deviation $D = max(|C_jC_j'|)$
8. Update I_i's tag by subtracting $2 * D$ (points owned by this interval got closer to the edge by $2 * D$)
9. Pick up all points inside intervals whose tag is less or equal to 0, and go to *4* to revisit them

number of intervals will increase and the workload involved by checking and updating the intervals at each iteration can increase significantly. If the value of *WIDTH* is to big, then the number of points for each interval increases, and even though the number of intervals is reduced, the performance loss is obvious when a big interval is marked for re-visiting.

One can easily observe that the quality of the final clusters is not affected by the proposed optimization. At each iteration, the composition of clusters is exactly the same as if we would have run the standard K-Means.

IV. Prototype Evaluation

To validate the potential of our K-Means optimization, we've implemented a prototype that runs on data sets composed of 2D points having their coordinates greater than 0 and less than 1. There have been randomly generated data sets of different sizes ranging from 100,000 points to 5,000,000 points. The random numbers generator used for generating the points coordinates, made use of an uniform distribution. The data sets were split into 4, 8 or 12 clusters. The points belonging to each cluster have been grouped into 0.1 wide intervals (*WIDTH = 0.1*). The running time of the optimized algorithm was compared to the running time of the standard K-Means processing of the same data sets in exactly the same conditions (same centroids, same execution environment). Experiments were conducted on a machine consisting of an Intel i7-4700MQ CPU, 8 GB RAM memory. Many runs were carried out for each use case, and the running times were averaged. Let us have a look at the results.

Table I presents the running time for both optimized and standard versions of the K-Means algorithm where the data

TABLE I
OPTIMIZED K-MEANS VS. STANDARD K-MEANS RUNNING TIMES - 4 CLUSTERS

Data set size	Running Time Standard K-Means (ms)	Running Time Optimized K-Means (ms)	Improved by(%)
100,000	6300	1967	68.77
250,000	14382	4528	68.51
500,000	40321	11402	71.72
750,000	55088	15943	71.05
1,000,000	73957	21436	71.01
2,000,000	140339	42962	69.38
5,000,000	420516	116630	72.26

TABLE II
OPTIMIZED K-MEANS VS. STANDARD K-MEANS RUNNING TIMES - 8 CLUSTERS

Data set size	Running Time Standard K-Means (ms)	Running Time Optimized K-Means (ms)	Improved by(%)
100,000	75664	28418	62.44
250,000	148472	57485	61.28
500,000	629777	224277	64.38
750,000	1004100	359230	64.22
1,000,000	1096291	396923	63.79
2,000,000	2798319	1006918	64.01
5,000,000	9685205	3355996	65.34

sets were divided into 4 clusters.

The running time has been reduced by up to 72.26%

Table II, presents the running time for both optimized and standard versions of the K-Means algorithm where the data sets were divided into 8 clusters. The running time has been reduced by up to 64.48%, which is less than the improvement shown in case of only 4 clusters. That can be explained by the fact that the more clusters we use, the higher the chances are for a bigger maximum centroid deviation ($\max(|C_iC_i)$, where $i = 1..K$), which decrease the chances of fulfilling inequality (4) causing more points to become part of the 'border' area.

Table III, presents the running time for both optimized and standard versions of the K-Means algorithm where the data sets were divided into 12 clusters. The best improvement we have got here rises up to 53.42% which again, is less than the improvement we have got for 4 and 8 clusters. These results confirm that a higher number of clusters results in wider 'border' areas, reducing the computational gain.

V. CONCLUSIONS AND FUTURE WORK

The paper introduces an optimized version of the K-Means algorithm. The optimization refers to the running time. Optimization comes from the considerable reduction of the data space that is re-visited at each loop.

The algorithm defines a 'border' area made of those points that are close enough to the edge of their cluster so that the next centroids move could cause them to switch clusters.

A prototype implementation of a domain specific data set has been evaluated. The implementation assumes the data set is made of 2D points with their coordinates between 0 and 1. The data set has been generated using a uniform distribution generator.

TABLE III
OPTIMIZED K-MEANS VS. STANDARD K-MEANS RUNNING TIMES - 12 CLUSTERS

Data set size	Running Time Standard K-Means (ms)	Running Time Optimized K-Means (ms)	Improved by(%)
100,000	53879	27388	49.16
250,000	186923	90140	51.77
500,000	323584	158888	50.89
750,000	681331	317328	53.42
1,000,000	809675	377522	53.37
2,000,000	1650657	776873	52.93
5,000,000	4835146	2324173	51.93

Running times for 4, 8 and 12 centroids have been compared to the running times of the standard K-Means algorithm, showing a reduction ranging from 49.16% to 72.26 %. At this stage we can not confirm that the improvement shown by the prototype will be held in all real-world use cases, but the results are certainly encouraging.

Our future research will focus on the domain-independent implementation and evaluation of the algorithm. The algorithm's scalability as well as data sensitivity (form and distribution) are to be analyzed with the purpose of concluding upon what would be the best and the worst environments (data and configuration) for the algorithm.

A natural question would be if the algorithm can be improved. One can easily note that the grouping intervals' width might be a point of vulnerability for the performance gain. The lower the width, the more intervals are to be checked; the higher the width, the more points are to be checked when their interval's distance to the edge goes bellow 0. A tradeoff has to be made here, therefore we will also focus on designing an auto-calibration algorithm for the interval width.

Implementations for parallel and distributed environments, as well as integration with existing frameworks (Hadoop, Mahout) are also on our goals list as they could lead our way towards big data sets.

REFERENCES

[1] Dolnicar, S, Using cluster analysis for market segmentation - typical misconceptions, established methodological weaknesses and some recommendations for improvement, Australasian Journal of Market Research, 2003, 11(2), 5-12.

[2] Ng, H.P., Ong, S.H.; Foong, K.W.C.; Goh, P.S.; Nowinsky, W.L. - Medical Image Segmentation Using K-Means Clustering and Improved Watershed Algorithm, 7th IEEE Southwest Symposium on Image Analysis and Interpretation, March 26-28, 2006, Denver, Colorado, pages 61-66

[3] Hongwei Xie, Li Zhang ; Jingyu Sun ; Xueli Yu - Application of K-means Clustering Algorithms in News Comments - The International Conference on E-Business and E-Government, May 2010, Guangzhou, China, pages 451-454

[4] kK Oyelade, O. J, Oladipupo, O. O, Obagbuwa, I. C - Application of K-Means Clustering algorithm for prediction of Students' Academic Performance, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 1, 2010, pages 292–295

[5] Burdescu, D.D.; Mihaescu, M.C., "Enhancing the Assessment Environment within a Learning Management Systems," EUROCON, 2007. The International Conference on "Computer as a Tool", vol., no., pp.2438,2443, 9-12 Sept. 2007

[6] Chang Wen Chen, Jiebo Luo, Kevin J. Parker - Image Segmentation via Adaptive K-Mean Clustering and Knowledge-Based Morphological Operations with Biomedical Applications, IEEE Transactions on Image Processing, VOL. 7, NO. 12, DECEMBER 1998, pages 1673 - 1683

[7] T. Kanungo, D.M. Mount, K.D. Piatko, N.S. Netanyahu, R. Silverman, A. Y. Wu - An efficient k-means clustering algorithm: analysis and implementation, Pattern Analysis and Machine Intelligence, IEEE Transactions on (Volume:24, Issue: 7), pages 881-892, July 2002, ISSN 0162-8828

[8] Fahim A.M., Salem A.M., Torkey F.A., Ramadan M.A. - An Efficient Enhanced K-means Clustering Algorithm Journal of Zhejiang University SCIENCE A, ISSN 1009-3095 (Print); ISSN 1862-1775 (Online), pages 1626 - 1633, 2006 7(10)

[9] Souptik Datta, Chris Giannella, Hillol Kargupta - K-Means Clustering Over a Large, Dynamic Network, Proceedings of the Sixth SIAM International Conference on Data Mining, April 20-22, 2006, Bethesda, MD, USA. SIAM 2006 ISBN 978-0-89871-611-5, pages 153 - 164

[10] Yufang Zhang, Zhongyang Xiong, Jiali Mao, Ling Ou - The Study of Parallel K-Means Algorithm, Proceedings of the 6th World Congress on Intelligent Control and Automation, June 21 - 23, 2006, Dalian, China, pages 5868 - 5871

[11] Jing Zhang, Gongqing Wu, Xuegang Hu, Shiying Li, Shuilong Hao - A Parallel K-means Clustering Algorithm with MPI, 4th Internation Symposium on Parallel Architectures, Algorithms and Programming, ISBN 978-0-7695-4575-2, pages 60-64, 2011

[12] Fazilah Othman, Rosni Abdullah, Nur'Aini Abdul Rashid, and Rosalina Abdul Salam - Parallel K-Means Clustering Algorithm on DNA Dataset, Parallel and Distributed Computing: Applications and Technologies, Lecture Notes in Computer Science Volume 3320, 2005, pp 248-251

[13] Jitendra Kumar, Richard T. Mills, Forrest M. Hoffman, William W. Hargrove - Parallel k-Means Clustering for Quantitative Ecoregion Delineation Using Large Data Sets, Proceedings of the International Conference on Computational Science, ICCS 2011, Procedia Computer Science 4 (2011) 1602-1611

[14] Reza Farivar, Daniel Rebolledo, Ellick Chan, Roy Campbell - A Parallel Implementation of K-Means Clustering on GPUs, Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications, PDPTA 2008, Las Vegas, Nevada, USA, July 14-17, 2008, 2 Volumes. CSREA Press 2008 ISBN 1-60132-084-1, pages 340-345

[15] Mario Zechner, Michael Granitzer - Accelerating K-Means on the Graphics Processor via CUDA, The First International Conference on Intensive Applications and Services INTENSIVE 2009, 20-25 April, Valencia, Spain, pages 7-15, ISBN 978-1-4244-3683-5

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 701–708

DOI: 10.15439/2014F59
ACSIS, Vol. 2

Handwritten Signature Verification with 2D Color Barcodes

Marco Querini, Marco Gattelli, Valerio M. Gentile, and Giuseppe F. Italiano
University of Rome "Tor Vergata",
viale del Politecnico 1, 00133, Rome, Italy
marco.querini@uniroma2.it
marco.gattelli@gmail.com
valeriomaria.gentile@gmail.com
italiano@disp.uniroma2.it

***Abstract*—Handwritten Signature Verification (HSV) systems have been introduced to automatically verify the authenticity of a user signature. In offline systems, the handwritten signature (represented as an image) is taken from a scanned document, while in online systems, pen tablets are used to register signature dynamics (e.g., its position, pressure and velocity). In online HSV systems, signatures (including the signature dynamics) may be embedded into digital documents. Unfortunately, during their lifetime documents may be repeatedly printed and scanned (or faxed), and digital to paper conversions may result in loosing the signature dynamics. The main contribution of this work is a new HSV system for document signing and authentication. First, we illustrate how to verify handwritten signatures so that signature dynamics can be processed during verification of every type of document (both paper and digital documents). Secondly, we show how to embed features extracted from handwritten signatures within the documents themselves (by means of 2D barcodes), so that no remote signature database is needed. Thirdly, we propose a method for the verification of signature dynamics which is compatible to a wide range of mobile devices (in terms of computational overhead and verification accuracy) so that no special hardware is needed. We address the trade-off between discrimination capabilities of the system and the storage size of the signature model. Towards this end, we report the results of an experimental evaluation of our system on different handwritten signature datasets.**

I. Introduction

BIOMETRIC recognition refers to the automatic identification of a person based on his/her anatomical (e.g., fingerprint, iris) or behavioral (e.g., signature) characteristics or traits. This method of authentication offers several advantages over traditional methods involving authentication tokens (including ID cards) or passwords: it ensures that the person is physically present at the point-of-identification; it makes unnecessary to remember a password or to carry a token. The most popular biometric traits used for authentication are face, voice, fingerprint, iris and handwritten signature.

In this paper, we focus on handwritten signature verification (HSV). Since people are used to signing documents in their everyday life, HSV is a natural and trusted method for user identity verification. HSV can be classified into two main categories, depending on the hardware used and on the method used to acquire data related to the signature: online and offline signature verification. Offline systems take handwritten signatures (represented as an image) from scanned documents. This means that offline HSV systems only process the 2D spatial representation (i.e., the shape) of the signature. On the contrary, online systems use specific hardware (e.g., pen tablets) to register pen movements during the act of signing. As a result, online HSV systems are able to process dynamic features, such as the time series of the pen's position, pressure, velocity, acceleration, azimuth and elevation. Online signature verification has been shown to achieve higher verification rate than offline signature verification [1], [2], [3], but unfortunately it suffers from several limitations.

First, the online approach works only for digital documents and it is currently unavailable for paper documents. In particular, during a document's life cycle, when a document is being printed, scanned or faxed, the signature dynamics are unavoidably lost. To overcome this limitation, there is an emerging need of designing new methods capable of embedding the signature dynamics within paper documents, along with the signature shape (the 2D spatial representation) which is the only feature usually preserved after printing. This will enable one to verify the authenticity of a document, regardless of its current (paper or digital) format, which is particularly important when the same document is repeatedly printed and scanned (or faxed) in a typical workflow.

Secondly, current online approaches raise privacy and security concerns since they store the genuine signatures of each user on a remote database server. Indeed, both commercial [4], [5], [6] and HSV systems proposed in the scientific literature [1], [7], [8] store genuine signatures of the users in a central database: during verification, specific signature data is retrieved from the database and compared to the actual signature. From the security viewpoint, an intruder who gains unauthorized access to the database containing dynamics of users' signatures can use this information to produce accurate forgeries. From the viewpoint of privacy, the recent news about the NSA surveillance program (see e.g., [9]) have definitely reduced our trust in providing sensitive data (such as signature features) to third parties. In order to address these privacy and security concerns, we need to design novel HSV

978-83-60810-58-3/$25.00

systems capable of supporting the online approach without using signature databases.

Thirdly, the online approach is often feasible only if special purpose hardware is available. Indeed, handwritten signatures are usually acquired by means of digitizing tablets connected to a computer, because smartphones and mobile tablets (that have worse sensitivity) may be not able to support the verification algorithms. As a result, the range of possible usages of the verification process is strongly limited by the hardware needed. To overcome this limitation, one needs techniques capable of verifying signatures acquired by smartphones and tablets in mobile scenarios.

To the best of our knowledge, there is no existing solution which is able to address all of those critical points simultaneously. The approach described in [1] addresses only the first point: by performing offline verification using online handwriting registration, the online approach is (partially) applicable for verifying signatures taken from paper documents, but this framework is not supported by mobile devices and requires an online signature database. Offline HSV solutions (such as [2], [10], [11]) address only the second point: they do not use remote signature databases, but unfortunately they are not able to take into account signature dynamics. Online HSV systems (such as [4], [5], [6], [7], [8]) address only the third point: they are supported by mobile devices, but cannot verify signatures taken from paper documents and are inherently based on remote database servers.

The goal of this paper is to address all of the above challenges by considering new mobile scenarios in which HSV can play a significant role. The novelties of our approach lie mainly in the following three aspects.

First, we present a new system to sign and verify documents so that the online approach is applicable for all kind of documents (including paper documents). It performs verification in a way that the signature dynamics can be used also when the signed documents are printed and scanned, thus allowing the online approach to operate in those cases where only the offline approach was available.

Secondly, we show how to embed features extracted from handwritten signatures within the documents themselves, so that no remote signature database is needed. To accomplish the embedding task, we make use of 2D barcodes. The main challenge here is to be able to store the signature dynamics (into documents), within the limited capacity of barcodes: on the one hand, we need to use a signature model whose size is small, while, on the other hand, we need to increase the capacity of state-of-art-barcodes. For this reason, we designed a color barcode denoted as High Capacity Colored 2-Dimensional (HCC2D) code [12], [13], which is well-suited for this framework because of its high data capacity (if compared with state of art barcodes). Specifically, it is capable of encoding about $4KB/inch^2$ (effective data density) with a success rate of 90% (reliability) [14], [15]. We designed a new barcode decoding algorithm based on graph drawing methods [16], which is able to run in few seconds even on mobile devices and to achieve nonetheless high accuracy in the recognition phase. The main idea of our algorithm is to perform color classification using force-directed graph drawing methods: barcode elements which are very close in color will attract each other, while elements that are very far will repulse each other. Figure 1 illustrates samples of HCC2D codes with 4 and 8 colors.

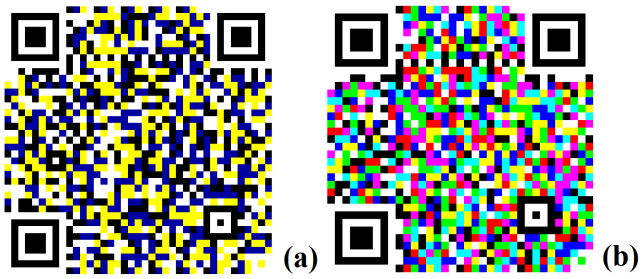


Fig. 1. Samples of the High Capacity Colored 2-Dimensional (HCC2D) code: (a) 4 colors and (b) 8 colors. Figure taken from [12]. (Viewed better in color).

Thirdly, we propose a method for the extraction and verification of signature dynamics which is compatible to a wide range of mobile devices (in terms of computational overhead and verification accuracy) so that no special hardware is needed. The main challenge here is to achieve a high verification performance, despite constrains due to the limited computational resources and pressure accuracy of mobile phones. For this reason, we designed a verification algorithm that can be run on mobile phones in fractions of a second and that weights the signature features based on the accuracy of the given device.

In order to assess the precision and recall of our HSV system, we conduct an experimental study whose results are reported for different data sets of signatures.

II. A Logical View of the HSV System

Our HSV system consists of three main modules, corresponding to three main phases. We next describe (from a logical point of view) the registration phase, the document signing phase and the document verification phase with our system.

A. The Registration Phase

The objective of the registration phase is to compute a compact representation of the signature dynamics of a given person. The process starts with the user writing his/her signature on the device's screen and ends with the generation of a biometric template representative of the user signatures. Figure 2 shows a high level, logical view of the registration procedure.

The registration phase consists of the following tasks. First, in order to take into account the variability among signatures produced by the same user, signature dynamics for at least three signatures are captured. Then, the features extracted from the various signatures are combined to form a template

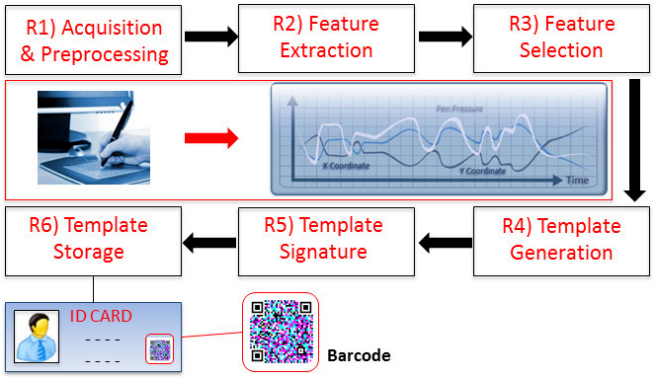


Fig. 2. Registration procedure for the proposed HSV system. (Viewed better in color).

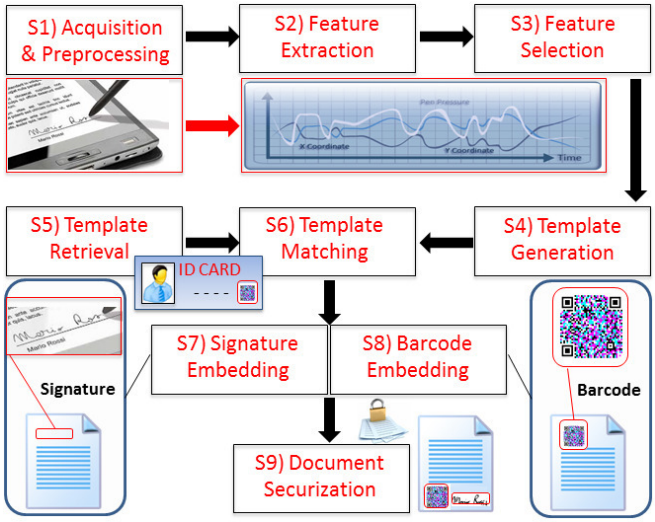


Fig. 3. Document signing procedure for our HSV system. (Viewed better in color).

representative of the given user. Finally, the template is digitally signed and securely stored in a barcode. The barcode is embedded within a card which is released only to the user to whom the template belongs, in order to address any privacy concern. We remark that in order to make this scenario possible, it is necessary that the barcode is capable of storing the whole template and the digital signature associated with it. We will describe in detail how to perform identity registration in Section III.

B. The Document Signing Phase

The objective of this phase is to sign documents on mobile devices so that: the signature dynamics are embedded within the documents themselves, (along with the spatial representation of the signature); the signature dynamics survive the document's life cycle, when a document is being printed, scanned or faxed.

Figure 3 shows a logical view of the document signing phase, which consists of the following steps.

1) Template Generation of the User to Be Verified (S1-S4): We need to generate a compact feature synthesis of the person to verify. In order to accomplish this task, we proceed through steps S1 to S4, which are equivalent to steps R1-R4 of the Registration phase, except for the fact that there is no need to capture three or more writings of the handwritten signature to be verified.

2) Retrieval of the Secure Template (S5): We retrieve the secure template (generated at the end of the registration phase) of the user corresponding to the claimed identity of the user to be verified.

3) Template Matching (S6): We compare the signature dynamics related to the identity to verify with the secure template retrieved at the previous step. This is a crucial step, because we do not allow unrecognized signatures to be embedded within documents. Note that in order to enhance security, we use strong authentication based on something the user has (i.e., a card storing the registered template) with something the user is (features of his/her handwritten signature captured at the moment).

4) Signature Embedding (S7): If the matching is successfully, we embed the spatial representation of the signature within the document.

5) Barcode Embedding (S8): If the matching is successfully, we embed the signature dynamics of the signature within the document by means of a high capacity barcode such as the HCC2D code.

The elements that need to be embedded by means of barcodes include the following:

- The template representing the signature dynamics.
- The timestamp of the signature.
- Information about the document the user is going to sign.
- The digital signature of all the above data.

The last three elements ensure that the barcode storing the signature dynamics cannot be copied and pasted on a new document for producing a forgery.

6) Document Securization (S9): The secure document is generated (with the signature image and the barcode). Because of the binding [signature features, document], no other document can be signed with features that are used for the given document.

C. The Document Verification Phase

The aim of the verification phase is to verify the authenticity of a handwritten signature. This phase ends with the authenticity of the document signature being accepted or rejected. Figure 4 shows a high level view of the document verification procedure, consisting of the following tasks.

1) Barcode Reading (V1): The objective of this step is to decode the HCC2D code which has been encoded in the document signing phase. This allows us to retrieve the secure template which has been stored within the barcode, along with the document metadata, the timestamp and the digital signature.

2) Template Retrieval (V2): The signature features used to sign the given document are retrieved from the barcode.

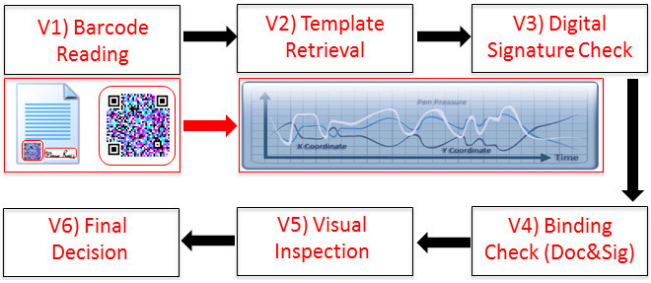


Fig. 4. Document verification procedure for our HSV system. (Viewed better in color).

3) Digital Signature Check (V3): The digital signature is checked before the retrieved data are processed, in order to ensure that they are not tampered with.

4) Binding Check (V4): The binding [signature features, document] is retrieved and showed to the user, in order to verify that those features were previously associated with the given document.

5) Visual Inspection and Final Decision (V5-V6): The signature matching is graphically showed to the user so that he/she is able to sense the similarity and the likelihood of a forgery. This is a crucial step because it ensures that the final decision is taken by the user.

In this stage, the shape of the signature is reconstructed from the signature dynamics stored in the barcode. We cannot completely trust the signature shape that is pasted on the document because it may be a forgery, but we can trust the reconstructed shape because data stored in the barcode are digitally signed.

III. The Signature Verification Algorithm

In this Section, we describe (from a technical point of view) our signature registration and verification methods.

A. Signature Registration Phase

The registration phase with our system starts with the user writing his/her signature on the device screen and ends with the generation of his/her biometric template (representative of his/her signature dynamics), which is embedded within a document (issued to the user) by means of a high capacity barcode. The procedure is as follows:

- First, the user is requested to make genuine signature 3 times.
- Once signatures are acquired, the system performs an analysis of the intra-person variability of the signature features. This allows the system to determine whether to accept or reject the 3 signatures upon which the user model has to be built: if variability is too high, the user is requested to repeat the acquisition step.
- If the intra-person variability is acceptable, for each signature, for each sampled point, the following elements (signature dynamics) are captured:
  - *Event Type*. The event which led to the generation of a sample can be of three different types: Pen-down, i.e., a pressed gesture has started (the motion contains the initial starting location); Pen-up, i.e., a pressed gesture has finished (the motion contains the final release location as well as any intermediate points since the last down or move event); Pen-Move, i.e., a change has happened during a press gesture (between a Pen-down and a Pen-up).
  - *Time*. The instant in which the pressed gesture has occurred (expressed as milliseconds since the first gesture event of the signature).
  - *X-Y coordinates*. The X and Y coordinates of the sampled point (expressed in pixels).
  - *Pressure*. The pressure with which the screen is pressed (expressed with a value ranging from 0, i.e., no pressure, to 1, i.e., maximum pressure).
- The signature dynamics of the three signatures are stored in a high capacity barcode which is embedded within a document.

Note that derivatives such as velocity or higher-order derivatives such as acceleration are not stored in this stage, as they can be computed at run-time during the verification phase. This is because we aim at minimizing the storage size of the signature model to be embedded by means of barcodes.

B. Signature Verification Phase

The verification phase determines the acceptance or the refusal of the claimed identity based on the similarity between the registered signature and the signature to verify. Our system can compute the similarity score in different ways depending on whether signatures are segmented or not (see below) before applying the verification algorithm, where signature segmentation refers to the process of partitioning the signatures into multiple strokes (i.e., segments). In our case, strokes are separated by discontinuities represented by pen-up events.

In the reminder of this section, we first describe the modes in which the system operates; then, we illustrate the verification algorithm. The same algorithm is used by the two modes: the main change is the way in which the algorithm is applied. We distinguish a segmented and an unsegmented mode.

- *Segmented Mode*. The matching is done by comparing each segment (or sub-sequence) of the signature to verify with the corresponding segment of the registered signature. This means that the verification algorithm is applied several times (for each pair of segments) and the final decision (acceptance or refusal) is taken according to the result of each segment pairing.
- *Unsegmented Mode*. The matching is done by comparing the whole sequence representing the signature to be verified with the corresponding registered signature. The verification algorithm is applied just once and the output is the final decision.

Now we turn to describe our verification algorithm, which is a scheme based on the Dynamic Time Warping (DTW)

algorithm. DTW is a popular and robust technique for comparing time series, capable of handling time shifting and scaling, which has been successfully used in literature for HSV (prevalently for online approaches like the methods proposed in [17], [18], but also for offline approaches such as the method described in [19]). We describe our verification stage for the segmented mode only, being the unsegmented mode a sub-case of the segmented mode in which the number of segments is exactly one.

The algorithm is as follows. First, for each feature f (such as X-Y coordinates, Pressure, X-Y Velocity, X-Y Acceleration) the following n-dimensional vector is computed, where n is the number of segments in which each of the three signatures is divided. S_j^i represents the i^{th} segment (related to feature f) of the j^{th} signature as a 1D time series, while $DTW(S_j^i, S_k^i)$ denotes the 1D Dynamic Time Warping method applied to the i^{th} segments of the j^{th} and k^{th} signatures.

$$\left\| \begin{matrix} f^1 \\ f^2 \\ ... \\ f^n \end{matrix} \right\| = \left\| \begin{matrix} \frac{DTW(S_1^1,S_2^1)+DTW(S_1^1,S_3^1)+DTW(S_2^1,S_3^1)}{3} \\ \frac{DTW(S_1^2,S_2^2)+DTW(S_1^2,S_3^2)+DTW(S_2^2,S_3^2)}{3} \\ ... \\ \frac{DTW(S_1^n,S_2^n)+DTW(S_1^n,S_3^n)+DTW(S_2^n,S_3^n)}{3} \end{matrix} \right\|$$

Then, once the $\|f\|$ vector is computed for each feature of interest, we get a $\|X\|$ and a $\|Y\|$ vector (x and y coordinates), a $\|P\|$ vector (pressure), a $\|V_x\|$ and a $\|V_y\|$ vector (velocity on x and y directions), a $\|A_x\|$ and a $\|A_y\|$ vector (acceleration on x and y directions).

Finally, we combine the metrics with the following weighted sums, by giving a weight to each of the kinds of signature features.

$$\left\| \begin{matrix} d^1 \\ d^2 \\ ... \\ d^n \end{matrix} \right\| = \left\| \begin{matrix} w_x \cdot X^1 + w_y \cdot Y^1 + w_p \cdot P^1 + ... + w_{a_y} \cdot A_y^1 \\ w_x \cdot X^2 + w_y \cdot Y^2 + w_p \cdot P^2 + ... + w_{a_y} \cdot A_y^2 \\ ... \\ w_x \cdot X^n + w_y \cdot Y^n + w_p \cdot P^n + ... + w_{a_y} \cdot A_y^n \end{matrix} \right\|$$

The weights w_x, w_y, w_p, ... , w_{a_y} must be experimentally determined as they are dependent on the device. For instance, even if the pressure change is generally a very discriminating feature (often leading to a high w_p coefficient), the influence of the capability of sensing pressure change (which is specific for each device) is significant and the weight should be lowered accordingly on low end devices.

The output distance vector $\|d\|$ represents the "distance" among the three signatures. The whole process is repeated twice; the first time using genuine registered signatures ($\|d_g\|$ as output), the second time using signatures to be verified ($\|d_v\|$ as output).

Finally, we compare the two distance vectors with each other. The similarity function is defined as follows.

$$Similarity\left(\left\| \begin{matrix} d_v^1 \\ ... \\ d_v^n \end{matrix} \right\|, \left\| \begin{matrix} d_g^1 \\ ... \\ d_g^n \end{matrix} \right\|\right) = \left(\sum_{i=1}^{n} F(d_v^i, d_g^i)/n\right)$$

where n is the number of segments and the $F()$ function is defined as follows (c is a tolerance coefficient).

$$f(d_v^i, d_g^i) = \begin{cases} 1 & \text{if } d_v^i \leq c \cdot d_g^i \\ 0 & \text{otherwise} \end{cases}$$

The higher the value of the similarity function, the more likely the claimed identity is correct. The final decision (accept or reject the claimed identity) depends on whether the similarity score is above or below of a given threshold.

Note that all the computation happens at verification time (including the tasks which process genuine signatures). We cannot move any computation at registration time, because the priority is to minimize the storage size of the output of the registration phase (to be embedded by means of barcodes).

IV. Experimentation

In this section we present experimental results concerning identity verification with our system. The accuracy of a recognition algorithm is generally measured in terms of two potential types of errors: false negatives (*fn*) and false positives (*fp*). False positives are cases where a claimed identity is accepted, but should not be, while false negatives are cases where a claimed identity is not accepted, while it should be. Two metrics building on true/false positives/negatives (*tp,fp,tn,fn*) are widely adopted: precision and recall. Recall ($tp/(tp+fn)$) is the probability that a valid identity is accepted by the system (i.e., true positive rate) while precision ($tp/(tp+fp)$) is the probability that a claimed identity which is accepted by the system is valid. F-measure (which is the harmonic mean of precision and recall) combines both metrics into a global measure ($f\text{-}measure = (2 \times prec \times recall)/(prec + recall)$). A more general f-measure is generally defined as function of a β parameter, which is used to weight f-measure in favor of precision ($\beta < 1$) or recall ($\beta > 1$).

A threshold on the similarity score must be identified for determining whether two signatures are similar (accept the identity) or significantly different (reject the identity). The higher the threshold, the higher the precision (i.e., the lower the risk of accepting invalid identities). However, a high threshold also decreases the recall of the system (i.e., the higher the risk to reject valid identities).

The performance of the proposed scheme has been assessed in terms of false positives, false negatives, precision, recall and f-measure on two kinds of dataset: first, on a standard dataset (i.e., the SVC database [20]), involving WACOM digitizing tablets, 100 sets of signature data, with 20 genuine signatures and 20 skilled forgeries for each set; secondly, on a custom dataset, built for this purpose using different smartphones (mainly from the Google Nexus family), involving 250 signatures partitioned into 5 sets of signature data, with 10 genuine signatures and 40 skilled forgeries for each set.

We start by describing the experimental set-up:

- For each user, 3 genuine signatures out of 20 (first dataset) or out of 10 (second dataset) were selected in rotation to form the template of the user.
- Every time a user template (building on 3 signatures) is selected, the remainder 7 genuine signatures were matched to the template itself in order to compute the false rejection rate (i.e., the false negatives rate) by means of the matching error. This process is repeated for every user.
- Given a user, the skilled forgeries (provided by users other than the one named) were matched to his/her template in order to compute the false acceptance rate (i.e., the false positives rate). This process is repeated for every user involved in the experiment.
- As for the first dataset, the SVC 2004 competition [20] consisted of two separate signature verification tasks, each of which was based on a different signature database. Each database of the SVC 2004 has 100 sets of signature data. Each set contains 20 genuine signatures from one signature contributor and 20 skilled forgeries from five other contributors. Contributors were asked to write on a digitizing tablet (specifically, a WACOM Intuos tablet).
- As for the second dataset, we used a custom dataset, built on data acquired by smartphones (i.e., no special purpose devices). We collected 250 handwritten signatures. Each user was requested to make genuine signature 10 times and to provide 10 (skilled) forgeries of any other user. The signatures were partitioned into 5 sets of signature data, where each set contains 10 genuine signatures of a specific user and 40 skilled forgeries of the signature of that user, produced by the other users involved in the experiment.

The experimental results in terms of precision, recall and f-measure (that vary according to the chosen thresholds) have been used for tuning the thresholds in order to get better performance (see Section IV-A). Then, once we fixed the threshold on the similarity score, we evaluated how subsampling the sequences (forming each signature) affected the overall precision and recall. This allowed us to identify the best trade-off between discrimination capabilities of the system and the storage size of the handwritten signature model, in order to ensure that the signature model fit into 2D barcodes (see Section IV-B).

The remainder of this section illustrates our results, which are in line with other work in the area, despite storage constrains due to barcode capacity and limitations in sensing and processing related to common devices such as smartphones and tablets.

A. Tuning the Thresholds to Enhance Precision and Recall

In this section we tune system thresholds by analyzing the curves of precision, recall and f-measure in order to get better performance (thresholds determine whether to accept or reject the claimed identity).

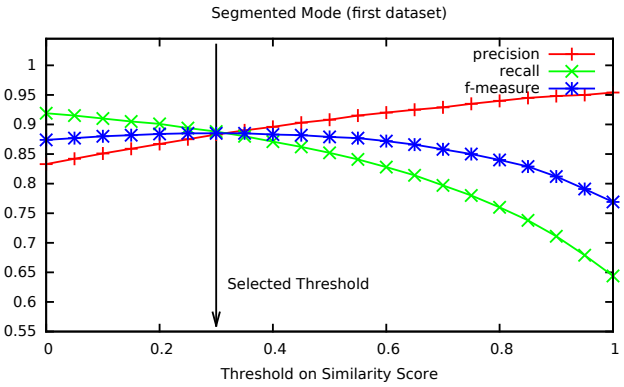


Fig. 5. 1^{st} dataset (SVC); segmented mode. Precision, recall and f-measure as functions of threshold on similarity score. (Viewed better in color).

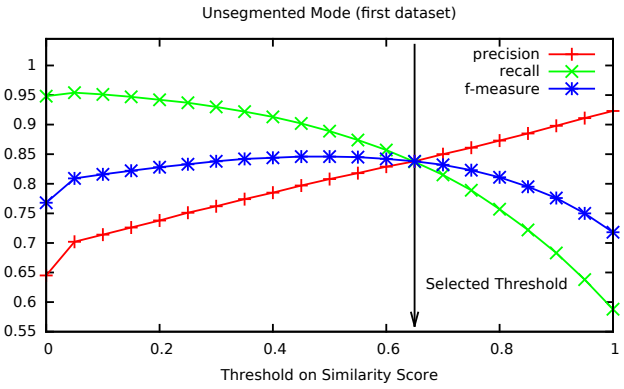


Fig. 6. 1^{st} dataset (SVC); unsegmented mode. Precision, recall and f-measure as functions of threshold on similarity score. (Viewed better in color).

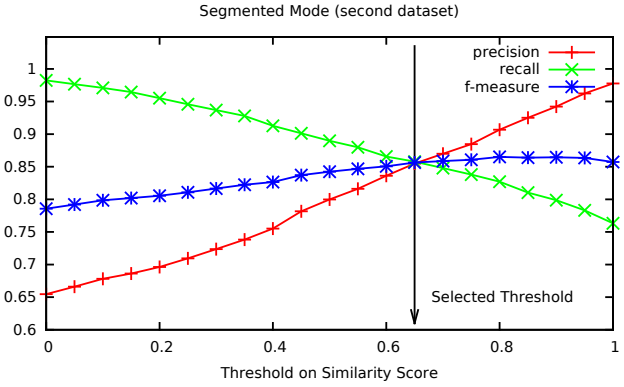


Fig. 7. 2^{nd} dataset; segmented mode. Precision, recall and f-measure as functions of threshold on similarity score. (Viewed better in color).

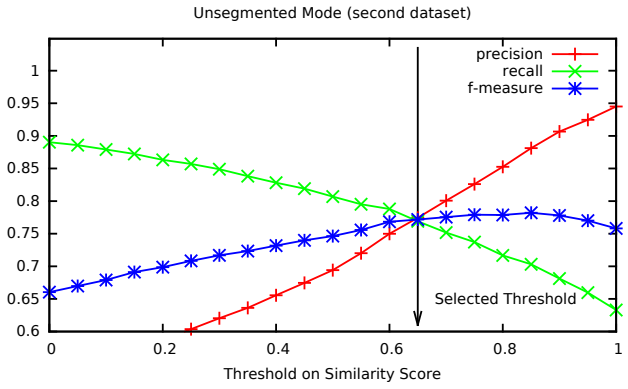


Fig. 8. 2^{nd} dataset; unsegmented mode. Precision, recall and f-measure as functions of threshold on similarity score. (Viewed better in color).

The four graphs of Figures 5, 6, 7 and 8 show the curves of precision, recall and f-measure as functions of threshold on the similarity score for each of the two working modes and for each of the two datasets. Claimed identities are accepted whenever the score is above the threshold, rejected otherwise. The higher the threshold, the higher the precision, but the lower the recall. The threshold which maximizes the f-measure is identified for each working mode and highlighted with an arrow in each graph. The best results (plotted in Figures 5 and 7) were achieved using the segmented mode. This is in agreement with the intuition that working at low granularity (i.e., comparing the subsequences segmented by pen-up events of a first signature with the corresponding subsequences of a second signature) leads to a more accurate comparison than just making a single comparison of the two signatures (considered in their entirety).

Precision and recall jointly reach a maximum of 89.50% in the graph of Figure 5 (SVC dataset). For this reason, at this operating point, where the curves of precision and recall cross each other, the false positive rate (also called the false acceptance rate - FAR) and the false negative rates (also called the false rejection rate - FRR) are both 10.50% as complementary values. This value is denoted as the equal error rate (EER), that is, the point where FAR equals FFR. We got good results, considering the EERs reported at the SVC competition: for instance, the EERs related to the SVC training set (with skilled, not random forgeries) range from a low of 5.50% to a high of 31.32% in the first verification task [21] and from 6.90% to 21.89% in the second verification task [22].

B. Trade-off between Precision/Recall and the Model Size

In this section we address the trade-off between precision and recall of the system and the space used for storing models of signatures. This allow us to show how to embed features extracted from handwritten signatures within the documents themselves by means of barcodes.

Consider that the size of a model depends on the number of samples with which we represent each handwritten signature. We expect that the more the space available for storing models of signatures, the more the precision and recall of the system are; at least, until a specific threshold value is reached, after which precision and recall remain almost constant. This is consistent with the intuition that getting more samples than needed (the limit is due to the precision of the acquiring devices) does not improve the overall performance.

Signatures data were subsampled as follows: 1:1 (all the samples captured by the device are kept), subsampled 2:1 (1 sample out of 2 is filtered out), subsampled 3:1 (2 samples out of 3 are filtered out), ..., subsampled 20:1 (19 samples out of 20 are discarded). For instance, from a signature which is 600 samples length we produce subsampled signatures whose length is 300 (subsampled 2:1), 200 (subsampled 3:1), ..., 30 (subsampled 20:1) samples. However, the actual signature length (expressed as number of samples) depends not only on the sampling rate of the device but also on the path length of the handwritten signature (expressed as length unit such as the millimeter).

In order to isolate the impact of the path lengths so that results do not depend on the length of the words forming the signature, we introduce the concept of samples density as the number of samples per unit length. In order to compute the samples density, the number of samples of a given signature is divided by the total path length of the signature itself. This is computed in pixels and is then converted from pixels to millimeters by referring to the number of points per inch (ppi) characterizing the device screen and to the (inches to millimeters) conversion factor. As a result, samples density is expressed here as number of samples per millimeter, which is a measure free from the signature word lengths and from device-dependent features such as the screen size.

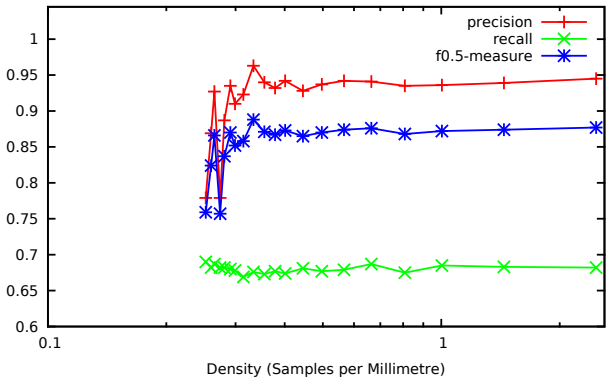


Fig. 9. Precision, recall and f-measure as functions of density (expressed as samples per millimeter) shown in log scale (Viewed better in color).

Figure 9 shows the precision, recall and f-measure of our system as functions of samples density. We used the f0.5-measure (with β equal to 0.5), which weights precision higher than recall. For our framework, increases in precision (that decrease the number of false positives) may be considered more important than increases in recall (that reduce the number of false negatives), since if a false instance is misclassified as

true (i.e., a false positive), a forgery is accepted as genuine, while, if a true instance is misclassified as false (i.e., a false negative), the user has only to re-enter the signature. The X-axis is on a log scale for visual clarity. On the far right side of the curves data are not subsampled (all the samples are kept), while the further we go towards the left side of the curves the more the data are subsampled (up to a maximum of 20:1). In absolute terms, the average path length of the signatures of our dataset was around 232 millimeters and the average number of samples (captured by the device for a single signature) was 581. Samples density ranges from around 2.5 to 0.125 samples per millimeter. This plot shows an interesting trend of decreasing precision and recall with decreasing samples density. The curves are almost constant initially (until a subsampling rate of 10:1 and a density of around 0.25), while at lower densities (or higher subsampling rates) the curves decrease sharply. This means that we are able to reduce the size of models of signatures by a factor of 10 (with respect to the sequence of samples acquired by the mobile device) without significant impacts on the overall precision and recall of the system. This, in turn, allows us to store models of signatures by means of barcodes, making our framework applicable to practical scenarios.

V. Conclusions

Our work presented a new HSV system for document signing and authentication, whose novelties lie mainly in the following aspects. First, we showed how to verify handwritten signatures so that signature dynamics can be processed during verification of every type of document (including paper documents). Secondly, we illustrated how to embed features extracted from handwritten signatures within the documents themselves, so that no remote signature database is needed. Thirdly, we proposed a method which is supported by a wide range of mobile devices so that no special hardware is needed. Finally, we showed how to reduce the size of models of signatures without significant impacts on the overall precision and recall of the system. In our experiments, Precision and recall cross at 89.50 (first dataset) and at 85.15% (second dataset). This is an interesting result, if noting that we used mobile devices (that is, no special-purpose hardware) to capture the signature dynamics needed by our experiments.

Acknowledgment

This work has been partially supported by Filas S.p.A. (Rome, Italy; regional agency for the promotion of development and innovation; under project MYME - My Mobile Enterprise).

References

[1] Y. Qiao, J. Liu, and X. Tang, "Offline signature verification using online handwriting registration," in *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07*. IEEE, 2007. doi: 10.1109/CVPR.2007.383263 pp. 1–8. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2007.383263

[2] M. K. Kalera, S. Srihari, and A. Xu, "Offline signature verification and identification using distance statistics," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 18, no. 07, pp. 1339–1360, 2004. doi: 10.1142/S0218001404003630. [Online]. Available: http://dx.doi.org/10.1142/S0218001404003630

[3] A. K. Jain, F. D. Griess, and S. D. Connell, "On-line signature verification," *Pattern recognition*, vol. 35, no. 12, pp. 2963–2972, 2002.

[4] Xyzmo, "Xyzmo signature solution," http://www.xyzmo.com/en/products/Pages/Signature-Verification.aspx, 2013, [Online; accessed 01-April-2014].

[5] SutiDSignature, "SutiDSignature," http://www.sutisoft.com/sutidsignature, 2013, [Online; accessed 01-April-2014].

[6] Andxor Corporation, "View2sign," http://www.view2sign.com/supported-signatures.html, 2013, [Online; accessed 01-April-2014].

[7] J. Trevathan and A. McCabe, "Remote handwritten signature authentication," in *ICETE*. Citeseer, 2005, pp. 335–339.

[8] M. Mailah and B. H. Lim, "Biometric signature verification using pen position, time, velocity and pressure parameters," *Jurnal Teknologi*, vol. 48, no. 1, pp. 35–54, 2012. doi: 10.11113/jt.v48.218. [Online]. Available: http://dx.doi.org/10.11113/jt.v48.218

[9] The Guardian, "NSA Prism program taps in to user data of Apple, Google and others," http://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data, 2013, [Online; accessed 01-April-2014].

[10] P. Kumar, S. Singh, A. Garg, and N. Prabhat, "Hand written signature recognition & verification using neural network," *International Journal*, vol. 3, no. 3, 2013.

[11] A. Pansare and S. Bhatia, "Handwritten signature verification using neural network," *International Journal of Applied Information Systems*, vol. 1, pp. 44–49, 2012. doi: 10.5120/ijais12-450114. [Online]. Available: http://dx.doi.org/10.5120/ijais12-450114

[12] M. Querini, A. Grillo, A. Lentini, and G. Italiano, "2D color barcodes for mobile phones," *International Journal of Computer Science and Applications (IJCSA)*, vol. 8, no. 1, pp. 136–155, 2011.

[13] A. Grillo, A. Lentini, M. Querini, and G. Italiano, "High capacity colored two dimensional codes," in *Proceedings of the 2010 International Multiconference on Computer Science and Information Technology (IMCSIT)*. IEEE, 2010, pp. 709–716.

[14] M. Querini and G. F. Italiano, "Color classifiers for 2D color barcodes," in *Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2013, pp. 611–618, full version submitted to the special issue of the conference in the Computer Science and Information Systems (ComSIS) journal.

[15] ——, "Reliability and data density in high capacity color barcodes," *Computer Science and Information Systems (ComSIS)*, 2014.

[16] D. Firmani, G. F. Italiano, and M. Querini, "Engineering color barcode algorithms for mobile applications," in *13th International Symposium on Experimental Algorithms (SEA)*, 2014.

[17] M. Faundez-Zanuy, "On-line signature recognition based on VQ-DTW," *Pattern Recognition*, vol. 40, no. 3, pp. 981–992, 2007. doi: 10.1016/j.patcog.2006.06.007. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2006.06.007

[18] O. Miguel-Hurtado, L. Mengibar-Pozo, M. G. Lorenz, and J. Liu-Jimenez, "On-line signature verification by dynamic time warping and Gaussian mixture models," in *International Carnahan Conference on Security Technology*. IEEE, 2007. doi: 10.1109/CCST.2007.4373463 pp. 23–29. [Online]. Available: http://dx.doi.org/10.1109/CCST.2007.4373463

[19] A. Piyush Shanker and A. Rajagopalan, "Off-line signature verification using DTW," *Pattern Recognition Letters*, vol. 28, no. 12, pp. 1407–1414, 2007.

[20] SVC, "Signature Verification Competition," http://www.cse.ust.hk/svc2004/, 2013, [Online; accessed 01-April-2014].

[21] ——, "First Task Results," http://www.cse.ust.hk/svc2004/results-EER1.html, 2013, [Online; accessed 01-April-2014].

[22] ——, "Second Task Results," http://www.cse.ust.hk/svc2004/results-EER1.html, 2013, [Online; accessed 01-April-2014].

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 709–717

DOI: 10.15439/2014F282
ACSIS, Vol. 2

Movement Tracking in Terrain Conditions Accelerated with CUDA

Piotr Skłodowski
Cybernetics Faculty at Military
University of Technology
ul. S. Kaliskiego 2,
00-908 Warsaw, Poland
Email: psklodowski@wat.edu.pl

Witold Żorski
Cybernetics Faculty at Military
University of Technology
ul. S. Kaliskiego 2,
00-908 Warsaw, Poland
Email: wzorski@wat.edu.pl

***Abstract*— The paper presents a solution to the problem of movement tracking in images acquired from video cameras monitoring outside terrain. The solution is resistant to such adverse factors as: leaves fluttering, grass waving, smoke or fog, movement of clouds etc. The presented solution is based on well known image processing methods, nevertheless the key was the use of an appropriate conduct procedure. In order to obtain a real-time system the CUDA technology was involved.**

I. Introduction

THE problem of movement detection in images [4] appeared relatively early [6]. The astronomy world struggled with objects detection in images [8] of the night sky acquired by telescopes long before the era of modern computers. In first systems images were alternately displayed in front of an operator who was able to perform detection of a motion celestial body. In such systems a natural subconscious human ability of movement detection was involved [9], [10].

Excluding cheap and simple movement detectors or sensors (passive infrared, ultrasonic, or microwave) the task of motion detection with the use of video cameras [11] is based on digital image processing [5]. Many present-day computer systems begin the work from the stage of a differential image of two images [14], which next undergoes a series of processes [16]. The detection of a movement [3] is not the only result – in modern systems a trajectory of a motion body can be determined [1] or even identification of the detected object may be performed [2], [22].

In this paper we consider the problem of object movement tracking [7] in images acquired from video cameras [12] monitoring outside terrain [13]. The assumption was to elaborate a solution resistant to such adverse factors as: leaves fluttering, grass waving, smoke or fog, movement of clouds etc. A set of well known image processing methods [19] is adopted, and the key was the use of an appropriate conduct procedure. In order to obtain a real-time system the CUDA technology was involved.

The CUDA (*Compute Unified Device Architecture*) technology appeared quite unexpectedly in 2007 as a result of new Nvidia's GPUs branded GeForce 8. CUDA gave the software developers direct access to the virtual instruction set and memory of the parallel computational elements in GPUs.

CUDA is a parallel computing platform and programming model [24] that makes using a GPU for general purpose computing simple and elegant. At present, there are two main CUDA architectures available: Fermi (see Fig. 1) and Kepler. The Maxwell architecture (20 nm technology node) is just about to be launched onto the market. From the programmer's point of view [25] the new architecture brings a set of features, both hardware and software, that is known as the compute capability of a device.

The idea of combining image processing methods or computer vision techniques with CUDA technology started relatively early [15] and going on, being very popular.

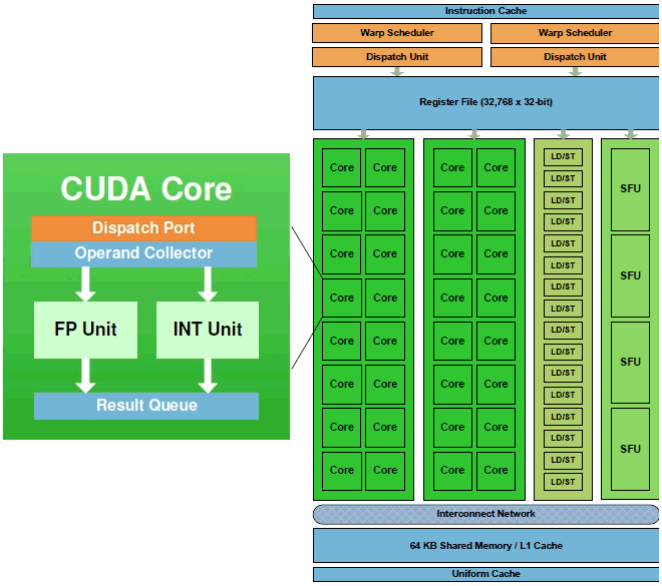


Fig. 1. CUDA core and Fermi SM (Streaming Multiprocessor) structure

978-83-60810-58-3/$25.00

This paper presents an announced earlier solution to movement detection and tracking, first elaborated using the Matlab environment, and finally independently implemented as the x86 and CUDA application.

The method was originally prepared for monitoring an airport's terrain, but for obvious reasons only neutral shots will be presented.

II. A Brief Presentation of the System

The used computer vision system consists of a PC equipped with a CUDA device (GTX 650 Ti, based on the Nvidia's Kepler architecture with compute capability 3.0), and an IP camera. It is supported by the Microsoft Visual Studio 2012 and CUDA 5.5 framework (the most important of it is the CUDA Toolkit component). Fig. 2 shows a visual scheme of the used computer vision system.

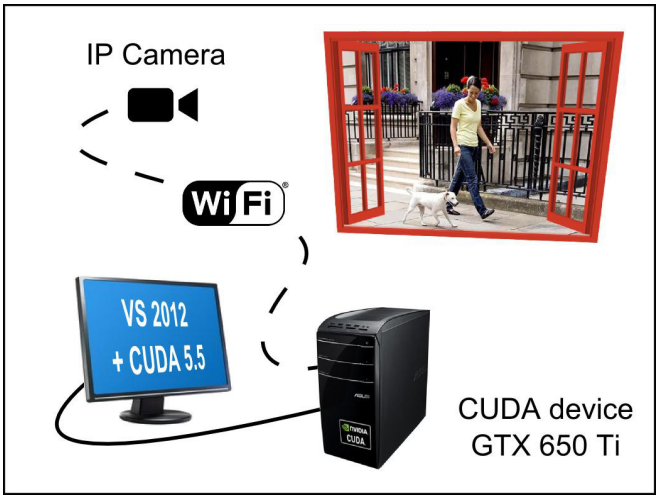


Fig. 2. Scheme of the used system

III. Matlab Implementation

The Matlab environment gives a possibility to elaborate the required procedure relatively fast. The amount of engineering tools included in the Matlab is impressive, nevertheless to obtain required speed the final implementation must be done with the CUDA technology.

A. Example source scenes

Source scenes have been acquired under various terrain and weather conditions. In this section two examples are presented (see Fig. 3). The first scene includes an object that is well visible, but also includes waving grass and clouds. The second scene is much more difficult, the object is comparatively small and there is a big tree with fluttering leaves. In both scenes a slight tilt effect is present between shots captured over a distance of a few seconds.

Fig. 3. The input source scenes

B. Compensation of the tilt effect

The initial obstacle is the tilt effect between shots, which may occur as a result of small vibration under the influence of wind or some mechanical reasons. To compensate the tilt one image is narrowed about a "frame" and matched with the second image in order to find a location with the smallest difference. Finally images of the scene are "framed" to guarantee the smallest difference between them. The source code in Fig. 4 gives details of the procedure. Fig. 5 shows (only) the cropping effect in the case of the second considered scene. The result will be visible in the case of difference images (the next section).

```
26    %IMAGE STABILIZATION by shift compensation
27 -  nop=5; %number of pixels (of the frame)
28 -  I1=I_in1(:,:,1);
29 -  I2=I_in2(:,:,1);
30 -  I2_framed=I2(nop+1:row-nop,nop+1:col-nop); %a "framed" image
31 -  d=255*row*col; bi=0; bj=0;
32    %searching for the smallest difference within range +-nop
33 -  for i=-nop:nop
34 -      for j=-nop:nop
35 -          I1_framed=I1(nop+1+i:row-nop+i,nop+1+j:col-nop+j);
36 -          difference=sum(sum(abs(I1_framed-I2_framed)));
37 -          if difference<d
38 -              bi=i; bj=j; d=difference;
39 -          end;
40 -      end;
41 -  end;
42 -  I1_framed=I1(nop+1+bi:row-nop+bi,nop+1+bj:col-nop+bj);
```

Fig. 4. Compensation of the tilt effect – the source code

Fig. 5. The cropping effect of the tilt compensation – frames are visible

C. Getting a difference image

The difference image generation [16] is the first processing stage for a scene. This approach is extremely popular in astronomy [17] and is commonly referred to as difference image analysis (DIA). Results (presented in negative) obtained for the considered scenes (after the tilt compensation) are visible in Fig. 6 and Fig. 7.

Fig. 6. The difference image for the first scene in Fig. 3

Fig. 7. The difference image for the second scene in Fig. 3

D. Removing unwanted artifacts

The received differenced images include moving objects as well as include some unwanted artifacts. In the case of the first scene (Fig. 6) a remnant of the tilt effect is still visible (e.g. contour of a building), and in the second scene (Fig. 7) a tree is well exposed. Some of the artifacts are heavy, what is shown in Fig. 8, which is a 3D visualization of the content of Fig. 7.

At the first glance the task of removing unwanted artifacts seems to be difficult. To solve the problem it is necessary to notice that tracked objects generate comparatively low frequencies and the unwanted artifacts rather high frequencies (see Fig. 8). As an outcome of many trials it turned out that erosion, a fundamental operation of morphological image processing [18], gives the best results.

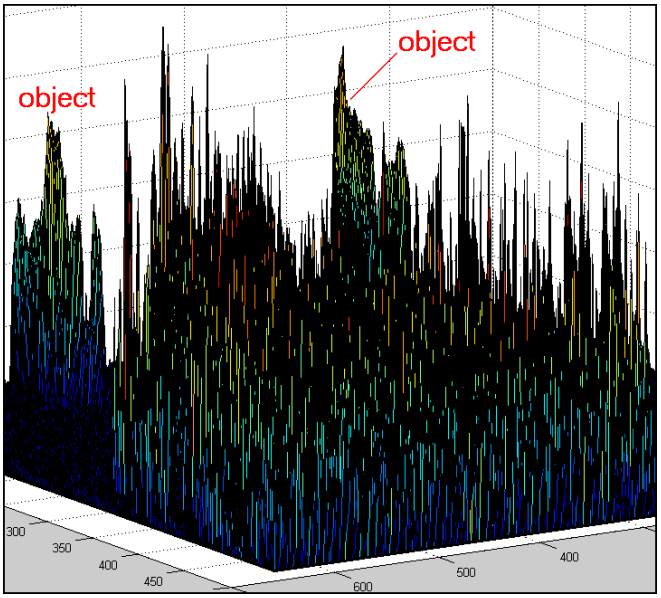


Fig. 8. A 3D visualization of the difference image from Fig. 7

The erosion operation is already available in Matlab as ready to use function (erode, imerode). Nevertheless, it was implemented "step by step" for gray-level images with the prospective x86 and CUDA implementations in mind. The source code in Fig. 9 shows details of the erosion implementation with a disk as the structuring element.

```
64    %erosion filter - "step by step"
65 -  SE=[  0     0     1     0     0;
66          0     1     1     1     0;
67          1     1     1     1     1;
68          0     1     1     1     0;
69          0     0     1     0     0]
70 -  [row col]=size(I_out1(:,:,1));
71 -  V=256*ones(5,5);
72 -  for i=1+2:row-3
73 -      for j=1+2:col-3
74 -          for k=-2:2
75 -              for l=-2:2
76 -                  if SE(k+3,l+3)==1
77 -                      V(k+3,l+3)=I_out1(i+k,j+l,1);
78 -                  end;
79 -              end;
80 -          end;
81 -          I_out2(i,j,1)=min(min(V));
82 -      end;
83 -  end;
84 -  I_out2(:,:,2)=I_out2(:,:,1);
85 -  I_out2(:,:,3)=I_out2(:,:,1);
```

Fig. 9. Matlab implementation of the erosion for gray-level images

The results obtained with the erosion are shown in Fig. 10 and Fig. 11. The objects are still well visible and the artifacts are predominantly filtered.

Fig. 10. The result after erosion for the first scene (compare Fig. 6)

Fig. 11. The result after erosion for the second scene (compare Fig. 7)

E. Exposing objects

The use of erosion filtering was beneficial to objects detection. It turned out that objects can be further exposed with the use of low-pass filtering. There are two possibilities: simple spatial filtering (neighborhood averaging) with a large mask 7x7 or just the standard transform FFT2. The second tool is faster and already available in majority of programming environments (including CUDA). In the case of Matlab we have two-dimensional convolution `conv2` and set of tools for two-dimensional discrete Fourier transform: `fft2`, `ifft2`, `fftshift`.

The process of FFT2 filtering is shown in Fig. 12, and a 3D result for the first scene is visible in Fig. 13, and for the second scene is presented in Fig. 14.

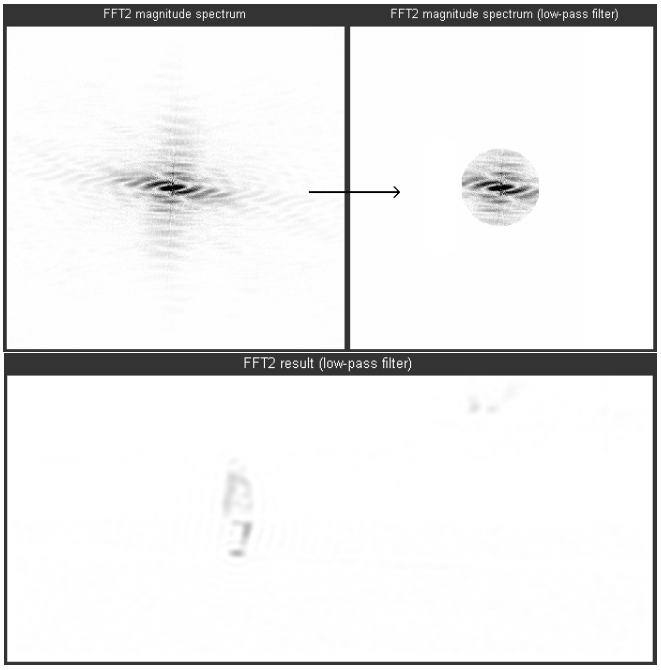


Fig. 12. The use of the FFT2 for the first scene (compare Fig. 10)

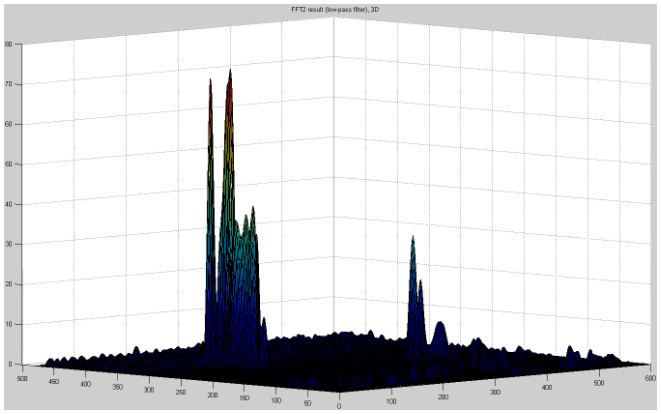


Fig. 13. The result of FFT2 for the first scene (see Fig. 12)

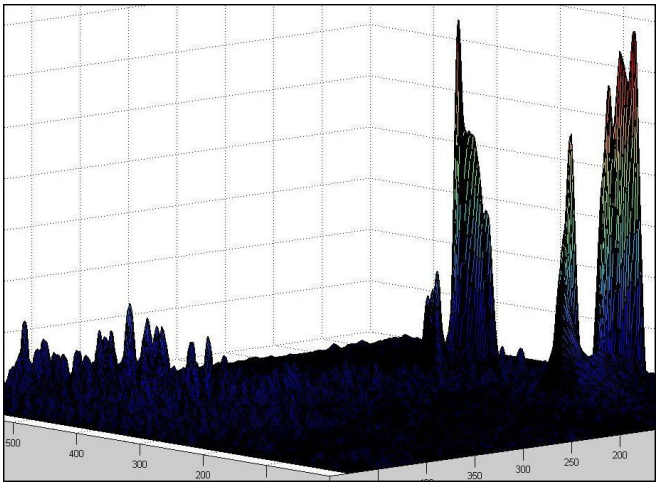


Fig. 14. The result of the FFT2 for the second scene (see Fig. 11)

F. Binarization and the decision

The results visible in Fig. 13 and 14 are rather satisfying. The last stage before the final decision about movement detection is binarization. The best results of binarization were received for threshold from the range of **20-50** of gray levels. The results of binarization are presented in Fig. 15.

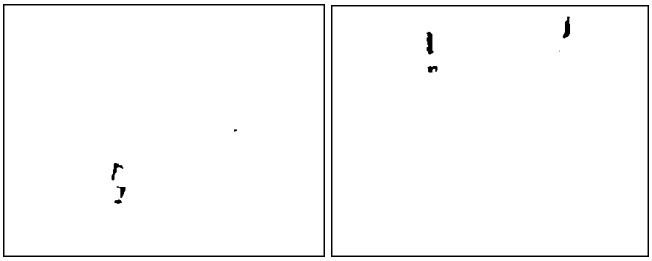


Fig. 15. The result of binarization for the considered scenes

The final decision about the movement detection is based on percentage size of objects in frames. In the case of Fig. 15 the percentage sizes of objects are respectively: 0,15% and 0,32%. The established threshold for the elaborated method is **0,1%**.

IV. x86 Implementation

The x86 implementation of the elaborated method has been made in C++ with the use of Visual Studio 2012. To speed up the implementation process the well known and free library called OpenCV was used. The library contains a set of ready to use computer vision algorithms (e.g.: linear filtering, cosine transform) as well as basic image processing functions (read/write images, conversion). Custom implementation has been made only for elements that are not included in the OpenCV or those which are poor optimized for the considered application.

C++ language has been chosen for both: x86 and CUDA implementation. Thanks to that it was possible to use exactly the same template project and therefore the execution is not overwhelmed by any language runtime.

A. Example source scenes

Source scenes were captured from an IP camera, flipped horizontally and then converted to 8 bits gray-scale images.

B. Compensation of the tilt effect

The tilt reduction function has been implemented independently in accordance with proposed algorithm and shown in following listing (Fig. 16). This function requires two images which we call “previous” and “current” frame. The previous frame is the frame captured first.

The only parameter required is *nop*. The *nop* stands for number of pixels. In our implementation we used constant value 5 which means that the previous frame is cropped by 5 pixels from all sides and the current frame is centered this way that the tilt effect (the images deference) to the previous frame is the smallest.

```
void Movement::tilt_reduction()
{
    int iw = this->iw - 2*nop,
        ih = this->ih - 2*nop,
        is = iw * ih;

    // crop the previous frame
    Mat roi_im1 = im1(Rect(nop, nop, iw, ih));

    int by = 0, bx = 0;
    int min = is*256;
    Npp8u *px;

    // find the best position in the current frame
    // where tilt is the lowest
    for (int y = 0; y < nop*2+1; y++)
        for (int x = 0; x < nop*2+1; x++)
        {
            Mat a = abs(roi_im1 - im2(Rect(x, y, iw, ih)));
            Scalar ssum = sum(a);
            int sum = ssum[0];
            if (sum < min)
            {
                by = y;
                bx = x;
                min = sum;
            }
        }

    Mat tmp_im = Mat::ones(this->ih, this->iw, CV_8U);
    Mat tmp_roi = tmp_im(Rect(nop, nop, iw, ih));

    // set "the previous frame"
    roi_im1.copyTo(tmp_roi);
    tmp_im.copyTo(im1);

    // set "the current frame"
    Mat roi_im2 = im2(Rect(bx, by, iw, ih));
    roi_im2.copyTo(tmp_roi);
    tmp_im.copyTo(im2);
};
```

Fig. 16. Compensation of the tilt effect in C++

C. Getting a difference image

The pixels from previous frame are subtracted from current frame then provided as an argument of *abs* function. The difference image is getting very easy using a ready function *abs* from OpenCV library and is coded as one line in movement detection function (see Fig. 17).

```
int Movement::detection(Mat &frame, Mat &result)
{
    frame.copyTo(im2);

    tilt_reduction();
    diff = abs(im1 - im2);
    erosion();
    low_pass_filter();
    binarization();

    im2.copyTo(im1);
    diff.copyTo(result);
    return 0;
}
```

Fig. 17. Movement detection in C++

D. Removing unwanted artifacts

To remove unwanted artifacts that might still persist in the processed image the erosion operator is applied. This has been made using our own implementation because we found it much faster than the option provided by OpenCV. The structuring element used in our implementation is disk inscribed in 5x5 matrix (see Fig. 18).

```
void Movement::erosion()
{
    unsigned disk[5][5] = {
        { 0, 0, 1, 0, 0 },
        { 0, 1, 1, 1, 0 },
        { 1, 1, 1, 1, 1 },
        { 0, 1, 1, 1, 0 },
        { 0, 0, 1, 0, 0 }
    };

    Mat out = Mat::zeros(ih, iw, CV_8U);
    Npp8u *data = (Npp8u *) diff.data,
        *out_d = (Npp8u *) out.data;

    for (int iy = 0; iy < ih-4; iy++)
        for (int ix = 0; ix < iw-5; ix++)
        {
            int min = 255, val;
            for (int dy = 0; dy < 5; dy++)
                for (int dx = 0; dx < 5; dx++)
                {
                    if (disk[dy][dx] == 1)
                    {
                        val = *(data + ((iy + dy) * iw + ix + dx));
                        if (val < min)
                        {
                            min = val;
                        }
                    }
                }
            *(out_d + (iy + 2) * iw + ix + 2) = min;
        }

    out.copyTo(diff);
};
```

Fig. 18. Erosion function in C++

E. Exposing objects

The last operation applied to the image before movement detection is convolution with 7x7 kernel of all ones.

```
void Movement::low_pass_filter()
{
  Mat kernel = Mat::ones(lowPassFilter, lowPassFilter, CV_64F), tmp1, tmp2;
  double mean = double(lowPassFilter * lowPassFilter);
  diff.convertTo(tmp1, CV_64F, 1./255);
  filter2D(tmp1, tmp2, -1, kernel);
  tmp2 = tmp2 / mean;
  tmp2.convertTo(tmp2, CV_8U, 255);
  tmp2.copyTo(diff);
}
```

Fig. 19. Low-pass filtering in C++

F. Binarization and the decision

```
void Movement::binarization()
{
    Mat to_compare = Mat::ones(ih, iw, CV_8U) * threshold;
    Mat out;
    compare(diff, to_compare, out, CMP_GT);
    out.convertTo(out, CV_8U);
    out.copyTo(diff);
}
```

Fig. 20. Binarization in C++

The result of all previous steps is still an grayscale image. Applying the threshold we got the final binarized image ready for the final step. This has been also achieved using one line ready to use function (see Fig. 20).

V. CUDA Implementation

Most of operations in proposed algorithm are available in Nvidia Performance Primitives [26]. The NPP is a collection of GPU-accelerated functions for image, video and signal processing. The library is freely available as a part of the CUDA Toolkit.

A. Use of the CUDA device structure

The only function that needed to be implemented independently was the tilt reduction. We couldn't match any function from NPP that would help us to achieve desired results therefore an own kernel has been implemented.

Although CUDA device allows to organize threads in 3D structure, 2D structure was enough. The X and Y axes responds to the position of pixels in the image. Block Index address pixels from “the previous” frame. Pixels from “the current” frame are further offset by the Grid Index. That makes two regions of interest (ROI) for each kernel iteration as shown in Fig. 21.

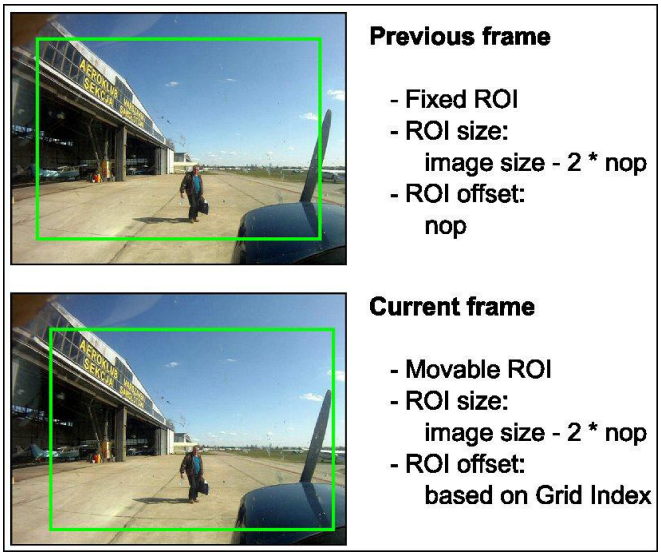


Fig. 21. Tilt reduction with regions of interest (ROI) in CUDA

Grid Size corresponds to *nop* parameter from the x86 implementation. Block Size is a parameter chosen empirically and have to be power of 2 for further reduction process. The source code of the CUDA kernel is presented in Fig. 22.

```
__global__ void tilt_reduction_kernel(Npp8u *im1, Npp8u *im2,
                                      int iw, int ih, int nop, int *ret)
{
    int iww = iw - 2*nop,
        ihh = ih - 2*nop;

    int thread = threadIdx.y * blockDim.x + threadIdx.x;

    __shared__ int sums[256];
    sums[thread] = 0;

    int offset1, offset2;
    int sum = 0;

    int iy = threadIdx.y,
        ix = threadIdx.x;

    while (ix < iww && iy < ihh)
    {
        // Offset in "the previous" frame
        offset1 = (iy + nop) * iw + ix + nop;
        // Offset in "the current" frame
        offset2 = (iy + blockIdx.y) * iw + ix + blockIdx.x;

        // Check the pixels difference
        int value = abs(*(im1 + offset1) - *(im2 + offset2));
        // Sum all differences for given ROI
        sum += value;

        ix += blockDim.x;
        if (ix >= iww)
        {
            ix = threadIdx.x;
            iy += blockDim.y;
        }
    }
    sums[thread] = sum;
    __syncthreads();

    // Reduction
    for (int i = blockDim.x * blockDim.y/2; i > 0; i /= 2)
    {
        if (thread < i)
        {
            sums[thread] += sums[thread + i];
        }
        __syncthreads();
    }

    // Save the result for given ROI
    if (thread == 0)
    {
        int return_offset = blockIdx.y * gridDim.x + blockIdx.x;
        *(ret + return_offset) = sums[0];
    }
}
```

Fig. 22. The source code for the CUDA kernel

B. CUDA implementation supported by the NPP library

The use of NPP library is relatively simply. The major difficulty is a preparation of the image data accordingly to NPP requirements. The NPP supports variety of data. Pixels may be provided as 8, 16 or 32 bits signed or unsigned integers or 32 bits floating point numbers. Unfortunately some functions don't support all data types. The choose should be made base on a function availability that need to be used.

What one need to remember is that the NPP is mainly C library. It is a reason that some features like function overloading are not available. One need to use functions that exactly match parameters types. To help to recognize functions a special function name convention has been introduced. Each NPP function begins with nppi. The data type that the function is dedicated for might be distinguish by its suffix. For example suffix R indicates the primitive operates only on a rectangular. Suffix I indicates that the primitive works "in-place". This is well described in the NPP documentation [26].

The image that is passed to the NPP is always described by three parameters: pointer to the image, image size (as ROI), and line step. Pointer to the image has to be the CUDA device pointer. Line step is the number of bytes between successive rows in the image. Fig. 23 shows the use of the NPP library.

```
void Movement::erosion()
{
    NppiSize oMaskSize = { 5, 5 };
    NppiPoint oAnchor = { 3, 3 };
    nppiErode_8u_C1R(pdDiff, iw, pdTmp, iw, oSizeROI,
        pdErodeMask, oMaskSize, oAnchor);
    swap(&pdTmp, &pdDiff);
};

void Movement::low_pass_filter()
{
    NppiSize oKernelSize = { lowPassFilter, lowPassFilter };
    NppiPoint oAnchor = { lowPassFilter, lowPassFilter };
    NppiSize oDstSizeROI = { iw + lowPassFilter, ih + lowPassFilter };
    int nDstStep = oDstSizeROI.width;
    int offset = lowPassFilter - lowPassFilter / 2;

    nppiCopyReplicateBorder_8u_C1R(pdDiff, iw, oSizeROI, pdTmp,
        nDstStep, oDstSizeROI, offset, offset);
    nppiFilter_8u_C1R(pdTmp, nDstStep, pdDiff, iw, oSizeROI,
        pdLPKernel, oKernelSize, oAnchor, lowPassFilter * lowPassFilter);
}

void Movement::binarization()
{
    nppiCompareC_8u_C1R(pdDiff, iw, (Npp8u) threshold, pdTmp,
        iw, oSizeROI, NPP_CMP_GREATER);
    swap(&pdTmp, &pdDiff);
}

void Movement::swap(Npp8u **a, Npp8u **b)
{
    Npp8u * tmp = *a;
    *a = *b;
    *b = tmp;
}
```

Fig. 23. CUDA implementation using the NPP library

VI. Movement Tracking

A basic extension to the issue of movement detection is the problem of object tracking. The simplest way of tracking can be performed by drawing a trajectory (a path) for the detected object as shown in Fig. 24-26.

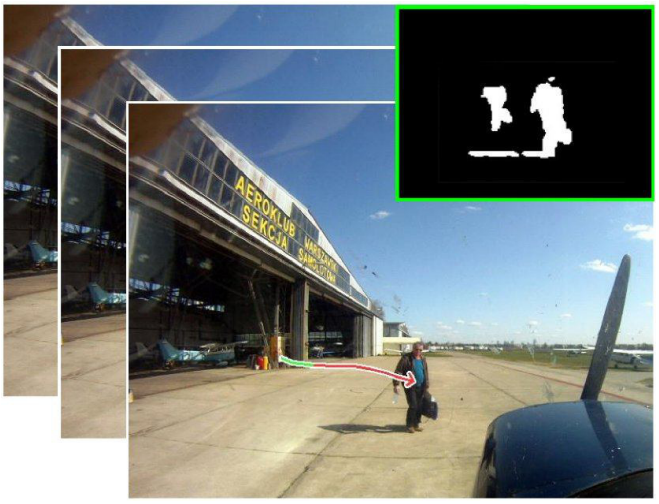


Fig. 24. An example of movement tracking in terrain conditions

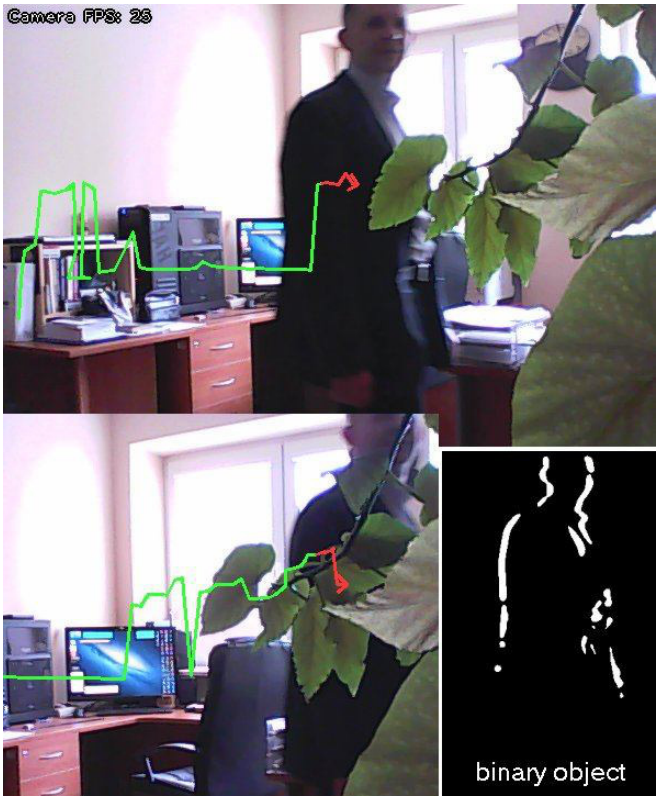


Fig. 25. An example of movement tracking inside a room

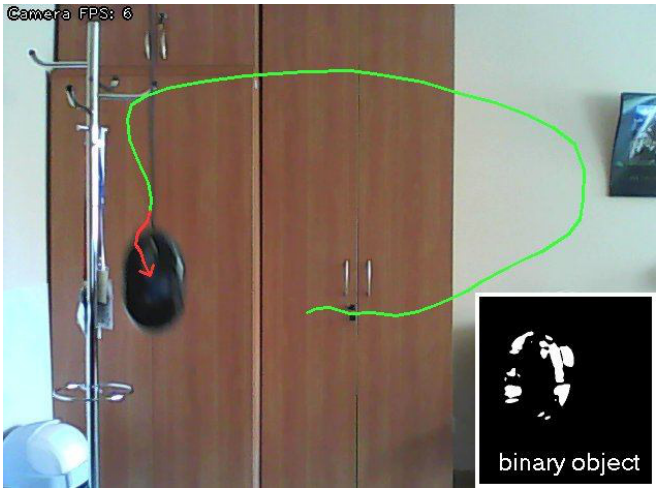


Fig. 26. An example of movement tracking of a small object (a mouse)

The suggested method allows to track only one object that is being detected. The object to be detected must occupy more than **0.1%** space of the binarized image. If that happens the object is surrounded by a rectangle and then center of its mass is calculated. If movement is detected in following frames then the track is plotted by joining calculated centers.

Aside of common template project, both implementations use the same algorithm for the considered stage, i.e. movement tracking. This is because of the algorithm simplicity which cause the CUDA implementation unnecessary. Thus, the function is common for x86 and CUDA implementation (see Fig. 27).

```
21  void process_result(Mat &result, t_detected_object &obj)
22  {
23      int iw = result.cols, ih = result.rows, is = iw * ih;
24      Npp8u * im = (Npp8u *) result.data;
25
26      obj.pt1.x = iw;
27      obj.pt1.y = ih;
28      obj.pt2.x = -1;
29      obj.pt2.y = -1;
30
31      int cnt = 0;
32      for (int y=0; y<ih; y++)
33          for (int x=0; x<iw; x++)
34          {
35              int offset = y * iw + x;
36              if (*(im + offset))
37              {
38                  cnt++;
39                  if (obj.pt1.x > x) obj.pt1.x = x;
40                  if (obj.pt1.y > y) obj.pt1.y = y;
41                  if (obj.pt2.x < x) obj.pt2.x = x;
42                  if (obj.pt2.y < y) obj.pt2.y = y;
43              }
44          }
45
46      if ((double) cnt / is > 0.001)
47      {
48          obj.detected = cnt;
49          obj.center.x = obj.pt1.x + (obj.pt2.x - obj.pt1.x) / 2;
50          obj.center.y = obj.pt1.y + (obj.pt2.y - obj.pt1.y) / 2;
51      }
52      else
53      {
54          obj.detected = 0;
55      }
56  }
```

Fig. 27. Movement tracking in C++

VII. Conclusion

There are two gains of the performed work that are fully concordant to the title of this paper: the elaboration of the method of movement tracking and its implementation in CUDA. The elaborated method can be described as a sequence of actions, what is shown in Fig. 28.

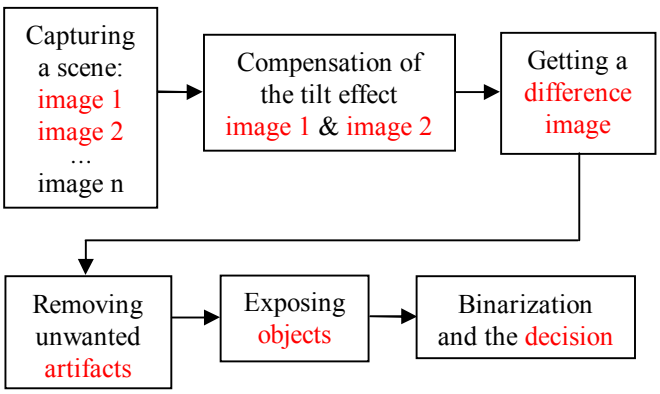


Fig. 28. A block diagram of processes for the elaborated method

CUDA and x86 implementations of the method were examined in details and optimized to receive the best performance. Performed benchmarks concerned only on selected portion of the source code directly responsible for

movement detection and tracking. Functions common for both implementations has been omitted in benchmarking.

The presented solution shows a significant performance difference between the implementation for x86 and the massively parallel implementation in CUDA. Both implementations give the same final result – confirmation that solution is correct. Performed benchmarks demonstrated substantial acceleration thanks to CUDA implementation which is suitable for a **real-time system**. It was possible to reach the speed of **25+ fps** for resolution **640x480**, at least **10 times faster** then in the case of x86 implementation. The upper limit velocity of tracked objects for the elaborated method is **4 m/s**. It is the outcome of the distance between adjacent frames. The lower limit velocity of tracked objects can be widely adjusted by the distance between analyzed frames.

In order to enrich the method an extension about identification of the detected object may be added using a method similar to one described in [23]. Another challenge is the problem of tracking multiple independent objects [20], [21].

REFERENCES

[1] M. Andriluka, S. Roth, B. Schiele, *People-tracking-by-detection and people-detection-by-tracking*, Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008, pp. 1-8. http://dx.doi.org/10.1109/CVPR.2008.4587583

[2] A. Bugeau, P. Perez, *Detection and segmentation of moving objects in highly dynamic scenes*, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007, pp. 1-8. http://dx.doi.org/10.1109/CVPR.2007.383244

[3] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V. K. Papastathis, M. G. Strintzis, *Knowledge-assisted semantic video object detection*, IEEE Transactions on Circuits and Systems for Video Technology Vol. 15, (10) 2005, pp. 1210–1224. http://dx.doi.org/10.1109/TCSVT.2005.854238

[4] Guofeng Zhang, Jiaya Jia, Wei Xiong, Tien-tsin Wong, Pheng-ann Heng, Hujun Bao: *Moving object extraction with a hand-held camera*, IEEE International Conference on Computer Vision, 2007, pp. 1-8. http://dx.doi.org/10.1109/ICCV.2007.4408963

[5] M. Heikkila, M. Pietikainen, *A texture-based method for modeling the background and detecting moving objects*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 28, (4) 2006, pp. 657–662. http://dx.doi.org/10.1109/TPAMI.2006.68

[6] R. Jain, H. Nagel, *On the analysis of accumulative difference pictures from image sequence of real world scenes*. IEEE Trans. Pattern Anal. Machine Intell., Vol. 1 (2) 1979, pp. 206-214. http://dx.doi.org/10.1109/TPAMI.1979.4766907

[7] Alper Yilmaz, Omar Javed, Mubarak Shah, *Object Tracking: A Survey*. ACM Computing Surveys, Vol. 38, No. 4, Article 13, Publication date: December 2006. http://doi.acm.org/10.1145/1177352.1177355

[8] K. J. Meech, *Astronomical image processing - applications to ultra-faint imaging of small, moving, solar system bodies: comets and near-Earth-objects*. Intelligent Processing and Manufacturing of Materials, Vol. 1, 1999. http://dx.doi.org/10.1109/IPMM.1999.792520

[9] G. Jahn, J. Wendt, M. Lotze, F. Papenmeier, M. Huff, *Brain activation during spatial updating and attentive tracking of moving targets*. Brain & Cognition, 78, 2012, pp. 105-113. http://dx.doi.org/10.1016/j.bandc.2011.12.001

[10] J. Ericson, J. Christensen, *Reallocating attention during multiple object tracking*. Attention, Perception & Psychophysics, 74, 2012, pp. 831-840. http://dx.doi.org/10.3758/s13414-012-0294-z

[11] K.A. Patwardhan, G. Sapiro, V. Morellas, *Robust foreground detection in video using pixel layers*, IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 30, (4) 2008, pp.746-751. http://dx.doi.org/10.1109/TPAMI.2007.70843

[12] Y. Wang, J.F. Doherty, R.E. Van Dyck, *Moving object tracking in video*. In proceedings of 29th Applied Imagery Pattern Recognition Workshop, 2000, pp. 95-101. http://dx.doi.org/10.1109/AIPRW.2000.953609

[13] D. Hurych, K. Zimmermann, T. Svoboda, *Fast Learnable Object Tracking and Detection in High-resolution Omnidirectional Images*. VISAPP, 2011, pp.521-530.

[14] Hironori Sumitomo, *Monitoring camera system, monitoring camera control device and monitoring program recorded in recording medium*. US 20030185419 A1, 2003.

[15] Z. Yang, Y. Zhu, and Y. Pu., *Parallel Image Processing Based on CUDA*. International Conference on Computer Science and Software Engineering 2008, Vol. 3, pp. 198–201. http://dx.doi.org/10.1109/CSSE.2008.1448

[16] Aisaka, *et al.*, *Image processing apparatus and method, and program*. United States Patent 8,577,137, November 5, 2013.

[17] D. M. Bramich, Keith Horne, M. D. Albrow, *et al.*, *Difference image analysis: extension to a spatially varying photometric scale factor and other considerations*. Monthly Notices of the Royal Astronomical Society, Volume 428, Issue 3, 2013, p.2275-2289. http://dx.doi.org/10.1093/mnras/sts184

[18] Frank Y. Shih, *Image Processing and Mathematical Morphology: Fundamentals and Applications*, CRC Press, 2009. http://dx.doi.org/10.1201/9781420089448

[19] Frank Y. Shih, *Image Processing and Pattern Recognition: Fundamentals and Techniques*, IEEE Press, 2010. http://dx.doi.org/10.1002/9780470590416

[20] P. Cavanagh, G. A. Alvarez, *Tracking multiple targets with multifocal attention*. Trends in Cognitive Sciences, 9, 2005, pp. 349-354. http://dx.doi.org/10.1016/j.tics.2005.05.009

[21] G. d'Avossa, G. Shulman, A. Snyder, M. Corbetta, *Attentional selection of moving objects by a serial process*. Vision Research, 46, 2006, pp. 3403-3412. http://dx.doi.org/10.1016/j.visres.2006.04.018

[22] W. Żorski, *Application of the Hough Technique for Irregular Pattern Recognition to a Robot Monitoring System*. Proceedings of the 11th IEEE International Conference MMAR 2005, pp.725-730.

[23] W. Żorski, K. Murawski, *Irregular patterns learning and matching in an example vision system*. Proceedings of the 18th IEEE International Conference MMAR 2013, pp.645-649.

[24] NVIDIA corporation, *CUDA C Programming Guide*, July 2013, PG-02829-001_v5.5: http://docs.nvidia.com/cuda/pdf/CUDA_C_Programming_Guide.pdf

[25] NVIDIA corporation, *CUDA C Best Practices Guide*, July 2013, DG-05603-001_v5.5: http://docs.nvidia.com/cuda/pdf/CUDA_C_Best_Practices_Guide.pdf

[26] NVIDIA corporation, NVIDIA Performance Primitives (NPP), Version 4.0, 2014: http://docs.nvidia.com/cuda/pdf/NPP_Library.pdf

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 719–728

DOI: 10.15439/2014F137
ACSIS, Vol. 2

Masking the Effects of Delays in Human-to-Human Remote Interaction

Fei Su, John Markus Bjørndalen, Phuong Hoai Ha, Otto J. Anshus
Department of Computer Science
UiT The Arctic University of Norway
fei.su@uit.no, jmb@cs.uit.no, phuong@cs.uit.no, otto@cs.uit.no

***Abstract*—Humans can interact remotely with each other through computers. Systems supporting this include teleconferencing, games and virtual environments. There are delays from when a human does an action until it is reflected remotely. When delays are too large, they will result in inconsistencies in what the state of the interaction is as seen by each participant. The delays can be reduced, but they cannot be removed. When delays become too large the effects they create on the human-to-human remote interaction can be partially masked to achieve an illusion of insignificant delays. The MultiStage system is a human-to-human interaction system meant to be used by actors at remote stages creating a common virtual stage. Each actor is remotely represented by a remote presence created based on a stream of data continuously recorded about the actor and being sent to all stages. We in particular report on the subsystem of MultiStage masking the effects of delays. The most advanced masking approach is done by having each stage continuously look for late data, and when masking is determined to be needed, the system switches from using a live stream to a pre-recorded video of an actor. The system can also use a computable model of an actor creating a remote presence substituting for the live stream. The present prototype uses a simple human skeleton model.**

***Index Terms*—Effects of Latency; Mask the effects of delays; Temporal Casual Synchrony; Remote Interaction.**

I. Introduction

IN DISTRIBUTED acting, actors at different stages, physically separated by distance, interact to create a coherent play. The interaction can be lazy, allowing for large delays without breaking the illusion of being at the same stage. This is, for example, the situation when actors do a relaxed handshake, or don't interact directly at all. The interaction can also be eager, where even small delays break the illusion. This is, for example, the case when actors do fast action/reaction with causally related movements between each other, or move in synchrony as done in dancing.

Fig. 1 depicts distributed acting. Three stages, in Tromsø, Porto, and Florence, have a total of four actors doing eager interaction, dancing together. In Tromsø, there are two actors physically present, while there is one actor in Porto and one in Florence. At each stage, each actor is represented by a remote presence in the form of an independent streaming video.

Distributed acting is complicated by each stage having a different clock, and by communication delays and jitter. The clock at each stage can easily be sufficiently synchronized with a reference clock, but delays and jitter are unavoidable and are the result of the finite speed of light, and of the technologies and systems applied to create a distributed stage gluing together the individual stages.

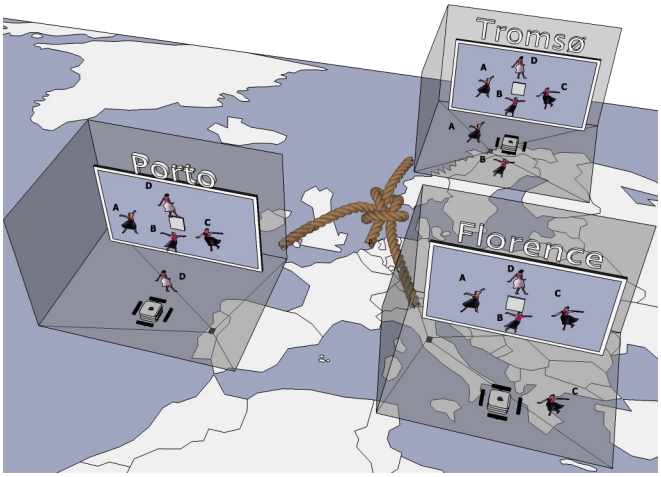


Fig. 1. Four dancers at different stages dance together. Each stage is equipped with sensors to detect actors and a display to visualize the remote presence of all the performers.

The speed of light defines the lower bound of a non-zero delay from an event happens until it can be observed. Table I shows the time needed for light to travel distances that may be typical in distributed acting. It takes about 3 microseconds between buildings, 30 milliseconds between cities and about 134 milliseconds around Earth's equator. The time it takes for light to travel from an actor to another and back is twice this amount of time. However, the actual delays experienced by actors interacting through a computer-based system are even higher.

TABLE I
Travel time at the speed of light

1 km	3.3 μs	Between buildings
1000 km	33 ms	Between cities
4000 km	134 ms	Around equator
2.4 x 10^{19} km	2.5M years	To Andromeda Galaxy

Figure 2 describes the total delay when observing a remote event. Delays are created by the sensors tracking actors, transfer of data from sensors to computers, processing of the sensor input, network transmission, on-route processing,

978-83-60810-58-3/$25.00

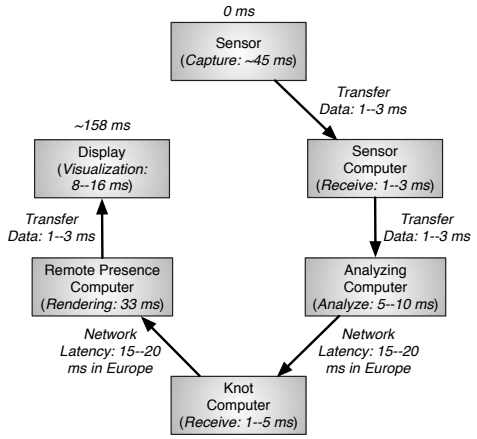


Fig. 2. Every Phase will add delay

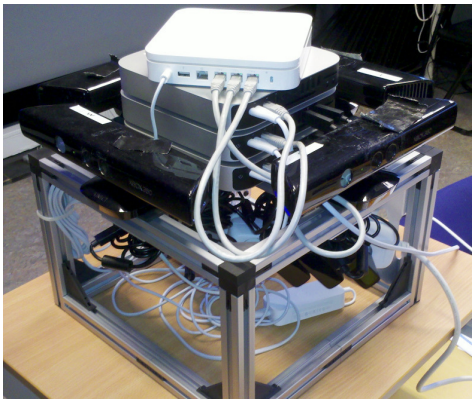


Fig. 3. The 360 degrees actor detection sensor rig comprising four 3D Kinect cameras, two Mac Mini computers, and wireless access point. One per stage is used.

receiving and processing the received data, and preparing and visualizing the data locally. The delays can be significantly larger than what is indicated in the figure if more processing is applied. These delays can be reduced and partially masked, but they can never be removed.

Delays are important when people interact. It has been documented [1], [2], [3], [4] that people accept delays below 200ms as insignificant when interacting tightly. When the delays grow beyond 200ms they become harder and harder to ignore, and actors can be expected to have problems interacting as if they were on the same physical stage.

The goal of the MultiStage system [5] is to aid actors at stages around the world in interacting with each other as if they were on the same stage. Each stage has a set of sensors, shown in figure 3, detecting and tracking the movements of the actors on the stage. The actors at the other stages are each represented by a remote presence. A remote presence is based upon having data about an actor available such that the actor's movements can be recreated remotely. A simple case is to have data representing a streaming video of the actor, and show it on a large display to visualize the actor in full scale. A more advanced case is when an actor's movements are used as input into a computation creating a remote presence of the actor. The remote presence can be visualized on a display or control a robot.

Several experiments were conducted to determine the objective and subjective performance of the system. Objective metrics include the delays in different parts of the prototype system, and processing and network resource usage. Subjective metrics include how much delay an actor will notice and tolerate when interacting, and when an actor experience that the switching of the masking in and out is smooth.

II. Masking Approaches

In [5], we define **loose temporal causal synchrony** to be when actions by actors happen causally in the correct order, but with no special demands on delays. **Interactive temporal causal synchrony** is when actions by an actor is seen in causal order and with delays as actors are used to when being on the same stage face to face. To achieve this even with delays and jitter being unavoidable, the idea is to mask the effects of delays as seen by the actors.

In the **Act-By-Actor**-approach, the actors react to the remote presences as if they were the actual actors. How the interaction looks and how it feels to actors and audiences depends on how large the delays are, by how much they vary, and by how good the actors are at compensating.

In the **Act-By-Director**-approach, a director keeps time and tells actors when to do actions according to a shared script or to a script for each actor. Even if the actors act on command it will seem to an audience as if they interact freely with each other.

A variant is to select a stage to be the **live** stage. The others are secondary stages. The start time for a performance at a secondary stage is the start time for the live stage minus the delay between them. Consequently, performances at secondary stages are started a little earlier than at the live stage such that when the live stage starts, the input from the secondary stages arrive. At the live stage the actors and an audience will experience a performance where local actors are in synchrony with the remote presences representing the remote actors. However, actors at a secondary stage will be out of sync with the remote presences. By switching which stage is the live stage at suitable points in the performance, each stage can be the live stage for a time.

A second variant of this approach is to **delay each local remote presence** at a stage. A local remote presence is the remote presence of an actor shown and heard at the stage where the physical actor also is. The effect is that an actor and an audience will experience a local and a remote event at the same time because they have both been delayed equally much. To make this approach practical, the delay cannot be so high as to make the actors and audience noticing it too much. Because delays between stages in practice tend to be different, this approach is most practical for just two stages with about equal delay between them.

A third variant is to **delay all remote presences** at a stage until data for the slowest remote presence arrives. With varying delays between the stages, they will soon be out of synchronization with each other. However, the local and remote presences at a stage will be in synchronization with each other. The delay waiting for the slowest can be long enough to be noticeable for actors and an audience. Consequently, the actors at a stage can be out of synchronization with the remote presences.

In the **Act-By-Wire**-approach, remote presences are manipulated to hide the effects of delays when delays reach predefined threshold values. Manipulations include just-in-time blending in of prerecorded videos of remote presences of actors, and just-in-time blending in of on-demand computed remote presences. A prerecorded and an on-demand computed remote presence will to a varying degree succeed in creating the illusion of low insignificant delays. If there is a script of what an actor should do at a given time, then a prerecorded remote presence can be created and played back at the correct time when delays go too high. When instead of using a static pre-recorded video a computation is run to create the remote presence, a wide range of possibilities are in principle available. These include blurring the movements of an actor such that delays are not so obvious, and predicting what an actor was going to do. We have not explored these possibilities yet.

III. Related Literature

Several systems try to enable interaction between local and remote users. The Distributed Immersive Performance (DIP) [6] and [7], is a multi-site interaction and collaboration system for interactive musical performances. In experiments, they artificially delayed the local stage and found out that (i) the tolerable latency for slow paced music is much higher than for fast paced music; (ii) to help performers pick up aural cues it is better having a low audio latency than synchronizing video and audio; and (iii) a roundtrip video delay of more than 230ms makes synchronization hard for the users. In [8], a series of experiments on the DIP system is described with focus on the audio delay, and how the delay affects musician's cooperation. An artificial delay of 50ms to the remote room's audio stream was tolerable. With the same latency added at both rooms it became possible to play easily together with a delay of up to 65ms. While they report on the effects of delays on audios, we report on the effects of delays on videos, and how they can be masked.

Other distributed collaboration systems include [9], [10], and [11]. These do not consider the effects of delays and how to mask them when users interact across distance.

Several techniques [12], [13], [14], [15] and [16], exist to reduce or hide network latency in network games and in distributed systems. The Dead-Reckoning (DR) technique is used in distributed simulations and to hide latency mostly in network games. Computers that own an entity will send unique information about the entity to other computers on the network. The information includes the position, velocity, and acceleration of the entity or more. Each computer simulates the movement of the entity. The computer which owns the entity will also simulate the entity as well as check the real state of the entity. When the simulated value and real value differs more than a threshold, the computer will send update information to the other computers. The dead-reckoning technique is a general way to decrease the amount of messages communicated among the participants.

IDMaps [17] measures the distance information on the Internet. This is used to predict latencies. King [18] uses recursive DNS queries to predict latency between arbitrary end hosts. In [19] a structural approach to latency prediction technique based on Internet's routing topology is proposed. In [20] the network latency is reduced based on estimates of the network path quality between end points. These approaches can be useful even if we don't mask latencies themselves, but the effects of delays. Predicting the very near future latency can be useful because we can start the masking right before large delays happen. The Local-Lag (LL) technique [21], provides for better fairness between local and remote players by making all see approximately the same delays. A local operation is delayed for a short time. During this short time period the operation is transmitted to remote computers participating in the game, and all computers can then execute the operation closer in time to each other. However, with more than two participants seeing significantly different latencies, the fairness cannot be maintained for all computers. In [22] and [23], the LL is integrated with DR to synchronize participants and keep better consistency among all computers.

In [14] and [22], some of the drawbacks of the above mentioned DR and LL techniques are identified. While the LL technique ensures fairness for two players, or for multiple with the same latencies between them, the fairness is not preserved when the latencies become too different. The same is the case for the DR approach because when a computer does an update, the time it takes to have data about this delivered at the other computers will vary depending on the latencies between the local computer and each of the other computers. This can result in a situation where a local player and some of the remote players can do actions earlier than other remote players.

Even if it is worthwhile to reduce network latencies and other delays, and do overlapping between communications and processing, delays cannot be removed. In this paper, we present several techniques to mask the effects of delays, and we also measure the cost of applying each technique.

There are several projects which have studied the effect of latency when remote users interact, including [24], [25], [2], [26], [4], [3], and [27]. When the latency from a user does an action until it is reflected in, say, a game, is more than 200ms, the user will notice the delay and his actions and scores are impacted by it. In a first person shooter game there is a 35% drop in shooting accuracy at 100ms of latency, and the accuracy drops sharply when the latency increases further. More than 200ms of latency should be avoided. For some sports and role-playing games a latency of 500ms can

be acceptable. Consequently, latency reduction and hiding techniques should aim at achieving end-to-end latencies less or equal to these numbers. When this cannot be achieved, then masking the effects of the various delays becomes interesting to apply as well.

In [28], a comparison is made between the end-to-end latency of an immersive virtual environment and a video conferencing system. The tolerable latency for verbal communication was found to be 150 ms. This was achieved by the teleconferencing system, but not the virtual environment system. A video was done capturing a person repeatedly moving an arm up and down. A video was also done of the same person as represented by the system. Synchronized cameras were used to be able to synchronize the two videos. The latency from the person moved an arm until it was reflected through the system was measured to be 100-120ms for the teleconferencing system, and 220-260ms for the virtual environment when the avatar for the user had been preloaded.

In [29] several techniques were used to reduce the latency for the head tracking system of an immersive simulation system. The techniques included disabling buffering and having a more direct path to the tracker hardware. This results in an almost 50% reduction in latency, from around 90ms to around 50ms.

Packet jitter [30] is the variation in the packet delay. Variations in packet size, buffer delay, and routing create packet jitter. The influence of the jitter in games is measured in [26], [31], [32], and [33]. They conclude that jitter had only a minor impact on the win probability, the scores and the user experience. However, when jitter increases, the tracking accuracy of a target, the users ability to keep a small and consistent distance between the center of the target and the cursor, declines.

In [34] they consider unfairness created by the cumulated errors between players. The system improves fairness by equalizing for all players, the errors of where an object of the game is placed and what it is doing. This resulted in a significant improvement in consistency between what players observed even for 100ms of delay between players at different computers.

IV. System Overview

Figure 4 shows the MultiStage system. The design and implementation is described in detail in [5].

The system is divided into a local and a global side. The local side of MultiStage primarily focuses on what is happening locally on a single stage. The global side is the glue binding stages together, taking care of distribution of data between stages, and doing analytics needing data from multiple stages.

The local stage monitoring (LSM) system detects local state at the stage, including actors and their movements, and streams it to the local stage analysis (LSA) system. The LSA analyzes the data to detect gestures (not expanded on in this paper), and forwards the data and data about detected gestures to the global side.

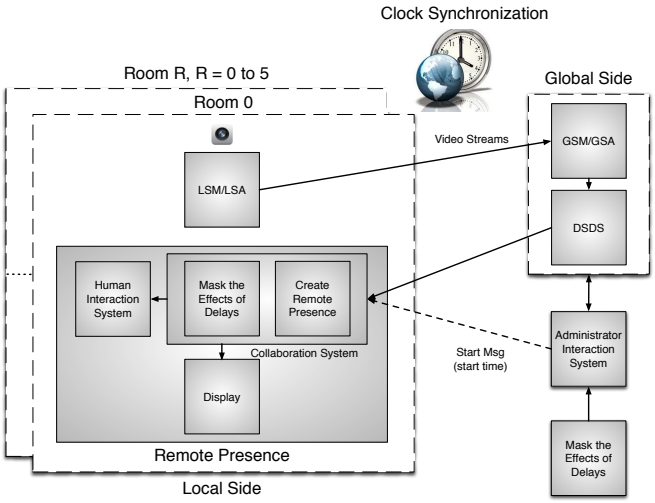


Fig. 4. The MultiStage system.

The LSM system produces an individual stream for each actor. This allows for great flexibility in treating each actor individually when looking for gestures, and where each actor's remote presence is manifested and located in relation to the others on the stages. In the present prototype, an individual stream for each actor is achieved by using a Kinect camera per actor. The system assumes that just a single actor is within the 3D field of view of the camera. All objects outside of this 3D space are ignored. The advantage of having a one-to-one relationship between actors and cameras is that it takes very little processing to create individual streams for the actors. This helps in reducing the delay from an actor moves until it is manifested in the remote presence at remote stages. The disadvantage is that when the number of actors increases, so must the number of cameras. Presently the prototype supports four actors per stage using four Kinect cameras arranged back to back. The back to back configuration avoids having the infrared dot cloud used by the cameras (to achieve depth information) to interfere with each other. While more Kinect cameras can be used, care must be taken to avoid interference. A more advanced sensor suite will help avoid this problem.

The remote presence system at each stage subscribes to streams from the global side. The data is used to locally create remote presences. In the prototype, remote presences are visualized on a big display. The visualization of a remote presence can take three forms. It can be 2D streaming videos based on color images captured by four Kinect cameras at each stage. Alternatively, 3D point streaming cloud videos can be used. These are created using color and depth images captured by the Kinect cameras. Finally, a remote presence can be visualized as an animated human skeleton created locally at each stage.

The LSM uses the Kinect cameras to sense actor movements. In principle, if the LSA identifies actor body movements, the data about this makes its way to the remote pres-

ence, and the computed human skeleton moves accordingly. In the present prototype, a script defining what each actor should do is used. When delays become to high, the human skeleton remote presence computation for an actor receives commands taken from the script. These commands are typically of the type "raise left arm" and "lower right arm". Computing a model of a human skeleton locally, and letting it react to just streaming movement commands, saves network bandwidth vs. distributing streaming videos.

The remote presence system includes the masking system. It looks for incoming data about remote actors being too delayed to do remote presences without the local actors noticing the delay. If the delays are too large, the masking system applies several techniques to mask the effects of the delays as seen by the actors. A limited form of masking is also done outside of the remote presence system. In this case the masking system provides information to the administrator interaction system (see below) such that it can tell the human interaction system at each stage what to do, like individual delayed start-up times of a performance for each stage.

The human interaction system at each stage informs actors what to do, and provides them a countdown for when they should start doing it. In the prototype, a display per stage is used to visualize this for the actors.

The global side monitoring (GSM) receives data streams from the local stages. It forwards the data to the global side analysis (GSA) system. The GSA system does analytics on the data streaming in from the stages looking for global state. An interesting global state is a collective gesture. It is comprised of several gestures done by several actors possibly at different stages. The idea is that when a given number of actors have done a certain gestures, this should result in actions taken at the stages, like, say, turning on a light or doing some modifications to the remote presences.

The GSA system forwards all data and information about global gestures to the distributed state distribution system (DSDS). The DSDS manages subscriptions from the remote presence system at each stage, and deliver streams to the subscribers.

The administrator interaction system lets a director manage the system, including setting and distributing to all stages the start time of a performance. Each computer in the system has a performance monitor measuring several metrics including latency between the computers and bandwidth. These measurements are made available to the administrator interaction system.

The sub-systems implementing the local side executes on computers local to a stage. This is done to achieve low local latencies, and reduce network bandwidth. It also distributes the global workload, and isolates the stages such that if one stage fails, the other stages have a higher probability of not being affected. The sub-systems implementing the global side executes on computers that are located relative to the stages to achieve high bandwidth and low latencies. The administrator interaction system is located on a computer which is convenient to use by an director.

Multistage is a distributed system, and the computers can have different clock values. The system assumes that all computers have the same time, and the clocks are therefore synchronized.

Experiments measuring the performance of the prototype have been done both with all stages locally on the same local area network, as well as kept more than 1500 km apart (Tromsø to Oslo and back). The system currently scales across the Internet with good performance to three stages, and comprises in total 15 computers, 12 cameras, and at least 12 outgoing and 36 incoming data streams.

The system was primarily implemented in Python. The OpenKinect Libfreenect library is used to fetch RGB and depth images from the cameras. The LSA motion detection using Python OpenCV is taken from Robin David on GitHub [35]. The human skeleton model is implemented in Python, using Pygame. Pygame is used to display the actor script. Python Tkinter module is used to display the Administrator Interaction System. The system runs on Linux and Mac OS X.

V. Design and Implementation of Masking the Effects of Delays

To do masking, several functionalities must be realized at each stage. A **shared clock** is assumed by the system. This is achieved with sufficient accuracy by using Network Time Protocol (NTP) [36] to set the local clocks. A **performance monitor** measures and computes the communication delays between all computers. To do so, every packet sent is time stamped. It also measures the clock differences between the computers at a stage and the DSDS distribution computer to determine if clock synchronization is needed to maintain the shared clock. The performance monitor is present at every computer of the system.

A **shared and individual performance start-times** are distributed by using the administrator interaction system to send a message with the performance start time to each stage. We assume that when needed there are predefined **actor scripts** available telling each actor what and when to do an action. In the prototype a display at each stage shows a count down until the next action is to be done, and visualizes with a simple drawing what the action is.

The following masking approaches are shown in figure 5. For all approaches we assume that the stages have already initiated subscriptions to data streams from each other, and that the streaming is in effect.

Live Stage: The administrator interaction system uses the performance monitor to measure the latency from the detection computer at each secondary stage to the distribution server. It also measures the latency from the distribution server to the remote presence computer at the live stage. The effective latency from a secondary stage to the live stage is the sum of these two latencies. A secondary stage's performance start time is the start time at the live stage minus the latency between the live and the secondary stage.

The administrator interaction system now sends a message to each stage with the start time of the performance and

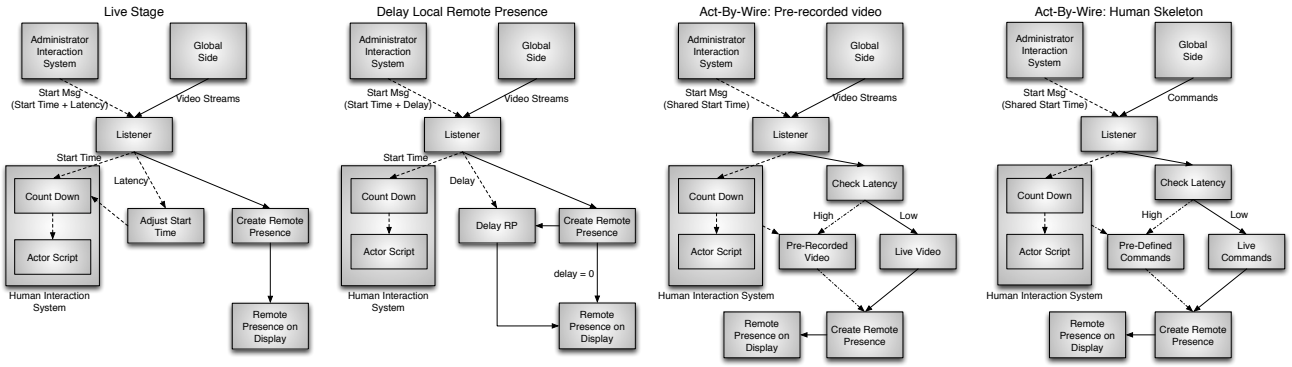


Fig. 5. Design and Implementation of the techniques to mask the effects of delays

the latency that should be decreased to the start time for that particular stage. The human interaction system at each secondary stage will now do a countdown with the start time of the live stage modified by the latency to the live stage. When the countdown ends, a visualization of what each actor should do is displayed. The human interaction system now acts as a director, counting-down to the next action of each actor, and then visualizing the action.

Delay Local Remote Presences: The administrator interaction system uses the performance monitor to measure the delay from the detection computer at each stage to the distribution server. It also measures the delay from the distribution server to the remote presence computer at the stage. If the delays are close an average delay is computed, and this approach to masking can be applied. The administrator interaction system sends a message to each stage with the start time of the performance and the average delay between the stages. The human interaction system starts a countdown at the given start-time. At a stage, each remote presence representing a local actor at the stage is locally delayed by the average delay. The remote presences from other stages are not delayed by the receiving stages.

Delay Locally the remote presences until data for the most delayed remote presence arrives: As for the Live Stage masking approach, the administrator interaction system uses the performance monitor to measure the delay from the detection computer at each stage to the distribution server. It also measures the delay from the distribution server to the remote presence computer at the stage. The effective delay from detection side of a stage to the display side of a stage is the sum of these two delays.

The administrative interaction system sends a message to each stage with the same start time, and the delay from every stage to the stage receiving the message. Each stage calculates by how much remote presences from each stage should be delayed to play back close in time to the remote presences coming from the stage with the longest delay. The human interaction system starts a countdown, and tells the actors what to do and when to do it. The create remote presence system creates remote presences as fast as it can, but remote presences from each stage are individually delayed by the calculated amount for each stage.

Act-By-Wire, blend in prerecorded video or compute a remote presence: The administrator interaction system sends the same start-time to the human interaction system at each stage. It starts a countdown and tells the actors what to do and when to do an action. For every image (or video frame) arriving to be used to create a remote presence, we check if the delta between the send timestamp of the image and the receive time is large enough to warrant masking. If more than a certain percentage of images are late, we start masking. If the percent goes down, we stop the masking. The threshold values used are based on subjectively trying the system on humans with different delay values, and determining when humans notice the delays in several settings, see later for more. We typically use a delay of about 280ms as the threshold for starting to do masking.

To mask short-term delays, the system check for delays over the last few seconds. The exact number of seconds used is tunable, depending upon how sensitive humans in a particular setting are to delayed remote presences.

The video used to mask the effects of delays is pre-recorded. The human interaction system does a countdown, and tells an actor what to do and when to do it, and a video is recorded. When later the same script is used during a performance, and the delays go above the threshold, the pre-recorded video blends in and takes over for the streaming video coming from a remote stage.

The masking system keeps ready the pre-recorded video in memory, and when masking is determined to be needed after checking the latency, it streams the pre-recorded video to the create the remote presence instead of the live streaming video.

Alternatively, instead of using a pre-recorded video, a model of an actor can be used. Instead of streaming a pre-recorded video to create a remote presence, the masking system streams the output from an implementation of the model. The model can receive input about detected body movements from the LSA (through the distribution server) of the remote stage. It can also use the script from the human interaction system to determine what an actor is meant to do. Presently, just a simple

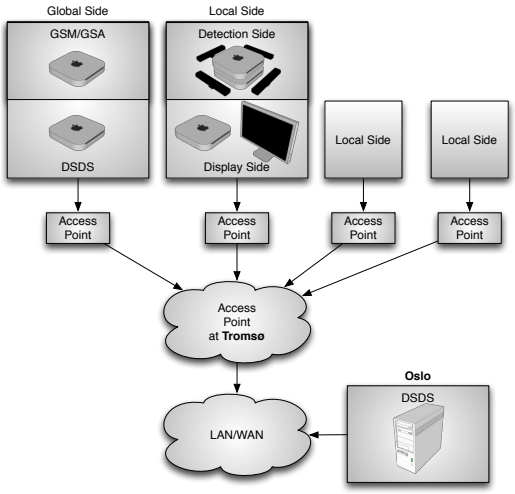


Fig. 6. The configuration of the experiments.

human skeleton model is used with arm movements taken from a script defining what an actor should do. It is future work to explore models and predicting actor behavior more fully.

VI. Evaluation

Several experiments were conducted to identify some of the effects of latency on the actors, and to document the measurable performance of the masking system. For the experiments the system was configured as given in figure 6. Computers used were Mac Minis at 2.7GHz and with 8GB 1333 MHz DDR3 memory. For all experiments all local side stages were on the same 1Gbit/s switched Ethernet LAN inside the Department of Computer Science at the University of Tromsø. The global side DSDS computer was either on the same LAN as the stages, or located on a Planetlab [37] computer at the University of Oslo, 1500km away.

System end-to-end one-way latency: The time it takes for a physical event happening on a stage to be picked up by the cameras and until a visualization of the actor is actually displayed on the same stage. We used a video camera with a high frame rate to record several videos of a user and the remote presence done on a display behind the user. We then counted frames to see how many frames it took from the user moved to the visualization caught up. On a LAN the end-to-end latency was between 90-125ms. With the DSDS at the computer in Oslo, the end-to-end latency was between 100-158ms. The variation in measured latency is because of several factors, including the distributed architecture of the prototype and the frame rate of the projector, video camera (240 fps) and the Kinects (30 fps), and other traffic on the LANs and WAN.

Global-to-Local round-trip latency: The latency going from the DSDS computer to a stage computer and back. We measured this by recording the time when we send a message from DSDS to a stage, and recording when a reply message comes back to DSDS. When all stages and the global side were on the same LAN, the round-trip latencies were between 1-2ms. When the DSDS system was on a computer in Oslo the round-trip latencies were around 32ms. This matches well with measurements reported by PingER [38] for Europe.

Actor-to-actor round-trip latency: The delay that actors will experience from when they do an action until they see the remote presence of another actor reacting. The typical latency between actors is two times the system end-to-end latency. Using the measured results from the system end-to-end one-way latency, the actor-to-actor round trip latency is from 180 to 316ms depending on where the DSDS computer is located.

Human response latency: The time it takes for a human actor to react to another actor's action. We used a high frame rate camera to record two actors' actions, and counted frames from when one actor initiated an action until the other actor responded to the action. The actions used were rapid and slow moving arm movements. The human response latency is about 345ms. We did not find that the latency varied significantly with the speed of an action.

Human noticeable latency: This is the latency at which a human actor will notice that an action is delayed. We simultaneously observed an actor and the corresponding remote presence. When the actor moves an arm, the remote presence moves an arm. In software we artificially added a delay to the remote presence until we noticed that the remote presence lagged behind the actor. When the added latency is more than 100ms, we did notice a difference of the movement between the actor and the remote presence.

Human tolerable latency: This is the latency an actor can tolerate before the illusion of being on the same stage with other actors breaks. We observed an actor shaking hands with another actor on the same stage. We then moved one of the actors to a remote stage, and repeated the shaking of hands. We now observed an actor shaking hands with a remote presence of the other actor. The delay between two actors were artificially increased until we subjectively decided that the handshake was not happening as fast as it did when the actors were physically on the same stage. We tried both rapid hand movement and slow hand movement. We subjectively decided that for a rapid hand movement, it is not tolerable when 150-200ms latency was added. The total actor-to-actor round-trip latency is in this case about 350-400ms. For slow hand movement, it is not tolerable when 600ms latency was added. The actor-to-actor roundtrip latency is about 800ms.

For handshake type of interaction, longer delays bordered on creating a feeling that the remote actor was being obnoxious by delaying just a bit too long before responding to a hand shake. However, this was not experienced unless we artificially added delays. This indicates that the prototype is able to maintain the illusion of being on the same stage for handshake type of interactions. However, we observe that the typical actor-to-actor round-trip latency in Europe is around 300ms or more. Consequently, when actors do fast and rapid interaction, the system can expect to have to mask the effects of the delays.

When to start masking: We simultaneously observed an actor moving an arm, and the corresponding remote presence.

In software we artificially added a random delay to every image used to create the remote presence. We tried different combinations of delays and for how many of the images were delayed. We found that when more than 50% of the received images during a period of three seconds were delayed 280ms or more there is a subjectively clearly visible lag in the remote presence vs. the actor. We therefore determine that when 50% of the images arrive 280ms late during the last three seconds, this is the threshold for when to start masking. This is a threshold that can be changed to customize for different usage scenarios.

When to stop masking: When masking is active, we need to establish a threshold for when to stop masking. We artificially create a situation where more than 50% of the images used to create a remote presence arrive too late. Consequently masking is done by the system. For the experiment we used the Act-By-Wire pre-recorded masking approach. We gradually decreased the percentage by 5% from 50% to 30%. We observe the switching back and forth between the live streaming of the remote presence and the pre-recorded stream. When 35-40% of the images arrive late the switch from the pre-recorded to the live streaming results in a transition without the observer noticing obvious effects of the delay. A higher percentage leads to a sooner switch, but the transition can be too fast and resulting in a blending in of the live streaming video with noticeable delays. A lower percentage results in keeping the pre-recorded video playing too long, and this can become noticeable by itself. The goal is to find a balance between when to start masking and when to stop. This can be different for different user activities and needs.

Above, we checked for late images during the last three seconds. A shorter period will lead to less delay in starting masking when needed, and a longer period is slower in starting masking. For shorter periods, a higher threshold for stopping the masking will reduce the likelihood of switching back and forth. For longer periods, a lower threshold for stopping the masking will increase the likelihood of switching back to the live streaming.

Cost of Masking: The CPU utilization at a remote presence computer without and with the masking technique active was measured. Two cameras were used sending images for two remote presences to a single remote presence computer. The CPU utilization without masking was about 22%. When masking was done for both remote presences using two pre-recorded videos the CPU utilization was basically the same, 22%. When masking was done using two human skeletons, the CPU utilization at the remote presence computer went down to 9%.

We explain this by observing that a significant part of the CPU load was consumed to display videos, making the masking itself insignificant. The very simple human skeleton approach is clearly less CPU demanding. We explain this by the simplicity of the model and that they use the display much less than the videos do.

The overhead of checking if masking is needed and to actually get the masking takes effect is about 40ms in average.

Table II shows the maximum system-end-to-end one-way latency at which each masking approach is in principle at least partially successful at masking the effects of delays.

VII. Discussion

Some of the masking techniques we applied need a synchronization of the clocks at every computer in, and consequently at, every stage of the system. The Network Time Protocol (NTP) provides time accuracy in the range of 1-30ms. The exact accuracy is highly dependent on the location of the computers vs. the NTP servers. If computers are on the same local area network, this will bring them close, around 1ms, to each other. If they are separated by the Internet, the clocks can be synchronized within tens of milliseconds to each other. However, network congestion and routing can cause the clock value used by each computer to be off hundreds of milliseconds. Therefore we do frequent NTP based clock settings and check explicitly for the clock difference between the computers to see if the clocks are more than 10ms off. If they are, we repeat using NTP to try to get all clocks within 10ms of each other. To further ensure that clocks are close enough, before the performance start time is sent to each stage, we again check the clock difference between the computer distributing data to all stages and the remote presence computers at every stage. The clock difference relevant for a stage is included in the message sent to each stage. A stage can then correct its performance start time accordingly if needed.

The experiments measured the objective metrics. No user studies were performed. The determination of thresholds was done naively based on the opinion of a few persons observing actors and remote presences.

The experiments used simple movements by an actor, primarily hand and arm movements. The results can be expected to be different for other actions done by actors, like body rotation, jumping, and dancing.

Different approaches to masking the effects of delays should be expected and to be needed based on what actors are doing. When actors do slow movements and the delays are low, the Act-By-Actor approach can be sufficient. However, it cannot mask the effects of larger delays. The Act-By-Director approach tells actors what to do and when to do an action. All actors are as such seen by an audience at a stage to be synchronized. This approach can mask the effects of large delays. The live stage approach will make just a single stage look synchronized. The other will typically be out of synchronization with the live stage and each other. The approach delaying the local remote presences by the amount of the delay to remote stages will make all stages synchronized if the artificial added delay is smaller than 65ms for audio and 300-400ms for video.

The approach of letting each stage do local delays of every remote presence waiting for the most delayed will make each stage to be in synchrony, but the stages will not be inter-stage synchronized. The Act-By-Wire approach can synchronize actors and remote presence of actors at all stages. However, it makes use of pre-recorded and creates on-the-fly remote

TABLE II
APPROACHES TO MASKING THE EFFECTS OF DELAYS. THE DELAY VALUES ARE THE MAXIMUM SYSTEM-END-TO-END ONE-WAY LATENCIES FOR WHEN AN APPROACH WILL BE AT LEAST PARTIALLY SUCCESSFUL AT MASKING THE EFFECTS OF DELAYS.

Approaches to masking the effects of delays	**Satisfactory synchrony between all remote presences at every stage**	**Satisfactory synchrony between all actors at every stage**	**Satisfactory synchrony between all actors and all remote presences at every stage**
Act-By-Actor	< 190-325ms	< 190-325ms	< 190-325ms
Act-By-Director	< 390-525ms	Any	< 390-525ms
Live Stage	Any (only at live stage)	< 390-525ms	Any (only at live stage)
Delay Local Remote Presence	Any	Any	< 390-525ms
Delay Locally All Remote Presences Waiting for the Slowest	Any	Any	< 390-525ms
Act-By-Wire (blend in pre-recorded remote presence)	Any	Any	Any
Act-By-Wire (blend in on the fly created remote presence)	Any	Any	Any

presences. These can be quite different from, say, a video of the actual actors.

All the masking approaches were tried in the prototype system. However, they are primarily documented as principles. To evaluate where they fit best in an actual interaction, they should be used, and the results should be studied.

The most advanced masking approach, Act-By-Wire using a model of the human to create the remote presence, can be applied with much more complex models than a human skeleton. This is future research. However, when a computable model of an actor is used, its execution should ideally produce results fast enough to not create further delays. If the model demands too long running time to create the needed output, a simpler model may have to be used. Alternatively, predictive techniques may be needed to have output ready when it is needed. The predictions can be based on pre-written scripts defining what a human is meant to be doing at any given time, or it can be based on analyzing the humans' actions in the near past. Predicting the behavior of an actor in the MultiStage system is future research.

VIII. CONCLUSION

In computer supported human-to-human interaction across distance, delays cannot be avoided. Consequently, while reducing the delays are well worth doing, sometimes they still become too large to ignore for humans. When this is the case, some of the effects of delays can be masked to create an illusion for the humans interacting, and for observers, that they are in the same room or on the same stage. However, the illusion created by masking has several limitations depending on which masking approach is used. There are two principally different types of masking. One type coordinates the interaction at suitable times to create a better illusion. The other frequently monitors the delays, and substitutes delayed data with data already available at each stage. Depending on the type of interaction, a suitable masking approach should be selected. The most complex approach, Act-By-Wire, will in all situations in principle create an illusion where interacting humans are fooled to believe that there are no significant delays perturbing the interaction. However, this approach can also create unexpected representations of remote humans, and when this happens it becomes clear that what is shown is only an approximation of the remote reality. The masking approaches we developed and did performance measurements on, demanded insignificantly more resources than not using them, and can even in the most complicated case when using Act-By-Wire, be switched in and out with insignificant delays.

Based on informal use of the system, we found that even 800ms of delay while interacting using slow movements in some cases were tolerable. However, the general case seems to be that delays above 200ms is noticeable when having remote presences based on vision and visualizations. We found that an actor-to-actor round-trip delay of above 200ms is frequently the case, and masking is consequently frequently needed.

ACKNOWLEDGMENT

Many thanks to the technical staff at the department. This work was funded in part by the Norwegian Research Council, projects 187828, 159936/V30, 155550/420, and Tromsø Research Foundation (Tromsø Forskningsstiftelse).

REFERENCES

[1] A. Pavlovych and W. Stuerzlinger, "Target following performance in the presence of latency, jitter, and signal dropouts," in *Proceedings of Graphics Interface 2011*. Canadian Human-Computer Communications Society, 2011, pp. 33–40.

[2] L. Pantel and L. C. Wolf, "On the impact of delay on real-time multiplayer games," in *Proceedings of the 12th international workshop on Network and operating systems support for digital audio and video*. ACM, 2002. doi: http://dx.doi.org/10.1145/507670.507674 pp. 23–29.

[3] [Online]. Available: http://www.measurepolis.fi/alma/ALMA%20Human%20Reaction%20Times%20as%20a%20Response%20to%20Delays%20in%20Control%20Systems.pdf

[4] X. Jiang, F. Safaei, and P. Boustead, "Latency and scalability: a survey of issues and techniques for supporting networked games," in *Networks, 2005. Jointly held with the 2005 IEEE 7th Malaysia International Conference on Communication*, vol. 1. IEEE, 2005. doi: http://dx.doi.org/10.1109/ICON.2005.1635458 pp. 6–pp.

[5] F. Su, G. Tartari, J. Bjørndalen, P. Ha, and O. Anshus, "Multistage: Acting across distance," in *Information Technologies for Performing Arts, Media Access, and Entertainment*, ser. Lecture Notes in Computer Science, P. Nesi and R. Santucci, Eds., vol. 7990, no. 978-3-642-40049-0. Springer Berlin Heidelberg, 2013. doi: http://dx.doi.org/10.1007/978-3-642-40050-6_20 pp. 227–239.

[6] A. A. Sawchuk, E. Chew, R. Zimmermann, C. Papadopoulos, and C. Kyriakakis, "From remote media immersion to distributed immersive performance," in *Proceedings of the 2003 ACM SIGMM workshop on Experiential telepresence*. ACM, 2003. doi: http://dx.doi.org/10.1145/982484.982506 pp. 110–120.

[7] R. Zimmermann, E. Chew, S. A. Ay, and M. Pawar, "Distributed musical performances: Architecture and stream management," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 4, no. 2, p. 14, 2008. doi: http://dx.doi.org/10.1145/1352012.1352018

[8] E. Chew, C. Kyriakakis, C. Papadopoulos, A. Sawchuk, and R. Zimmermann, "Distributed immersive performance: Enabling technologies for and analyses of remote performance and collaboration."

[9] A. Basu, A. Raij, and K. Johnsen, "Ubiquitous collaborative activity virtual environments," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 2012. doi: http://dx.doi.org/10.1145/2145204.2145302 pp. 647–650.

[10] A. Tang, M. Pahud, K. Inkpen, H. Benko, J. C. Tang, and B. Buxton, "Three's company: understanding communication channels in three-way distributed collaboration," in *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, 2010. doi: http://dx.doi.org/10.1145/1718918.1718969 pp. 271–280.

[11] H. H. Baker, N. Bhatti, D. Tanguay, I. Sobel, D. Gelb, M. E. Goss, W. B. Culbertson, and T. Malzbender, "Understanding performance in coliseum, an immersive videoconferencing system," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 1, no. 2, pp. 190–210, 2005. doi: http://dx.doi.org/10.1145/1062253.1062258

[12] [Online]. Available: http://www.gamasutra.com/view/feature/3230/dead_reckoning_latency_hiding_for_.php

[13] Y. W. Bernier, "Latency compensating methods in client/server in-game protocol design and optimization," in *Game Developers Conference*, vol. 98033, no. 425, 2001.

[14] Z. Li, X. Tang, W. Cai, and S. J. Turner, "Fair and efficient dead reckoning-based update dissemination for distributed virtual environments," in *2012 ACM/IEEE/SCS 26th Workshop on Principles of Advanced and Distributed Simulation (PADS)*. IEEE, 2012. doi: http://dx.doi.org/10.1109/PADS.2012.18 pp. 13–22.

[15] T. K. Capin and I. S. Pandzic, "A dead-reckoning algorithm for virtual human figures," in *Virtual Reality Annual International Symposium, 1997, IEEE 1997*. IEEE, 1997. doi: http://dx.doi.org/10.1109/VRAIS.1997.583066 pp. 161–169.

[16] V. Y. Kharitonov, "Motion-aware adaptive dead reckoning algorithm for collaborative virtual environments," in *Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*. ACM, 2012. doi: http://dx.doi.org/10.1145/2407516.2407577 pp. 255–261.

[17] P. Francis, S. Jamin, C. Jin, Y. Jin, D. Raz, Y. Shavitt, and L. Zhang, "Idmaps: A global internet host distance estimation service," *Networking, IEEE/ACM Transactions on Networking (TON)*, vol. 9, no. 5, pp. 525–540, 2001. doi: http://dx.doi.org/10.1109/90.958323

[18] K. P. Gummadi, S. Saroiu, and S. D. Gribble, "King: Estimating latency between arbitrary internet end hosts," in *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*. ACM, 2002. doi: http://dx.doi.org/10.1145/637201.637203 pp. 5–18.

[19] H. V. Madhyastha, T. Anderson, A. Krishnamurthy, N. Spring, and A. Venkataramani, "A structural approach to latency prediction," in *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*. ACM, 2006. doi: http://dx.doi.org/10.1145/1177080.1177092 pp. 99–104.

[20] Y. Lee, S. Agarwal, C. Butcher, and J. Padhye, "Measurement and estimation of network qos among peer xbox 360 game players," in *Passive and Active Network Measurement, doi: http://dx.doi.org/10.1007/978-3-540-79232-1_5*. Springer, 2008, pp. 41–50.

[21] P. Huang, Y. Ishibashi, N. Fukushima, and S. Sugawara, "Interactivity improvement of group synchronization control in collaborative haptic play with building blocks," in *Proceedings of the 9th Annual Workshop on Network and Systems Support for Games*. IEEE Press, 2010, p. 2.

[22] Y. Zhang, L. Chen, and G. Chen, "Globally synchronized dead-reckoning with local lag for continuous distributed multiplayer games," in *Proceedings of 5th ACM SIGCOMM workshop on Network and system support for games*. ACM, 2006. doi: http://dx.doi.org/10.1145/1230040.1230071 p. 7.

[23] A. Malik Khan, S. Chabridon, and A. Beugnard, "A dynamic approach to consistency management for mobile multiplayer games," in *Proceedings of the 8th international conference on New technologies in distributed systems*. ACM, 2008. doi: http://dx.doi.org/10.1145/1416729.1416783 p. 42.

[24] S. Aggarwal, H. Banavar, A. Khandelwal, S. Mukherjee, and S. Rangarajan, "Accuracy in dead-reckoning based distributed multiplayer games," in *Proceedings of 3rd ACM SIGCOMM workshop on Network and system support for games*. ACM, 2004. doi: http://dx.doi.org/10.1145/1016540.1016559 pp. 161–165.

[25] T. Yasui, Y. Ishibashi, and T. Ikedo, "Influences of network latency and packet loss on consistency in networked racing games," in *Proceedings of 4th ACM SIGCOMM workshop on Network and system support for games*. ACM, 2005. doi: http://dx.doi.org/10.1145/1103599.1103622 pp. 1–8.

[26] M. Bredel and M. Fidler, "A measurement study regarding quality of service and its impact on multiplayer online games," in *Proceedings of the 9th Annual Workshop on Network and Systems Support for Games*. IEEE Press, 2010. doi: http://dx.doi.org/10.1109/NETGAMES.2010.5679537 p. 1.

[27] M. Claypool and K. Claypool, "Latency and player actions in online games," *Communications of the ACM*, vol. 49, no. 11, pp. 40–45, 2006. doi: http://dx.doi.org/10.1145/1167838.1167860

[28] D. Roberts, T. Duckworth, C. Moore, R. Wolff, and J. O'Hare, "Comparing the end to end latency of an immersive collaborative environment and a video conference," in *Proceedings of the 2009 13th IEEE/ACM International Symposium on Distributed Simulation and Real Time Applications*. IEEE Computer Society, 2009. doi: http://dx.doi.org/10.1109/DS-RT.2009.43 pp. 89–94.

[29] G. Papadakis, K. Mania, and E. Koutroulis, "A system to measure, control and minimize end-to-end head tracking latency in immersive simulations," in *Proceedings of the 10th International Conference on Virtual Reality Continuum and Its Applications in Industry*. ACM, 2011. doi: http://dx.doi.org/10.1145/2087756.2087869 pp. 581–584.

[30] [Online]. Available: http://www.serviceassurancedaily.com/2008/06/latency-and-jitter/

[31] A. Pavlovych and C. Gutwin, "Assessing target acquisition and tracking performance for complex moving targets in the presence of latency and jitter," in *Proceedings of the 2012 Graphics Interace Conference*. Canadian Information Processing Society, 2012, pp. 109–116.

[32] M. Dick, O. Wellnitz, and L. Wolf, "Analysis of factors affecting players' performance and perception in multiplayer games," in *Proceedings of 4th ACM SIGCOMM workshop on Network and system support for games*. ACM, 2005. doi: http://dx.doi.org/10.1145/1103599.1103624 pp. 1–7.

[33] G. Armitage and L. Stewart, "Limitations of using real-world, public servers to estimate jitter tolerance of first person shooter games," in *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology*. ACM, 2004. doi: http://dx.doi.org/10.1145/1067343.1067377 pp. 257–262.

[34] S. Aggarwal, H. Banavar, S. Mukherjee, and S. Rangarajan, "Fairness in dead-reckoning based distributed multi-player games," in *Proceedings of 4th ACM SIGCOMM workshop on Network and system support for games*. ACM, 2005. doi: http://dx.doi.org/10.1145/1103599.1103608 pp. 1–10.

[35] R. David, "Motion-detection-opencv." [Online]. Available: https://github.com/RobinDavid/Motion-detection-OpenCV

[36] [Online]. Available: http://www.ntp.org/

[37] [Online]. Available: https://www.planet-lab.eu/

[38] [Online]. Available: http://www-wanmon.slac.stanford.edu/cgi-wrap/pingtable.pl

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 729–734
DOI: 10.15439/2014F181
ACSIS, Vol. 2

Pong Game on FPGA with CRT or LCD Display and Push Button Controls

Roland Szabó, Aurel Gontean
Politehnica University Timişoara,
Faculty of Electronics and
Telecommunications, Timişoara,
România
Email: roland.szabo@etc.upt.ro,
aurel.gontean@upt.ro

***Abstract*—This paper presents the creation of the Pong game on an FPGA using a computer display. The game is optimized to work on both LCD and CRT displays too. The game is implemented from scratch, it has the two pads drawn, the ball is also drawn and the background is painted in white. Everything is implemented in hardware on FPGA, so this way at the end we are able to create the ASIC with the game. The goal is to create the Pong game on a single chip.**

I. Introduction

THE creation of games was always in the mind of the engineers. From the start of engineering scientists always thought about how to create something to entertain themselves, not only the required research work.

The beginning of the electronics gave big opportunities to the engineers to create games. The first games were also blinking lights and LEDs, but the real games could be made after the introduction of the displays.

One of the first games created was the predecessor of the Pong game, which origi-nal name is Tennis For Two, developed on oscilloscope and created by William Higinbotham in 1958.

On oscilloscope there were created two paddles and a ball, which could be hit by the paddles, if the ball was missed, the other player got a point.

We wanted to reproduce the history, but on FPGA and after on a standalone chip.

We wanted to recreate the Pong game from scratch, by drawing the paddles and the ball, panting the background and displaying it on a computer display.

We chose this game, because it one of the big classic games and it needs not so much drawing, but it's really much fun to play.

II. Problem Formulation

Our task was clear and simple; we needed to refresh the history by recreating one of the pioneers of the computer gaming history, the Pong game.

To make it a little more difficult and interesting, we set the task to create the game on FPGA, because in the future we plan to create the stand alone ASIC, the chip with the Pong game.

The hardware what we had was the NEXYS 2 development board (Fig. 1) with Spartan-3E FPGA (Fig. 3). With this we had also other hardware that we could use like a CRT or LCD computer display to be able to have the whole computer game system. A block diagram of the experimental setup can be seen on Fig. 2, we can see that all the Pong game is on the FPGA board which is connected via VGA port to a PC monitor. The game could be controlled with mouse, keyboard or as in our implementation with the 4 push buttons from the FPGA board.

III. Problem Solving

A. Theoretical Background

We needed to study the whole structure of the development board in order to be able to create the game and to be able to build the correct driver for the VGA monitor, even the Spartan-3E pins.

On Fig. 4 we can see the Spartan-3E other pins which can be connected to the switches, to the push buttons, to the LEDs or to the 7 segment display from the development board. In our Pong game we used the 4 push buttons, 2 for one paddle and 2 for the other paddle for moving them up and down.

The VGA port on the NEXYS 2 development board is interesting. On Fig. 5 we can see the structure of the VGA port which can be found on the NEXYS 2 development board.

Nexys2 board uses 10 FPGA lines to create a port with 8-bit color VGA and two standard lines of synchronization (HS - horizontal sync, VS - vertical sync). Color signals use a resistive divider circuit which together with the 75 Ω termination VGA display, create eight levels of signals, red and green lines, and four signal levels on the blue line.

The VGA timing needs to be set correctly. A VGA controller circuit must generate vertical sync signals – VS and horizontal sync signals – HS and a coordinate delivery of video data on a pixel clock. The pixel clock defines the time available to display one pixel of information. VS signal defines the frequency of the refresh of the screen and is common to all the information on the screen when is redrawn. The minimum refresh frequency is a function of

the intensity of the phosphor screen and electronic spot. Basically refresh frequency is in the range of 50-120 Hz. For a screen of 480 lines with 640 pixels per line, using a 25 MHz pixel clock and a refresh of 60 + /- 1Hz, signal timings are shown in Fig. 6. Synchronization time pulse width for intervals of “front” and “back porch” (these intervals are times pre-and post-synchronization, during when information cannot be displayed) are based on observations taken from actual VGA monitors.

A VGA controller circuit (Fig. 7) decodes the output horizontal sync counter, which is driven by the pixel clock, to generate horizontal sync HS times. This counter can be used to locate any pixel on a given line. Similarly, the output vertical sync counter that increments with each HS pulse can be used to generate VS vertical syn-chronization times and this number can be used to locate any given line. These two counters (continuously operating) can be used to address a memory.

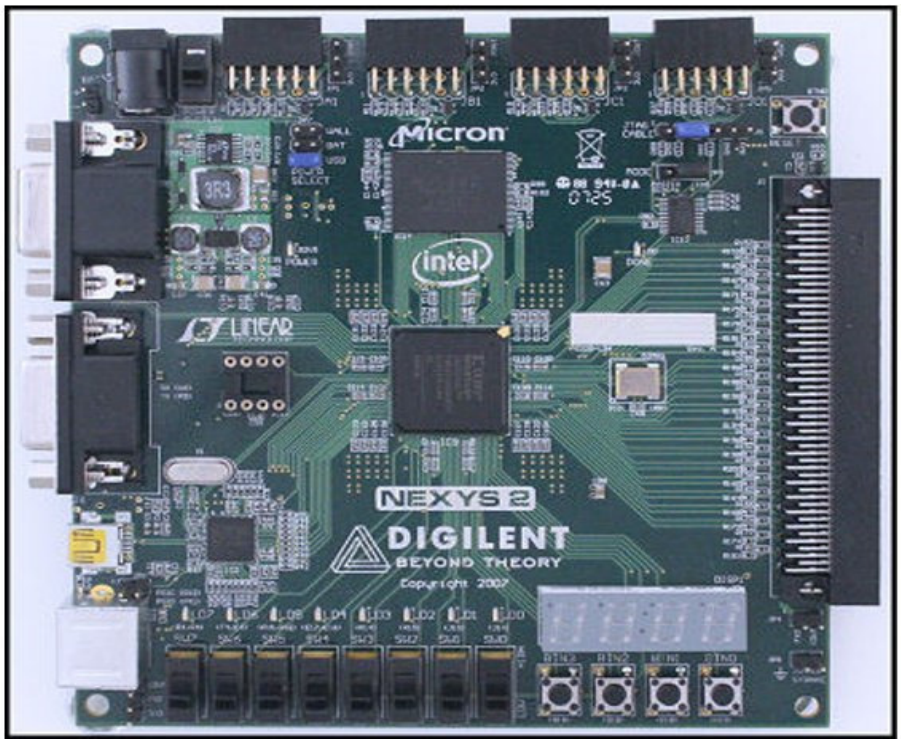


Fig. 1 The NEXYS 2 development board from Digilent with Spartan-3E FPGA

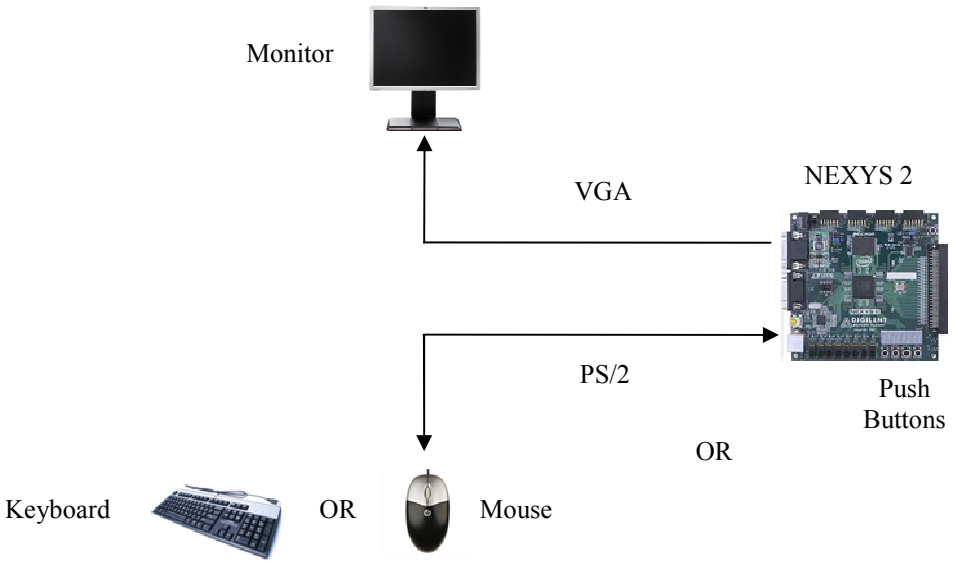


Fig. 2 The block diagram of the experimental setup

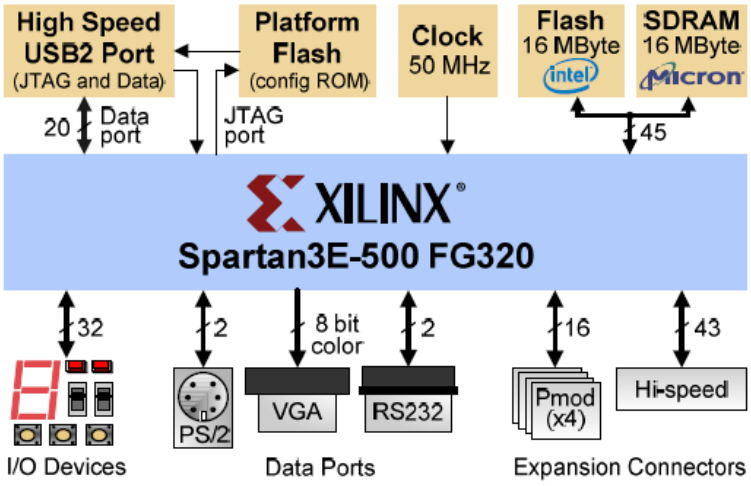


Fig. 3 The Spartan-3E FPGA from the NEXYS 2 development board

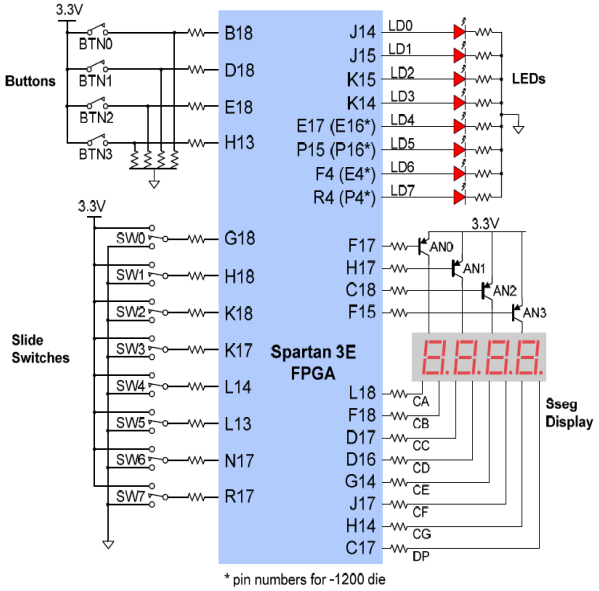


Fig. 4 Spartan-3E pins

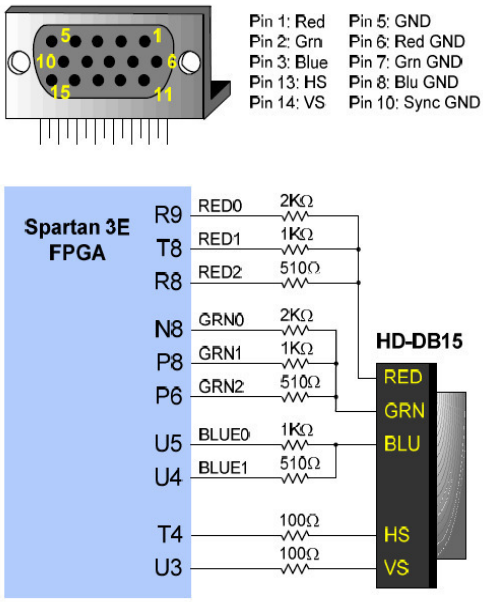
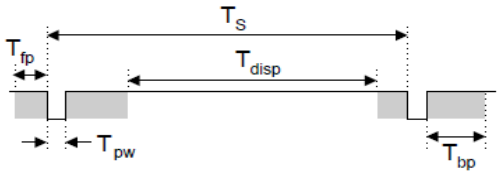


Fig. 5 The VGA port on the NEXYS 2 development board



Symbol	Parameter	Vertical Sync			Horiz. Sync	
		Time	Clocks	Lines	Time	Clks
T_S	Sync pulse	16.7ms	416,800	521	32 us	800
T_{disp}	Display time	15.36ms	384,000	480	25.6 us	640
T_{pw}	Pulse width	64 us	1,600	2	3.84 us	96
T_{fp}	Front porch	320 us	8,000	10	640 ns	16
T_{bp}	Back porch	928 us	23,200	29	1.92 us	48

Fig. 6 The timing for a resolution of 640x480

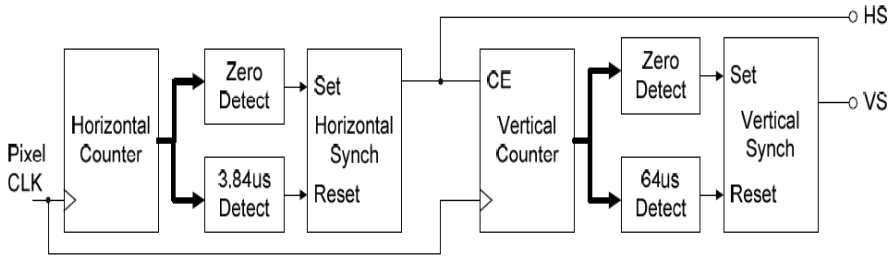


Fig. 7 The circuit for a VGA controller

B. Implementation of the Pong Game

The Circuit of the Game is not very simple. The chip of the Pong game can be seen on Fig. 8.

The inside structure of the Pong game can be seen on Fig. 9. As we can see that the inside structure is quite simple for our circuit. We have only 3 chips inside, one for the VGA sync and one for the graph part for drawing the balls and paddles. We made even the ball to be round by loading the binary values from ROM memory. The last chip is a D-latch for creating a delay for timing.

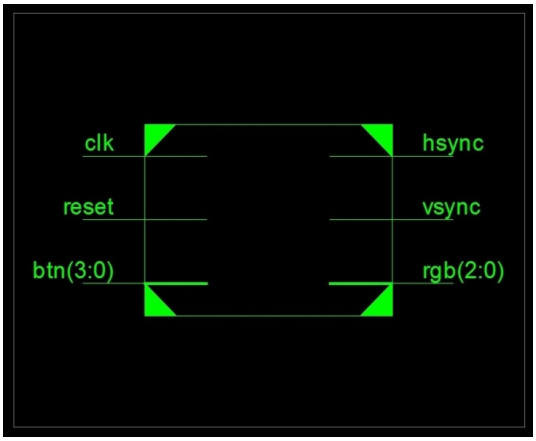


Fig. 8 The Pong game chip

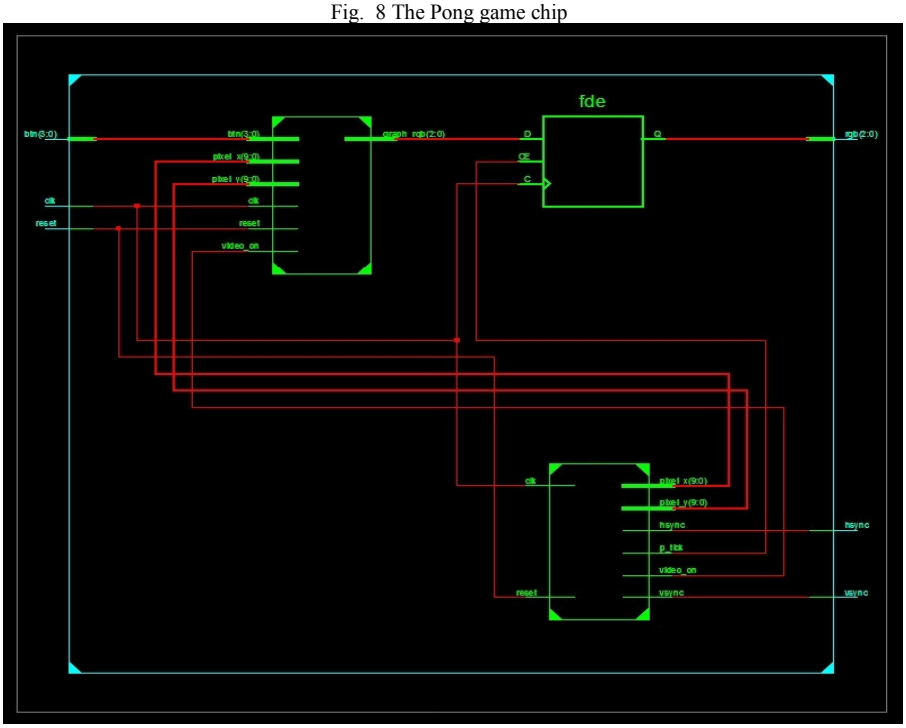


Fig. 9 The inside structure of the Pong game chip

Explanation of the game rules will be presented next. On Fig. 10 we can see the result of the Pong game on a CRT monitor with VGA input. The paddles are controlled with the push buttons from the NEXYS 2 development board.

The game consists of two blades and a ball. Background color of the paddles and the ball is configurable in software. The background is white, the paddles are red and blue, and the ball is green.

Two players can play the game with the 4 push buttons on the board. If the board is oriented in normal position, with the 4 push buttons right below, the first player, with blue paddle, will have the first 2 push buttons (BTN0 movement up, BTN1 movement down) and the second player, with red paddle, will have the last 2 push buttons (BTN2 movement up, BTN3 movement down).

The ball moves automatically hitting the top and bottom edge of the screen and the paddles when the ball it's hit by a player. If a player misses the ball, than he gets a goal, so the ball goes off the screen to the right or left side. For example, if a player misses the ball on the right, then left player gave a goal to the right player. After this it will be a new ball on the left, so the left player serves. The ball will move straight from left to right as long as the right player manages to hit the ball. If the right player scores a goal, then the ball starts its automatic movement from the right. The game is repeated until it it's stopped. The player, who manages to give the most goals, wins.

Fig. 10 The result of the Pong game on a CRT monitor

IV. Conclusion

As we could see we created a Pong game on FPGA. Our goal was to reproduce one of the most entertaining games of the computer game history. The Pong game was one of the first electronic games and maybe made the bases of the computer games history.

It was a good choice to create it on FPGA, because we had a suitable development board which has VGA port for interfacing a computer monitor and push buttons for controls. Making it on FPGA was a good idea also, because this way we had an embedded system, this way we don't have other dependencies not even operating sys-tem dependency. We also chose this platform, because this way we could create an ASIC, a standalone chip, by converting the VHDL code to Verilog with Mentor Graphic tools and we could create the layout of the chip for the final product. This way we had a game implemented on a chip, on hardware; this means that we have a standalone device with no speed issues, like if it would be made in software on a microcontroller.

In future we plan to extend the controls from push buttons to mouse and keyboard on PS/2 interface. We plan also to port the project to other platform too, like the ATLYS board, and create the video drivers on DVI or HDMI interfaces, for newer monitor types and create the keyboard and mouse drivers on USB interface or even add joystick drivers on USB interface.

References

[1] R. Szabó, A. Gontean, "Creating a Serial Driver Chip for Commanding Robotic Arms," *Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2013, pp. 671–674.

[2] R. Szabó, A. Gontean, "Creating a Programming Language for the AL5 Type Robotic Arms," *36th International Conference on Telecommunications and Signal Processing (TSP)*, 2013, pp. 62–65, http://dx.doi.org/10.1109/TSP.2013.6613892.

[3] R. Szabó, A. Gontean, I. Lie, "Smart Commanding of a Robotic Arm with Mouse and a Hexapod with Microcontroller", *18th International Conference on Soft Computing (MENDEL)*, 2012, pp. 476–481.

[4] Ye Xien, Ye Zhiquan, "Implementation of PS/2 Mouse and Keyboard Based on Windows CE," *International Forum on Computer Science-Technology and Applications (IFCSTA)*, vol. 3, 2009, pp. 318–321, http://dx.doi.org/10.1109/IFCSTA.2009.318.

[5] Punj Pokharel, Binod Bhatta, Anand D. Darji, "Optimized drivers for PS/2 and VGA using HDL," *IEEE International Conference on Computer Science and Automation Engineering (CSAE)*, vol. 3, 2011 pp. 262–266, http://dx.doi.org/10.1109/CSAE.2011.5952677.

[6] Guoping Zhang, Man-de Xie, "Design of visual based-FPGA Ping-Pang game with multi-models," *Second Pacific-Asia Conference on Circuits, Communications and System (PACCS)*, vol 2, 2010, pp. 31–34, http://dx.doi.org/10.1109/PACCS.2010.5626899.

[7] Deep Vardhan Bhatt, D. Du Toit, Gerhard P. Hancke, "Design of a controller for a universal input/output port," *IEEE International Instrumentation and Measurement Tech-nology Conference (I2MTC)*, 2012, pp. 647–652, http://dx.doi.org/10.1109/I2MTC.2012.6229492.

[8] Kyu-Sam Sam Lim, Kon-Woo Kwon, Heeju Ju Park, Jeong-Hun Hun Kim, Suki S. Kim, Jun-Jea Sung, Kwang-Hyun Hyun Baek, "Design an infrared wireless optical mouse system and a dual-band infrared receiver," *15th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, 2008, pp. 810–813, http://dx.doi.org/10.1109/ICECS.2008.4674977.

[9] Noriyuki Aibe, Moritoshi Yasunaga, "Reconfigurable I/O interface for mobile equipments," *IEEE International Conference on Field-Programmable Technology*, 2004, pp. 359–362, http://dx.doi.org/10.1109/FPT.2004.1393299.

[10] Liu Kai, Yang Yuliang, Zhu Yanlin, "Tetris game design based on the FPGA," *2nd In-ternational Conference on Consumer Electronics, Communications and Networks (CECNet)*, 2012, pp. 2925–2928, http://dx.doi.org/10.1109/CECNet.2012.6201435.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 735–740
DOI: 10.15439/2014F250
ACSIS, Vol. 2

Image Hashing Secured With Chaotic Sequences

Relu-Laurentiu Tataru
Faculty of Electronics, Telecommunications and Information Technology,
POLITEHNICA University of Bucharest, 1-3, Iuliu Maniu Bvd., Bucharest 6, Romania
Email: tataru.relu@yahoo.com

***Abstract*—This paper presents an image hashing algorithm using robust features from jointed frequency domains. Extracted features are enciphered using a secure chaotic system. The proposed hashing scheme is robust to JPEG compression with low quality factors. This scheme also withstands several image processing attacks such us filtering, noise addition and some geometric transforms. All attacks were conducted using Checkmark benchmark. A detailed analysis was conducted on a set of 3000 color and gray images from three different image databases. The security of the method is assured by the robustness of the chaotic PRNG and the secrecy of the cryptographic key.**

I. Introduction

THE INCREASING popularity of mobile devices (smartphones, tablets and other gadgets) with high quality cameras and the development of high resolution digital cameras raised the number of digital images published over the Internet. An important contribution to this evolution was due to fast development of social networks and services dedicated to media sharing. All these services increased the number of images released in the online environment. This evolution raised the need of addressing some important issues such as: ownership protection, content authentication, fast image indexing and retrieval.

Digital watermarking was proposed as a first solution for these issues. Watermarking techniques are used to embed binary signatures (watermarks) modifying directly the content of the image. Watermarked images are identified detecting the presence of the watermark inside the image. However, many watermarking methods do not solve the problem of content identification. Most of the embedded watermarks are independent of the image content. An alternative of these problems was provided by the concept of image hashing.

Perceptual image hashing was used in the last years to solve ownership disputes, authentication and image retrieval problems. An image hashing scheme usually provides a sequence of values defining the visual characteristics of the image. The result is an image fingerprint usually protected with cryptographic techniques. Compared with conventional hash functions from cryptography, perceptual hash functions designed for images tolerate those modifications which do not affect the content of the image (i.e. compression, filtering, noise addition etc.). Thus, images with different representations, but with the same visual content, provide the same or very close hash values.

In the current work we propose a new algorithm which includes a perceptual hashing scheme for digital images secured with chaotic sequences. The interest was to find a suitable image representation space providing robust features to create an image hash for serving authentication of digital images, copyright protection and easy image database management. Our idea was to combine efficiently existent spaces in frequency domain for feature extraction.

This paper is organized as follows: Section II describes the design principles and properties of image hashing regarding some state of the art principles, Section III presents the construction of the image hashing scheme, Section IV illustrates the simulation results, Section V points some practical applications of the proposed algorithm and Section VI concludes the work.

II. Image Hashing – Design Principles and Properties

A. Design Principles

It is widely accepted that the basic components of perceptual image hashing are image pre-processing, feature extraction, feature post-processing and randomization.

By image pre-processing a new scaled version of the digital image is obtained. This is usually achieved reducing the representation space of the original image without losing the significant features of the content. The result is usually a scaled version of the input image, facilitating operations with low computational cost.

Feature extraction is the next phase in the construction of the image hash. This is an important stage because the feature space influences directly the robustness of the hash function. Depending on the application type of the image hashing scheme, a certain domain for feature extraction could be imposed. Robust algorithms are usually relied on frequency transforms for feature extraction. A scheme resilient to JPEG compression may use Discrete Cosine Transform (*DCT*) for the feature extraction stage. In [3], the authors proposed a scheme based on the statistical modeling

978-83-60810-58-3/$25.00

of DCT coefficients as a Gaussian distribution. The authors assert that invariance of DCT coefficients achieves robustness against attacks such as JPEG compression, filtering, scaling, brightness adjustment, histogram equalization and even small angle rotations. In [6], Fridrich and Gojan illustrate the advantage of considering low frequencies in DCT domain for feature extraction. Their reasoning is based on the properties of low frequency DCT coefficients which preserve the significant information of the image. Any modification in these frequencies is noticeable on the host image. Other transforms are also used in achieving perceptual image hashing. Guo and Dimitros propose the extraction of a robust feature set by using Discrete Wavelet Transform (DWT) followed by Radon transform [9]. The hash value is generated using a probabilistic quantization. This hash value is resilient to image compression, filtering, scaling and rotations. The authors assert good results even for image tampering. When high robustness against geometric attacks is required, the feature set may be extracted using transforms such as Discrete Fourier Transform or Mellin Fourier Transform. Swaminathan et al. [7] propose in their work an image hashing algorithm based on Mellin Fourier Transform. They claim to obtain good results against rotation operations up to 10^o and 20% cropping. This class of methods usually performs well against this type of attacks. However, they may be less robust against other common attacks such as noise addition.

The feature extraction process could be also realized in other transform domains such as Singular Value Decomposition (SVD) [5] or Fast Johnson-Lindenstrauss Transform (FJLT) [10].

The feature set extracted from a transform domain is generally built to assure the goals of the image hashing scheme.

Post-processing stage is usually a compression of the previously extracted features. A feature reduction technique is usually applied in this purpose in order to obtain a final binary feature set. This step is commonly realized using one of the following techniques: random projection of the feature set in another space, direct compression of the feature set, feature set quantization, clustering or by computing a cryptographic hash of the feature set.

Randomization is the last step in achieving the final perceptual hash value of the image. This step is mandatory and assures the unpredictability of the hash value obtained for each digital image, using a secret key.

B. Properties of Image Hashing

The final hash algorithm should provide the following features: one-way i.e. hard to recover the input from the hash value, collision resistant i.e. perceptually different images provide totally different hash values and key-dependence i.e. the hash value is highly dependent on the secret key.

The use of the hash value in verifying an image with a pair is resumed to the direct comparison of the two binary hashes. Few or zero differences between the hash values validate the authenticity of one image with respect to the other one.

The goal of this paper is to propose a robust hash function which respects both design principles and general features of image hashing algorithms. The proposed algorithm is potentially capable of solving copyright disputes, authenticating similar images and retrieving the image content from large image databases.

III. Proposed Image Hashing Scheme

As most of the image hashing algorithms, our scheme computes a global set of features from a digital image. A feature set is used to compute a perceptual hash value. The feature set is enciphered using a chaotic system. The novelty of the proposed algorithm is given by the feature set construction and the use of a proven secure chaotic system for the feature set encryption. A description of the proposed image hashing scheme is illustrated in the following subsections.

A. Image Pre-Processing

A color digital image is converted to grayscale and resized to a default size $m \times m$. The resizing procedure allows fast operations on the grayscale image. Comparing to the original input image, the content of the new image is not changed.

B. Feature Extraction

For the feature extraction step, the grayscale image is converted in frequency domain. This is the most significant stage in computing a robust feature set. A feature set built in frequency domain provides good robustness to certain classes of attacks.

Discrete Wavelet Transform (DWT) and Discrete Cosine Transform (DCT) are jointly used for extracting the global feature set. The first level DWT decomposition assures the separation of the information from the grayscale image in frequency sub-bands LL_1 (low-low), HL_1 (high-low), LH_1 (low-high) and HH_1 (high-high). The LL_1 sub-band carries most of the information from the grayscale image. For this reason, we consider the LL_1 sub-band in the feature generation process. A n –level decomposition is performed, considering LL_{n-1} sub-band ($n \geq 2$) at each iteration. At the n-level decomposition, the LL_n sub-band is obtained. This sub-band provides a matrix preserving most of the correlations from the original grayscale image. The DCT transform is applied for this sub-band on blocks of size $k \times k$. The Wavelet distribution from the LL_n sub-band is changed at the block level, and the new distribution follows the properties of DCT transform. Most significant frequencies are positioned in the top-left corner of the block, and the less significant frequencies are grouped in the bottom-right,

according to the DCT distribution. The first term of each DCT block, i.e. the (0,0) frequency, integrates the most important part of information from the block. This frequency, also called DC term, is extracted from each DCT block.

C. Feature Post-Processing

A feature vector containing the DC's of all DCT blocks computed from the LL_n sub-band is obtained at the previous step. At the current step, we apply a binarization technique for the feature vector. This is achieved by comparing each component of the feature set with the global mean of the feature set. Binary 0 is used to represent DC values under the mean and binary 1 is used to represent DC values above the mean. Thus, we obtain a binary fingerprint of the digital image.

D. Feature Randomization

This step is mandatory in order to assure the confidence of the binary feature set. The security of the feature set is obtained by direct enciphering with a recently proposed chaotic system. The chaotic generator proposed by Vlad et al. [8] is used as a stream cipher. This generator is based on tent-map and the running-key principle. The tent-map has the following formula:

$$x_{n+1} = f(x_n) = \begin{cases} \frac{x_n}{a}, & 0 \le x_n \le a \\ \frac{1-x_n}{1-a}, & a < x_n \le 1 \end{cases} \quad (1)$$

where $a \in (0,1) \setminus \{0.5\}$ is the control parameter of tent-map, x_n is the n^{th} value of the chaotic sequence generated using the tent-map and x_0 is the initial value from $(0,1)$ range. Binary sequences Z_i are generated using the X_i real value sequences generated with the tent-map, and binarization threshold c.

$$z_{i,j} = \begin{cases} 0, & 0 \le x_{i,j} \le c \\ 1, & c < x_{i,j} \le 1 \end{cases}, \quad (2)$$

where $z_{i,j}$ is the j^{th} element of the chaotic binary sequence Z_i and $x_{i,j}$ is the j^{th} element of the chaotic non-binary sequence X_i.

A running-key procedure is applied for typical Z_i binary sequences in order to obtain binary i.i.d. sequences compatible with the fair coin model.

The enciphering key used for the proposed image hashing algorithm is a binary sequence based on five additions of typical sequences Z_i, as shown in equation 3.

$$Y = \sum_{i=0}^{4} Z_i \bmod 2 \quad (3)$$

According to [8], the binarization threshold was considered equal to the control parameter $(c = a)$. The security of the method was theoretically and experimentally proven for a control parameter a in the range $(0.39, 0.61) \setminus \{0.5\}$ for 5 modulo 2 additions.

The secret key K of the system is given by the initial values of each non-binary sequence X_i and the control parameter:

$$K = (x_{00} \| x_{01} \| x_{02} \| x_{03} \| x_{04} \| a) \quad (4)$$

Note: As already suggested in [8], the additions number could be increased, extending the range of the control parameter a. This result leads to a larger selection of the secret key. The construction principle of the chaotic system used to generate the pseudo-random key to encipher the feature set is presented in Fig. 1.

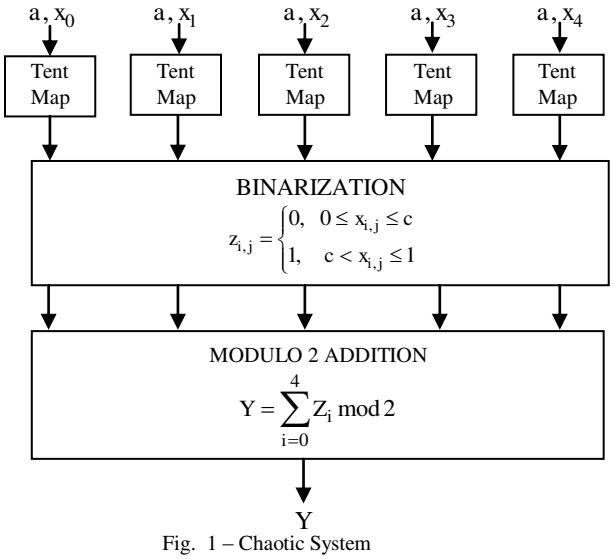


Fig. 1 – Chaotic System

The step-by-step construction of the perceptual image hashing scheme according to the description is next presented:

1. The input color image is transformed to grayscale.
2. The grayscale image is resized to a standard mxm dimension using the bicubic interpolation.
3. A n –level DWT Haar decomposition is applied on the grayscale image.
4. LL_n sub-band is divided in non-overlapping kxk blocks and DCT transform is applied on each block.
5. DC coefficients are extracted from all DCT blocks and the vector V containing the features of the image is built. The

length of the feature vector V is given by the formula:

$$l = \left(\frac{m}{k \cdot 2^n} \right)^2 \quad (5)$$

6. The mean value m_{dc} of the feature vector is computed.

7. The feature vector V is binarized and a new binary feature vector W is obtained according to the formula:

$$w_i = \begin{cases} 0, & v_i < m_{dc} \\ 1, & v_i \geq m_{dc} \end{cases} \quad (6)$$

where $V = (v_i)_{i=1...l}$ and $W = (w_i)_{i=1...l}$

8. A pseudo-random sequence $Y = (y_i)_{i=1...l}$ is generated using the chaotic system presented in Fig. 1, with the secret key K.

9. The binary feature vector W is enciphered using the pseudo-random sequence Y and the final hash value $H = (h_i)_{i=1...l}$ is obtained, where: $h_i = w_i \otimes y_i$, $i = 1...l$

At the end of all steps, a hash value with $l-$ bits length is obtained. The proposed system is illustrated in Fig. 2.

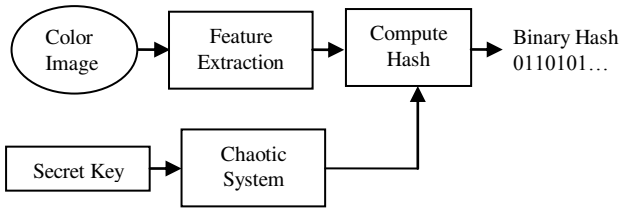


Fig. 2 – Proposed Hashing System

IV. Experimental Results

A. Simulations description

Several parameters were tested during the simulations. This parameters were the dimension of the resized image (M), the $n-$ level of the DWT transform and the block size kxk of the DCT transform. In this paper we illustrate the performances of the proposed algorithm for the following parameters:

a) $M = 256,\ n = 1,\ k = 8,\ l = 256$

b) $M = 256,\ n = 2,\ k = 4,\ l = 256$

The investigation of the proposed scheme was performed for hash values with constant binary lengths $l = 256$. All feature sets were encrypted using the 256–bit length chaotic sequences generated according to [8]. Each bit error from the binary hash contributes to the final error with the value $\frac{1}{l}$ (i. e. $err \approx 0.0039$).

For our purposes we used three different databases with resized images between 512x384 and 1024x1024. The investigation of the proposed algorithm was conducted using the following databases: Uncompressed Color Image Database (UCID) (color images), Break Our Steganographic System (BOSS) database (color images) and Break Our Watermarking System (BOWS) database (gray images). 1000 uncompressed images with different formats (tif, bmp and pgm) were randomly chosen from each database to create our testing set containing 3000 images. Images from BOSS database were converted from CR2 (Cannon Raw file format) in bmp format.

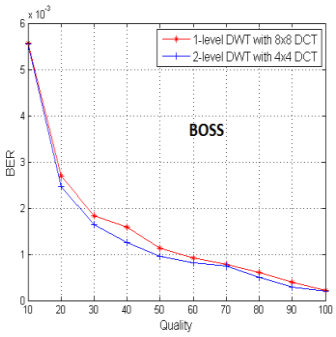
To define the similarity between the reference hash and the target hash, we use the Bit Error Rate (BER) as a measure of number of differences. The BER value for two hashes is given by the ratio between the number of erroneous bits and the total number of bits. A perfect similarity equates with a 0 BER value and two completely different images should provide a BER value close to 0.5 (not similar).

B. Robustness against JPEG Compression

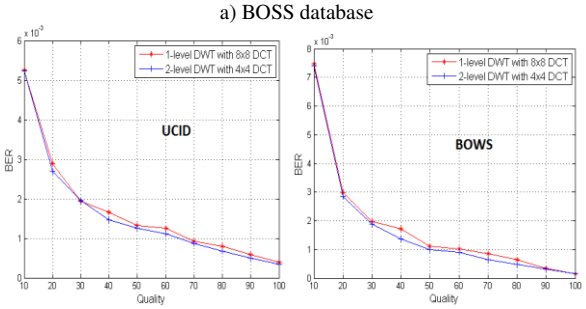
Our tests for the proposed hash function aimed primarily the resilience of the method to the JPEG compression with different quality factors $(Q = 10, 20...100)$. A number of 1000 uncompressed digital images from each database were compressed with different quality factors, from 10 to 100. All hash values computed from uncompressed images were compared with hash values calculated for the corresponding JPEG image compressed with quality factor Q. The robustness achieved under JPEG compression for all three databases are independently illustrated in Fig. 3. Our results prove the robustness of the proposed method for both color and gray images at compression factors down to $Q = 10$. A DWT 2-level decomposition of the image jointed with the DCT 4x4 decomposition proved slightly better results in terms of robustness against JPEG compression for all three image sets.

C. Robustness against other image processing attacks

The proposed hashing algorithm exploits the advantage of transform domain and also provides some robustness against common image processing attacks such as filtering, noise addition and some geometric transforms. The BER value calculated between the hashes was also used as metric to measure the similarity between the original and attacked images. Several attacks were applied on the Lena image stored in JPEG format compressed with $Q = 90$ and with size 512x512. All attacks were conducted using the Checkmark framework [11] with the specific parameters and the results are presented in Table 1. The proposed hashing scheme performed well under filtering, noise addition and some geometric attacks. However, the method proved to be vulnerable against geometric manipulations such as rotations greater than 2^o and certain cropping attacks.



a) BOSS database



b) UCID database c) BOWS database

Fig. 3 - Robustness against JPEG compression

TABLE I.
ROBUSTNESS AGAINST CHECKMARK ATTACKS FOR IMAGE LENA

Attacks [11]	**Bit-Error-Rate (BER)**
Gaussian Noise(1)	0
Hard Thresholding(1)	0.0039
Soft Thresholding(1)	0.0039
Wiener Filtering	0.0039
Median Filtering(1)	0.0156
Sharpening	0.0117
Shearing(1)	0.1133
Stirmak(1)	0.0156
JPEG compression 10	0
Wavelet Compression 10	0
Denoising with Remodulation(1)	0.0039
Sample Down(1)	0.0117
Template Remove	0.0039
NullLineRemove(1)	0.0078
Rotation 2°	0.0273
Rotation 10°	0.4961
Scale 40%	0.0117
Crop 40%	0.1796

(1) This test is available in Checkmark Benchmark in several variants. Worst result is presented.

D. Robustness against malicious attacks

An attacker may perform two types of malicious attacks. The former implies the counterfeiting of both digital image and hash value. A second type of manipulation is by direct modification of the image content, while retaining the hash value of the image. The first class of attacks is unfeasible for the proposed scheme due to the secrecy of the enciphering key of the chaotic system. The resilience of the proposed algorithm against this class of attacks is given by the strength of the chaotic system and the secrecy of the key. For the latter class of attacks, the image may be maliciously distorted using the following techniques: object addition, object removal and object altering. Our block based hashing scheme is less sensitive to local modifications. Changing small parts of the image is reflected by the DC coefficients obtained for the DCT transform applied for the LL sub-band. However, these changes are not fully reflected in the binary feature vector. This is because the averaging procedure used to collect the feature set is not always sensitive to this type of modification. The use of a threshold very close to 0 may assure a partial robustness against this class of attacks.

A feasible solution for images containing text elements (letters, numbers, visible watermarks etc.) is using character identification techniques. All extracted text elements may be hashed using a robust cryptographic hash function. This hash value is concatenated to the perceptual hash value and a final hash is built. The use of SHA-256 as cryptographic hash function assures a 512 bit length of the final hash value.

E. Collision Resistance

A perceptual hashing scheme should provide different hashes for dissimilar images. The proposed algorithm complies with this requirement and provides different hash values for different images. In order to illustrate the collision resistance property of the proposed image hashing scheme an example is illustrated in Fig. 4 for images AudiA4_1.jpg and AudiA4_2.jpg (source: www.autovit.ro) The BER value calculated for the images presented in figure 0.5078 indicates the total difference between the hashes of the two distinct images.

Fig. 4 – a) AudiA4_1.jpg b) AudiA4_2.jpg
BER = 0.5078

However, the BER value of dissimilar images is not always close to 0.5. In Fig. 5 we illustrate the discriminative capability of the proposed algorithm, by computing the probability density function of BER values for dissimilar images. This result was obtained for 1000 image pairs, randomly extracted from the test databases. The BER values calculated between perceptual hashes of distinct images have a Gaussian distribution, with the mean 0.4763, which is close to the theoretical value 0.5.

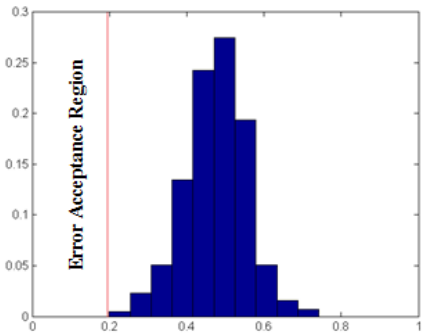


Fig. 5 – Probability density function for 1000 dissimilar image pairs

The minimum value (i. e. 0.1992) is far enough from most of the values obtained for Lena image and its attacked versions using Checkmark Benchmark. A BER threshold value fixed at 0.05 assures very good performances for our perceptual hashing method.

III. Practical Applications of the Proposed Hashing Scheme

The goal of the proposed image hashing system was to cover the following three topics: image authentication, image retrieval and copyright protection. In all three cases the reference image is required. For each target image the hash value is computed using the same secret key as for the reference image. Two hashes are computed at the bit level in order to determine the similarity level.

The verification system is built according to Fig. 6.

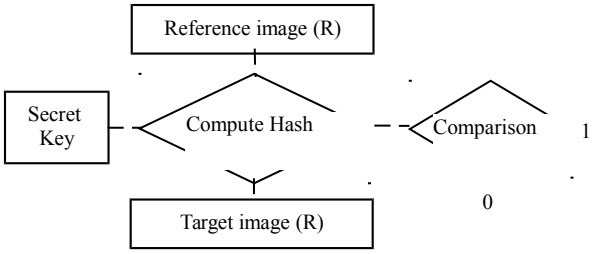


Fig. 6 Verification System

After comparing the two hashes at bit level, a *BER* value is computed. Depending on the sensitivity of the application integrating the image hashing scheme, different threshold value of the BER should be chosen. *BER* values above the threshold outputs a binary 0 (non-authentic image) and *BER* values under the threshold outputs 1 (authentic image). A *BER* value should be close to 0 when very high sensitivity is required (e.g. authentication of tampered images) and the may be increased when the application is not very restrictive (e.g. applications with content identification such as *TinEye*, *Google Images*).

IV. Conclusions and Future Works

In this paper we investigated the concept of image hashing in frequency domain secured with the aid of chaotic sequences. Our feature points are extracted using jointed *DWT* and *DCT* transforms. The feature set is enciphered using a robust chaotic map. A large set of natural images was tested in order to measure the performances of the proposed method. Experimental results illustrated a good robustness of the proposed method against known attacks such as compression, filtering, noise addition and slight geometric transformations. The proposed scheme may be applicable for image authentication, copyright protection and image retrieval. In part of future research, we will concentrate on an alternative approach which is more robust against geometric transforms and tampering attacks.

Acknowledgment

The author thanks Professor Adriana Vlad for the guidance concerning the use of the chaotic system from this work and for all technical advices and discussions about chaos based cryptography.

References

[1] V. Monga and B. Evans, "Robust perceptual image hashing using feature points," *in Proc. IEEE Int. Conf. Image Processing, Singapore*, pp. 677-680, 2004, doi: 10.1109/ICIP.2004.1418845.

[2] L. Weng and B. Preneel, "A secure perceptual hash algorithm for image content authentication," in *Proc. Of IEEE International Conference on Signal Processing and Communications*, pp. 1063-1066, 2007.

[3] F.-X. Yu, Y.-Q. Lei, Y.-G. Wang and Z.-M. Lu, "Robust image hashing based on invariance of DCT coefficients," JIH-MSP 2010, vol.1, pp.286-291.

[4] M. K. Mihcak and R. Venkatesan, "New iterative geometric methods for robust perceptual image hashing," in *Proc. ACM Workshop Security and Privacy in Digital Rights Management*, Philadelphia, 2005, doi: 10.1007/3-540-47870-1_2.

[5] S. S. Kozat, K. Mihcak, and R. Venkatesan, "Robust perceptual image hashing via matrix invariances," *Proc. IEEE Conf. on Image Processing*, pp. 3443-3446, 2004, doi: 10.1109/ICIP.2004.1421855.

[6] J. Fridrich and M. Goljan, "Robust hash functions for digital watermarking," in *Proc. IEEE Int. Conf. Information Technology: Coding Computing*, pp. 178-183, 2000.

[7] A. Swaminathan, Y. Mao and M. Wu, "Image hashing resilient to geometric and filtering operations," in *Proc. IEEE Workshop on Multimedia Signal Processing*, Siena, Italy, Sep. 2004, doi: 10.1109/MMSP.2004.1436566.

[8] A. Vlad, A. Luca, O. Hodea and R. Tataru, "Generating chaotic secure sequences using tent map and a running-key approach," in *Proc. of The Romanian Academy*, Series A, vol. 14, pp. 292-302, 2013.

[9] X. X. Guo and H. Dimitrios, "Content based image via wavelet and radon transform," in *Proc. Of the 8th Pacific Rim Conference on Multimedia*, Hongkong, China, vol. 4810, pp. 755-764, 2007, doi 10.1109/ICIEA.2007.4318736.

[10] X. Lv and Z. J. Wang, "Fast Johnson-Lindenstrauss Transform for Robust and Secure Image Hashing," *Proc. of the IEEE 10th Workshop on Multimedia Signal Processing (MMSP)*, pp: 725-729, 2008, doi:10.1109/MMSP.2008.4665170.

[11] S. Pereira, S. Voloshynovskiy, M. Madueno, S. Marchand-Maillet and T. Pun, "Second generation benchmarking and application oriented evaluation," in *Information Hiding Workshop*, Pitsburgh, PA, USA, 2001.

[12] G. Schaefer and M. Stich, "Ucid – An uncompressed colour image database," in *Proc. SPIE, Storage and Retreval Methods and Applications for Multimedia*, pp. 472-480, San Jose, U.K, 2004.

[13] P. Bas, T. Filler and T. Pevny, "Break our steganographic system – the ins and outs of organizing BOSS," in *Proc. of Information Hiding Conference*, Prague, 2011, doi: 10.1007/978-3-642-24178-9_5.

[14] T. Furon and P. Bas, "Broken arrows," *EURASIP Journal on Information Security*, 2008, doi:10.1155/2008/597040.

2nd Workshop on Scalable Computing in Distributed Systems and 7th Workshop on Large Scale Computations on Grids

The Large Scale Computing in Grids (LaSCoG) workshop originated in 2005, and when it was created we have stated in its preamble that:

"The emerging paradigm for execution of large-scale computations, whether they originate as scientific or engineering applications, or for supporting large data-intensive calculations, is to utilize multiple computers at sites distributed across the Internet. In particular, computational Grids are collections of distributed, possibly heterogeneous resources which can be used as ensembles to execute large-scale applications. While the vision of the global computational Grid is extremely appealing, there remains a lot of work on all levels to achieve it."

While, it can hardly be stated that the issues we have observed in 2005 have been satisfactorily addressed, a number of changes has happened that expanded the world of large-scale computing. Today we can observe emergence of a much more general paradigm for execution of large-scale applications, whether they originate from scientific or engineering areas, or they support large data-intensive calculations. These tasks utilize computational Grids, cloud-based systems and resource virtualization. Here, collections of distributed, possibly heterogeneous resources, are used as ensembles to execute large-scale applications.

This being the case, we have decided to keep the LaSCoG workshop tradition alive, but to co-locate it with a conference which will have an appropriately broader scope. This is how the Workshop on Scalable Computing in Distributed Systems (SCoDiS'14) emerged.

The LaSCoG-SCoDiS'14 pair of events shares a joint Program Committee and is envisioned as a forum to promote an exchange of ideas and results aimed at addressing sophisticated issues that arise in developing large-scale applications running on heterogeneous distributed systems.

Topics

Covered topics include (but are not limited to):

- Large-scale algorithms and applications
- Symbolic and numeric computations
- High performance computations for large scale simulations
- Large-scale distributed computations
- Agent-based computing
- Data models for large-scale applications
- Security issues for large-scale computations
- Science portals
- Data visualization
- Performance analysis, evaluation and prediction
- Programming models
- Peer-to-peer models and services for scalable Grids
- Collaborative science applications
- Business applications
- Data-intensive applications
- Operations on large-scale distributed databases
- On-demand computing
- Computation as a service
- Federation of compute capacity
- Virtualization supporting computations
- Self-adaptive computational / storage systems
- Volunteer computing
- Large scale computation with GPU
- Cloud computing architectures and models
- Load Balancing in large-scale distributed systems
- Intelligent resource allocation
- Cloud Security, Privacy, Confidentiality and Compliance
- Mobile Cloud
- High Performance Cloud Computing
- Green Cloud Computing
- Economic, Business and ROI Models for Cloud Computing
- Performance, Capacity Management and Monitoring of Cloud Configuration
- Cloud Interoperability and Portability
- Cloud Application Scalability and Availability
- Big Data Cloud Service

Event Chairs

Gusev, Marjan, University Sts Cyril and Methodius, Macedonia

Paprzycki, Marcin, Systems Research Institute Polish Academy of Sciences, Poland

Petcu, Dana, West University of Timisoara, Romania

Ristov, Sasko, University Sts Cyril and Methodius, Macedonia

Program Committee

Anderson, David, University of California, Berkeley, United States

Bass, Len, NICTA, Australia

Brodnik, Andrej, University of Ljubljana, Faculty of Computer and Information Science, Slovenia

Camacho, David, Universidad Autonoma de Madrid, Spain

D'Ambra, Pasqua, ICAR-CNR, Italy

Feldmann, Anja

Filippone, Salvatore, University Rome Tor Vergata, Italy

Ganzha, Maria, University of Gdańsk and Systems Research Institute Polish Academy of Sciences, Poland

Gepner, Paweł, Intel Corporation, United Kingdom

Gordon, Minor, Software development consultant, United States

Gorgan, Dorian, Technical University of Cluj-Napoca, Romania

Goscinski, Andrzej, Deakin University, Australia

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 743–752
DOI: 10.15439/2014F234
ACSIS, Vol. 2

Synthesis of Real Time Distributed Applications for Cloud Computing

Stanisław Deniziak
Cracow University of Technology, Department of Computer Engineering
Warszawska 24, 31-155 Cracow, Poland
Kielce University of Technology, Department of Computer Science
Al. Tysiaclecia Panstwa Polskiego 7, 25-314 Kielce, Poland, Email: sdeniziak@pk.edu.pl

Sławomir Bąk
Cracow University of Technology, Department of Computer Engineering
Warszawska 24, 31-155 Cracow, Poland
Email: sbak@pk.edu.pl

***Abstract*—This paper presents the methodology for the synthesis of real-time applications for the Infrastructure as a Service (IaaS) model of cloud computing. We assume that the function of the application is specified as a set of distributed echo algorithms with real-time constraints. Then our methodology schedules all tasks on available cloud infrastructure minimizing the total costs of the IaaS services, while satisfying all real-time requirements. It takes into account limited bandwidth of communication channels as well as the limited computation power of server nodes. The optimization is based on the iterative improvement algorithm, which has the capability of escaping from the local extrema, giving better results than greedy algorithms. The method starts from the fastest solution and in the next steps modifies the solution to reduce the cost of hiring the cloud infrastructure. We also present a sample application, that shows the benefits of using our methodology.**

***Index Terms*—cloud computing, Infrastructure as a Service, real-time system, distributed systems, system synthesis.**

I. Introduction

CLOUD computing recently has received significant attention as a new computing infrastructure. A cloud environment often has hundreds of thousands of processors with numerous disks interconnected by dedicated high-speed networks. There are three deployment models of cloud computing [1]. The first is the private cloud, it works specially for organization with private security and exclusive network. The second is the public cloud, it gives the maximum efficiency level in shared resources and it is protected by the cloud service provider. The third is the hybrid cloud, it combines the private and public. Cloud computing supports three types of services [2]:

- IaaS (Infrastructure as a Service) offers end users direct access to processing, storage, and other computing resources. IaaS allows users to configure resources, to run operating systems and to run application software on them. Examples of IaaS are: Amazon Elastic Compute Cloud (EC2), Rackspace and IBM Computing on Demand,
- PaaS (Platform as a Service) offers an operating system as well as suites of programming languages and software development tools that customers can use to develop their own applications. Examples of PaaS are: Microsoft Windows Azure and Google App Engine. PaaS gives end users control over application design, but does not give them control over the physical infrastructure,
- SaaS (Software as a Service) offers final applications that end users can access through a thin client (web browser). Examples of SaaS are: Gmail, Google Docs. The end users (customers) do not exercise any control over the design of the application, servers, networking and storage infrastructure.

Cloud computing is really changing the way, how and where the computing is going to be performed. More and more Internet-enabled devices are now available (mobile phones, smart TVs, navigation systems, tablets, etc.). It is expected that in a few years, almost each product may be identified and traced in the Internet using RFID (Radio Frequency IDentification), NFC (Near Field Communication) or other wireless communication methods. Smart device not only incorporates sensing/monitoring and control capabilities, but also may cooperate with other devices and with Internet applications. For example an adaptive car navigation system may interact with an Internet system, controlling and monitoring the traffic in a city, to avoid traffic jam. In such case cloud applications are used to process requests sent by smart devices implementing client applications. Usually responses to the device should be sent during the limited time period. Therefore, this class of application is a real-time system.

Distributed Internet application requires an expensive network platform, consisting of servers, routers, switches, communication links etc., to operate. The cost of the system may be reduced by sharing the network infrastructure between different applications. This is possible by using the Infrastructure as a Service (IaaS) model [3] of the cloud computing services [4]. IaaS together with a real-time cloud environment [5] seems the ideal platform for many real-time cloud applications. But to guarantee the quality of service

978-83-60810-58-3/$25.00

and minimize the cost of the system, efficient methods of mapping real-time applications onto IaaS should be developed.

Some studies [6], [7] consider resource allocation for cloud applications. The common focus of these works is the optimization of resource allocation from IaaS in respect of the cost. One of the previous method selecting resources from a cloud is based on the conception of the game theory [8]. The method optimizes the cost and the performance. This conception reflects the common characteristics of the physical position and bandwidth available between job and resources, and emphasizes on establishing a scheduling relationship between near entities. In resource scheduling, a choice of near and low-cost resources is a key criterion. Paper [9] also describes the scheduling algorithm for cloud computing. In this cost-based method, the set of computing resources with the lowest price are assigned to the user, according to the current supplier resource availability and a price. Another method, proposes scheduling of resources, based on genetic algorithm [10]. In this method, scheduling scheme is coded using integer sequence and a fitness function is based on influence degree. The genetic operations include selection, crossover, mutation and elitist selection. None of the above methods consider real-time requirements.

The use of the cloud infrastructure for real time computing is a quite new concept. Current work concerning Real Time Cloud Computing mainly concentrates on 2 domains: adopting existing web technologies to this new paradigm and developing software architectures for real-time applications. Recent studies [11]–[14] have been performed on the allocation of resources for real time tasks. Aymerich *et al.* [11] developed an infrastructure for a real-time financial system based on cloud computing technologies. Liu *et al.* [12] showed how to schedule real-time tasks with different utility functions. The real-time tasks are scheduled non-preemptively with the objective to maximize the total utility by using time utility function (TUF). Tsai *et al.* [13] discuss about a real-time database partitioning on cloud infrastructures. Kim *et al.* [14] investigate power-aware provisioning of resources for real-time cloud services. In their work the real-time constraint is specified in a Service Level Agreement (SLA) between customers and cloud providers. SLAs specify the negotiated agreements, including Quality of Service (QoS), such as deadlines. In such cloud models the service provider is responsible for the allocation resources. Their work examines power management while allocation of resources should meet the SLA. None of these studies consider a cost-efficient selection, from a set of different types of resources available in clouds, for real-time tasks.

The closest work to ours is that of Kumar *et al.* [15]. They develop an algorithm of resource allocation for applications with real-time tasks. They propose an EDF-greedy scheme and a scheme considers temporal overlapping to allocate resources efficiently. Unfortunately an EDF-greedy strategy may not give the lowest total cost, because of their tendency to be trapped in local minima of the cost.

In our work, we consider the IaaS model of the real-time cloud computing, where the user pays the cost of using the resources supported by the service provider. We present the methodology for the synthesis of reactive, real-time cloud applications specified as a set of distributed echo algorithms. The goal of our methodology is to find the distributed architecture of the application which will satisfy all user requirements. We developed an iterative improvement algorithm, which is able to escape from the local extrema, giving much better results than constructive algorithms. Presented method also minimizes the cost of IaaS services required for running the real-time application in the cloud environment.

The next section presents our assumptions and it defines the concept of real-time cloud computing used in our methodology. In section 3 the method of synthesis will be described. Section 4 presents example and experimental results demonstrating the advantages of the methodology. The paper ends with conclusions.

II. Problem statement

System synthesis is a process of automatic generation of the system architecture, starting from the formal specification of functional and non-functional requirements. Functional requirements define functions that should be implemented in the target system. Nonfunctional requirements usually define constraints that should be fulfilled, e.g., time constraints define the maximal time for execution of the given operations, cost requirements define the maximal cost of the system, etc.

Functions of distributed systems are usually specified as a set of communicating tasks or processes. Since we consider real-time systems, hence time constraints are the main set of requirements. The model of the system specification used in our methodology will be described in p.1.

We use existing network infrastructure, hired from a cloud (IaaS), consisting of servers, routers and connections. If the current architecture does not guarantee that all time requirements will be met, the infrastructure should be extended by adding some components, i.e. additional resources should be hired from cloud providers. Thus, it should be possible to specify architectural requirements that have to be satisfied by the target system. The model of the target architecture will be described in p.2, while requirements that are used in our methodology will be presented in p.3.

1. Functional specification

We assume that a real-time cloud application will process requests received from clients. The system should be able to process all requests during the required time period, i.e., for each real-time request a response should be sent before the specified deadline. We consider soft real-time processing [16], ensuring that the process will be completed at a given time depending on the constraints of quality of service. In

case of a large number of requests and a long time of processing, real-time processing will be possible only if massive parallel computing will be applied. Therefore, the functional specification of the system should represent the function as a distributed algorithm [17], developed according to the following requirements:

(1) parallel model of computations: system should be specified as a set of parallel processes using message passing communication,
(2) parallel request handling: huge number of requests may cause the communication bottleneck, to avoid this, simultaneous requests should be handled by different processes.

We assume that the system is specified as a collection of sequential processes coordinating their activities by sending messages. Specification is represented by a graph $G = \{V, E\}$, where V is a set of nodes corresponding to the processes and E is a set of edges. Edges exist only between nodes corresponding to communicating processes. Tasks are activated when required set of events will appear. As a result, the task may generate other events. External input events will be called requests (Q), external output events are responses (O) and internal events correspond to messages (M). The function of the system is specified as finite sequences of activation of processes. There is a finite set of all possible events

$$\Lambda = Q \cup O \cup M = \{\lambda_i : i = 1, \ldots, r\} \quad (1)$$

For each event λ_i communication workload $\omega(\lambda_i)$ is defined. System activity is defined as the following function:

$$\Phi : C \times V \rightarrow \omega \times 2^{\Lambda} \quad (2)$$

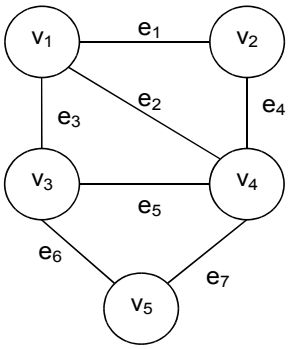
where C is an event expression (logical expression consisting of logical operators and Boolean variables representing events) and ω is the workload of the activated process.

Using function Φ it is possible to specify various classes of distributed algorithms. Fig. 1 presents sample echo algorithm [18] consisting of 5 processes. The algorithm consists of 10 actions. Each action is activated only once, when the corresponding condition will be equal to true. All actions except A_1 and A_6 contain alternative sub-actions. Only the first action, for which the condition will be satisfied, will be activated. According to the echo algorithm specification, process v_1 is the initiator, messages $m1_1$, ..., $m14_7$ are explorer messages, while $m15_1$, ..., $m25_7$ are echo messages (indices are added only for readability, mx_i means that message mx is associated with edge e_i in the graph, for the same reason, edge names in the event expressions mean any received message corresponding to this edge, e.g., $e_1 = m1_1 \mid m4_1 \mid m15_1$, $e_2 = m2_2 \mid m9_2 \mid m20_2$, etc.). Events x_1, ..., x_{11} are internal events, used for storing the state of processes between successive executions.

Since different requests may be processed by distinct algorithms, the function of a system may be specified using a set of functions Φ sharing the same processes. Each function has only one initiator (process activated by the request). Processes may be activated many times, but the algorithm should consists of the finite number of actions and infinite loops are not allowed.

2. Real Time IaaS Architecture

The proposed architecture of RTCCI (Real Time Cloud Computing Infrastructure) is composed of two layers (Fig. 2): Network Layer (*NL*) and Server Layer (*SL*).



A_1: $\Phi(v_1,\{q_1\}) \rightarrow (5, \{m1_1,m2_2,m3_3\})$

A_2: $\Phi(v_2,\{m1_1\}) \rightarrow (4,\{x_1,m5_4\}) \mid \Phi(v_2,\{m10_4\}) \rightarrow (4,\{x_2,m4_1\})$

A_3: $\Phi(v_3,\{m3_3\}) \rightarrow (7,\{x_3, m7_5,m8_6\}) \mid \Phi(v_3,\{m11_5\}) \rightarrow (7,\{x_4,m6_3,m8_6\})$
$\mid \Phi(v_3,\{m13_6\}) \rightarrow (7,\{x_5,m6_3,m7_5\})$

A_4: $\Phi(v_4,\{m2_2\}) \rightarrow (6,\{x_6,m10_4,m11_5,m12_7\})\mid$
$\Phi(v_4,\{m5_4\}) \rightarrow (6,\{x_7,m9_2,m11_5,m12_7\})\mid$
$\Phi(v_4,\{m7_5\}) \rightarrow (6,\{x_8,m9_2,m10_4,m12_7\})\mid$
$\Phi(v_4,\{m14_7\}) \rightarrow (6,\{x_9,m9_2,m10_4,m11_5\})$

A_5: $\Phi(v_5,\{m8_6\}) \rightarrow (5,\{x_{10},m14_7\}) \mid \Phi(v_5,\{m12_7\}) \rightarrow (5,\{x_{11},m13_6,\})$

A_6: $\Phi(v_1,\{e_1\&e_2\&e_3\}) \rightarrow (10, \{r_1\})$

A_7: $\Phi(v_2,\{x_1\&e_1\&e_4\}) \rightarrow (4,\{m15_1\}) \mid \Phi(v_2,\{x_2\&e_1\&e_4\}) \rightarrow (4,\{m16_4\})$

A_8: $\Phi(v_3,\{x_3\&e_3\&e_5\&e_6\}) \rightarrow (3,\{m17_3\}) \mid \Phi(v_3,\{x_4\&e_3\&e_5\&e_6\}) \rightarrow (3,\{m18_5\}) \mid \Phi(v_3,\{x_5\&e_3\&e_5\&e_6\}) \rightarrow (3,\{m19_6\})$

A_9: $\Phi(v_4,\{x_6\&e_2\&e_4\&e_5\&e_7\}) \rightarrow (5,\{m20_2\}) \mid \Phi(v_4,\{x_7\&e_2\&e_4\&e_5\&e_7\}) \rightarrow (5,\{m21_4\}) \mid \Phi(v_4,\{x_8\&e_2\&e_4\&e_5\&e_7\}) \rightarrow (5,\{m22_5\}) \mid$
$\Phi(v_4,\{x_9\&e_2\&e_4\&e_5\&e_7\}) \rightarrow (5,\{m23_7\})$

A_{10}: $\Phi(v_5,\{x_{10}\&e_6\&e_7\}) \rightarrow (2,\{m24_6\}) \mid \Phi(v_5,\{x_{11}\&e_6\&e_7\}) \rightarrow (2,\{m25_7\})$

Fig 1. Sample specification of the echo algorithm

Layer *NL* consists of Communication Channels (*CC*) composed of routers and communication links. For each CL_i the available bandwidth $B(CC_i)$ is defined.

Layer *SL* contains servers (*S*) consisting of computational nodes N_i. Each N_i is characterized by performance P_i reserved for RTCC system, and it may be equipped with a network interface. Thus, each computational node may be connected to another communication link.

The goal of our methodology is to find the cheapest system architecture for an application that fulfills all time constraints and uses the existing network infrastructure available in a cloud. All servers (nodes) and communication channels, that are used in the target architecture $\Pi_T = \{S, CC\}$ of the system, will be outsourced to the cloud provider.

The method starts from the initial architecture $\Pi_I = \{S', CC'\}$ consisting of the fastest resources. Next, the architecture is optimized by performing some modifications of Π_I, only resources supported by cloud providers are considered here. Our methodology minimizes the cost of hiring the network infrastructure by achieving the maximal utilization of all resources and by allocating cheapest components that satisfy all time constraints. Each available resource is characterized by properties defining the performance and the cost of the corresponding IaaS service. Specifications of all available resources constitute the database of resources $L = \{CC'', S''\}$.

Communication channels $cc_i \in CC''$ are characterized by the maximal available bandwidth $B(cc_i)$, bandwidth $B_r(cc_i)$ reserved for the application and the price of communication service $Cr(cc_i)$ for each available bandwidth. Communication channel connects any pair of network interface ports. Thus, the time of transmission of packet D_i through communication channel cc_j is the following:

$$T(D_i)=\frac{l(D_i)}{B_r(cc_j)} \tag{3}$$

where $l(D_i)$ is the length of packet D_i.

We assume that each server s_i may consist of any number of nodes, i.e., a multiprocessor or a cluster architecture of the server. Each node may execute all assigned tasks sequentially. Thus, the following properties characterize the server:

– n_S - the number of nodes, hence server s_i may be represented as a set $\{N_1, ..., N_{ns}\}$ of nodes,

– $Cr(s_i)$ - the cost of the computing services, the cost depends on the number of nodes allocated to the application, usually the cost function is not linear.

– $\{P_1, ..., P_{ns}\}$ - performance of each node.

The time required for executing process τ_i by the node N_j equals:

$$T(\tau_i)=\frac{w(\tau_i)}{P_j} \tag{4}$$

where $w(\tau_i)$ is the workload of task τ_i.

Fig. 2 presents a sample target architecture of RTCCI.

3. Requirements and constraints.

Let $\rho(\lambda_x,\lambda_y)$ be a sequence of actions $A_1,...,A_s$ such, that λ_x is the request, λ_y is the response, and:

$$A_1:\Phi(v_i,\lambda_x)\rightarrow\{\omega_1,\{\lambda_1\}\},\ A_s:\Phi(v_j,\lambda_s)\rightarrow\{\omega_s,\{\lambda_y\}\},\ \underset{1<k<s-1}{\forall} A_k\rightarrow A_{k+1} \tag{5}$$

where v_i, v_j are any processes and $A_k \rightarrow A_{k+1}$ means that action A_k generates events activating action A_{k+1}. Then, the time of execution of the given sequence of actions is defined as a sum of the execution times of all processes and a time of inter-process communication:

$$t(p(\lambda_x,\lambda_y))=\sum_{i=1}^{s}\frac{\omega(A_i)}{P(A_i)}+\sum_{i=1}^{s}\frac{\omega(m_i)}{B_r(m_i)} \tag{6}$$

where: $\omega(A_i)$ is the workload of the process activated by action A_i, $P(A_i)$ is the performance of the server executing this process, $\omega(m_i)$ is the communication size, $B_r(m_i)$ is the reserved bandwidth of the channel used for sending the message. If processes activated by actions A_k and A_{k+1} are executed by the same server, then $\omega(m_k) = 0$ for any message sent between these processes.

The time constraint is the maximal period of time that may elapse between sending request and receiving the response. Since the request may activate different sequences of actions until the response will be obtained, therefore the time constraint (deadline) is defined as:

$$t_{max}(\lambda_x,\lambda_y)=\underset{i}{MAX}(t(p_i(\lambda_x,\lambda_y))) \tag{7}$$

During the synthesis, processes and transmissions are scheduled and assigned to network resources. The method first assigns processes and transmissions to the fastest re-

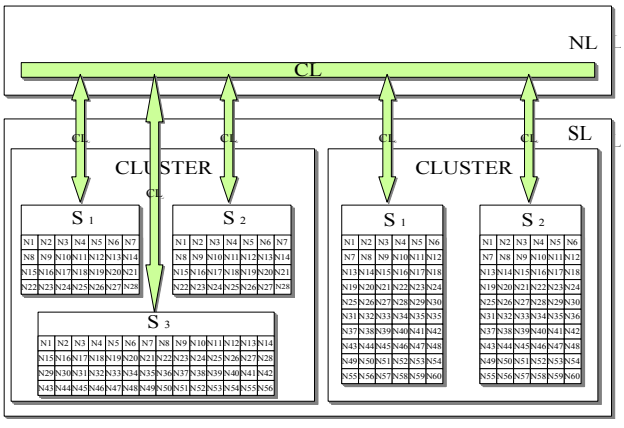


Fig 2. Sample target architecture

sources supported by cloud providers. This step verifies if it is possible to find the network infrastructure which fulfills all time constraints. Next, the cost is minimized by performing the following modifications:

- change communication channel to cheaper one, decreasing bandwidth $B_r(cc)$, for any allocated communication channel, in this way the cost of communication service $Cr(cc)$ may be reduced,
- change server s to cheaper one or reduce the number of allocated nodes, in this way the cost of computing services $Cr(s)$ will be reduced.

Only modifications that do not violate time constraints are considered. The optimization process will stop, when each considered modification of the architecture will cause violation of time requirements. Hence, the total cost of the IaaS service will be the following:

$$C_M = \sum_i Cr(cc_i) + \sum_j Cr(s_j) \tag{8}$$

The goal of our methodology is to minimize C_M.

III. Synthesis

Our method of synthesis starts from the formal specification of the system (as described in p. II.1) and tries to produce the optimal target architecture of the system, that satisfies all constraints. The method minimizes the cost of outsourcing the network infrastructure to the IaaS cloud provider.

1. Assumptions

The method is based on the worst case design. We assume that the workload of each action and sizes of all transmissions are estimated for the worst cases. All time constraints should also satisfy the following condition:

$$t_{max}(\lambda_q, \lambda_o) \leq \frac{1}{f_{max}(\lambda_q)} \tag{9}$$

where $f_{max}(\lambda_q)$ is the maximal frequency of requests λ_q, λ_o is response to the request. Otherwise, the system will be not able to process requests in real-time.

The system specification consists of a set of distributed algorithms (tasks). Our scheduling method is based on the assumption that the worst case is when all tasks will start at the same time, this corresponds to the simultaneous appearance of all requests. Thus, all tasks are scheduled in a fixed order and are activated in certain time frames. When the system will receive new request, it will be processed during the next activation of the corresponding task. Therefore, time constraints should include this delay, i.e., task should be scheduled with period equal:

$$\frac{t_{max}(\lambda_q, \lambda_o)}{2} \tag{10}$$

The goal of optimization is the minimization of the cost of outsourcing the network infrastructure to cloud providers. The method schedules tasks and transmissions on available cloud resources. We use an efficient iterative algorithm for finding the (sub-)optimal solution.

2. Dynamic task graph

The algorithm should be able to verify if after scheduling next task, it is still possible to obtain the valid system. For this purpose the dynamic task graph (DTG) is created. All tasks are simultaneously analyzed according to their order of execution, assuming that processes and transmissions will be executed by the fastest resources. Since in the system specification only the first message received by a process is relevant, all other messages are temporarily neglected. In this way the specification is converted into task graph. Next, the task graph is scheduled using ASAP (As Soon As Possible) method.

3. Algorithm of the synthesis

In our earlier works [19], [20] synthesis is performed using the greedy algorithm, that schedules processes according to their priority. However, it is constructive algorithm and the obtained results are far from optimal, because the method is prone to be trapped in the local minima of the cost. In this paper, we present the iterative improvement algorithm, which is able to escape from the local minima, giving better results than constructive algorithm. An outline of the algorithm is shown on Fig. 3.

Gain is the difference of quality of the new solution and the current one. The quality of the solution is determined on the basis of several features of the target architecture. In our case, the quality of solutions is based on the cost of the system that satisfies all time constraints, i.e., the optimum is the cheapest system that meets the time constraints.

The algorithm starts from the initial architecture where all processes are assigned to resources with the highest performance, according to the rule: the biggest task to the fastest processor. For transmissions, also communication services with the highest bandwidth are reserved. Next, while any time constraint is not violated, the method tries to reduce the cost of IaaS services by modifying the network infrastructure using iterative improvement methods (refinement of the current result). The methodology repeats the following steps:

- remove the node or replace the resource with a cheaper one,

```
Generate initial solution Π^CUR
do{
  Π^BEST = Π^CUR;
  gain = 0;
  while((Π'=refine(Π^BEST)) != Φ){
   gain = Q(Π') - Q(Π^CUR);
   if(gain > 0)
     Π^CUR = Π';
  }
}while(gain > 0);
```

Fig 3. An outline of the iterative improvement algorithm.

- create a dynamic task graph,
- schedule all processes and transmissions.

The solution giving the best gain is chosen to the next step. The algorithm terminates when there is no solution that can improve the total cost of the system.

The quality of the solution determines the gain of the improvement. The aim of the algorithm is to minimize the cost, thus the main system feature determining the gain should be the cost of the system. However, driving refinement only according to the optimisation goal usually leads to trapping the algorithm in local minima (the greedy approach). Hence it is appropriate to define the quality of the solution using also other features of the solution. It should inhibit the greed of the algorithm. For this purpose laxity L is introduced. The laxity is defined as follows:

$$L = t_{max} - t_{cur}, L \geq 0 \tag{11}$$

where t_{cur} is the execution time for the current solution.

At each step, various modifications of the current system are considered. Each modification may change the cost and/or the latency of the solution. Quality (Q) of the modified system is defined as follows:

$$Q = \begin{cases} \frac{C_{BEST} \cdot L_{CUR}}{C_{CUR} \cdot L_{BEST}}, when\ \Delta C < 0 \text{ and } L_{CUR} \geq 0 \\ 0, when\ \Delta C > 0 \text{ or } L_{CUR} < 0 \end{cases} \tag{12}$$

where:

$$\Delta C = C_{CUR} - C_{BEST} \tag{13}$$

L_{CUR} and C_{CUR} are the features of the current result, L_{BEST} and C_{BEST} are the features of the best result, found in the previous iterations.

The quality is defined as the ratio of the previous cost to the cost after modification. If the latency is also changed, then the quality is modified by the percentage of the latency increase. If there is no reduction of the system cost, then the quiality equals 0, i.e., modification will be rejected. This condition guarantees that the algorithm is convergent. The quality is also equal to 0 when a time constraint is violated, i.e., $L_{CUR} < 0$.

Solutions that do not lead to a gain greater than 0 are rejected. The quality prefers solutions with the greatest reductions of cost and greater increase of the performance of the system. If all modifications do not reduce the cost of the system ($\Delta C=0$) then the solution with the greatest increase of the system performance is taken as the best to the next step.

At each step of the algorithm, various modifications of the current solution are considered and solution that gives the highest quality is chosen to the next step. Since the quality Q depends also on the increase of laxity, therefore the greed of the algorithm will be reduced, i.e., instead of the modification reducing cost the algorithm may select modification that more reduces the laxity. Higher laxity means more possiblilities of improvements in the next steps.

In order to minimize the cost of the system, in the algorithm the following modifications are considered:

(1) Change the node from the cloud to cheaper one and move tasks to it.
(2) Replace the communication link to cheaper one.
(3) Remove one node and move all assigned tasks to other nodes.

In the case where more than one task is allocated to the resource, it is necessary to schedule tasks. The FIFO scheduling method is used for this purpose. The refinement process is presented on Fig. 4. It consists of 3 loops, each loop evaluates all possible modifications of the system architecture. Only systems with quality greater than 0 are taken into consideration. We assume that the process *refine* returns the architecture, then after next activation it continues its execution. The process terminates after analyzing all possible modifications of the initial architecture. Architecture with the highest quality is taken as an input to the next step of the algorithm.

IV. Example

As an example demonstrating our methodology we present the design of an adaptive navigation system for a smart city [19]. We assume that all cars are equipped with GPS navigation devices (GD), that are able to communicate with the Internet using wireless communication (we assume that the network of access points covers the whole city). GD devices send requests to RTCC system. Requests contain information about current position, the destination and user preferences. Then, the system finds the optimal route and sends response to GD device. Since GD expects the response in a reasonable time, then the system should satisfy real-time constraints. We assume that the time in which the GD device has to get an answer must be no longer than 5 seconds. The idea of such system is based on the adaptability, i.e., the system may take into consideration traffic information, traffic impediments (e.g., car accidents) and it may construct different routes for the same destinations to avoid traffic jams.

Since the system may receive thousands of requests per second the centralized system may not be able to handle all requests due to the communication bottleneck. Therefore, we propose the distributed system. The city is partitioned into sectors, routes through each sector are computed by different processes. Each process also receives requests and sends responses from/to positions in the corresponding sector. Thus, the function of the system may be specified as a set of distributed algorithms, similar to the echo algorithm. In our example the specification consists of 6 to 12 tasks, in each task another process is the initiator. The initiator receives all requests coming from the corresponding sector, computes all possible routes to adjacent sectors and sends the information about routes to adjacent processes. When messages will reach the destination sector, then the best route is selected and information about it is sent back to the initiator.

```
refine(Π^CUR){
    for each X_i ∈ Π^CUR do {
      Π′ = Π^CUR − X_i;
      for all X_j in IaaS do
         if C(X_j ) < C(X_i) then {
           Π″ = Π′ + X_j;
           for each v_k ∈ X_i do  //transfer tasks from X_i to X_j
             Assign v_k to X_j;
           if Q(Π″) > 0 then
              return Π″;
         }
    }
    for each CL_i ∈ Π^CUR do {
      Π′ = Π^CUR − CL_i;
      for all CL_j in IaaS do
        if C(CL_j ) < C(CL_i) then {
           Π″ = Π′ + CL_j;
          for(cl_k ∈ CL_i){  //transfer transmission from CL_i to CL_j
             Assign cl_k to Cl_j;
           if Q(Π″) > 0 then
              return Π″;
        }
    }
    for each X_i ∈ Π_CUR do {
     Π″ = Π^CUR − X_i;
     for each v_k ∈ X_i do {  //transfer tasks from X_i to other resource from Π^CUR
       Find resource X_j ∈ Π″ such, that L(Π″)is maximal after
          assigning v_k to X_j;
       Assign v_k to X_j;
     }
     if Q(Π″) > 0 then
       return Π″;
    }
   return Φ;
}
```

Fig 4. Synthesis algorithm for cost minimization.

Assume that a cloud provider offers 13 servers and 4 bandwidths for communication services (Fig. 5), and assume that parameters of available resources are known. In Table I available bandwidths of communication links and the cost of IaaS communication services are presented. Table II shows parameters of servers available in the cloud and costs of IaaS computing services are presented. The time constraint t_{max} equals 5 s.

Some dynamic tasks graph constructed for the best solution are presented on Fig 6. On Fig 7 the Gantt chart presenting the scheduling of all processes is presented. We may observe high utilization of all servers.

TABLE I COST OF AVAILABLE IAAS COMMUNICATION SERVICES

Link	Bandwidth (Mbps)	Per hour
Lx1	1	0.0001$
Lx2	5	0.0010$
Lx3	10	0.0028$
Lx4	20	0.0069$

TABLE II: COST OF AVAILABLE IAAS SERVICES FOR ONE SERVER.

Server	Processor	Per hour
S1	1.7 GHz	0.004 $
S2	2.4 GHZ	0.008 $
S3	2 × 1.7 GHz	0.007 $
S4	2 × 2.4 GHz	0.014 $
S5	4 × 1.7 GHz	0.013 $
S6	4 × 2.4 GHz	0.025 $
S7	4 x 1.5 GHz	0.012 $
S8	4 x 1.2 GHz	0.01 $

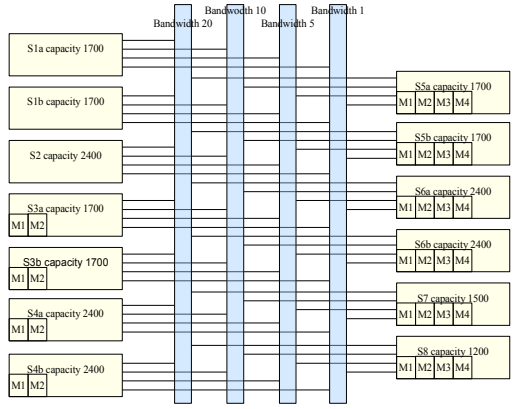


Fig 5. Database of resources available in the cloud.

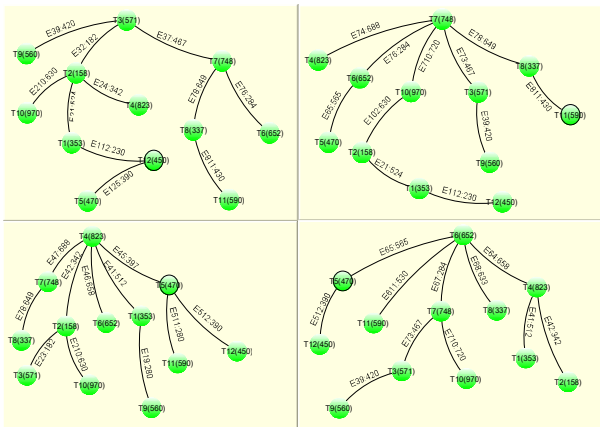


Fig 6. Several sample tasks graph

The frequency of task activation depends on the number of requests appearing during the given time period. For a large number of requests the system will require more computing power. Thus, the cost of IaaS services strongly depends on the maximal estimated traffic in the city. Fig. 8 and Table III present the dependence between the number of requests and the cost of IaaS services for greedy and iterative improvement algorithm. To allocate resources from the IaaS, iterative improvement algorithm produced much better results than greedy algorithm. The comparison is shown in Table III. Our algorithm allows the end users of IaaS to reduce the cost of hiring cloud resources by over 50%.

V. CONCLUSIONS

We analyze the problem of allocating resources for real-time tasks such that the cost is minimized and all the deadlines are met. In this paper the methodology for the synthesis of reactive, real-time cloud applications accordant with the Cloud Computing concept, was presented. We developed the architectural model of the reactive RTCC system and we proposed the method of specification for such systems, in the form of a set of distributed Echo algorithms.

Next, the method of synthesis that guarantees the fulfillment of all time requirements was proposed. The method schedules all processes and transmissions on network resources supported by cloud providers, while the cost of IaaS services is minimized. Finally, we presented the design process of the sample RTCC system, which underlines the advantages of our methodology above greedy algorithm.

In our approach we use iterative improvement algorithm for scheduling and allocation of new resources and we show its advantage over the heuristic greedy algorithm. In the future work we will consider developing a more sophisticated method of optimization as well as more advanced methods for the worst case analysis. Reactive RTCC systems are a new challenge for future Cloud Computing. We believe that in the future, RTCC systems will constitute an important class of Cloud Computing systems, thus efficient design methods will be very desirable.

REFERENCES

[1] IBMSmartCloud http://www.ibm.com/cloud-computing/us/en/what-is-cloud-computing.html. Last access, April 2014

[2] C. S. Yoo, “Cloud Computing: Architectural and Policy Implications”. *Review of Industrial Organization*, June 2011, 38.4: 405-421, http://dx.doi.org/10.1007/s11151-011-9295-7.

[3] A. Amies, H. Sluiman, QG. Tong and GN Liu, “Infrastructure as a Service Cloud Concepts. Developing and Hosting Applications on the Cloud” IBM Press, 2012.

[4] R. Buyya, J. Broberg, A. Goscinski, “Cloud Computing: Principles and Paradigms" New York, USA: Wiley Press., 2011. pp. 1–44, http://dx.doi.org/10.1002/9780470940105.

[5] D. Kyriazis et al, “A Real-time Service Oriented Infrastructure” *Annual International Conference on Real-Time and Embedded Systems (RTES 2010)*. November 2010, Singapore. pp. 39-44, http://dx.doi.org/10.5176/978-981-08-7654-8_R-47.

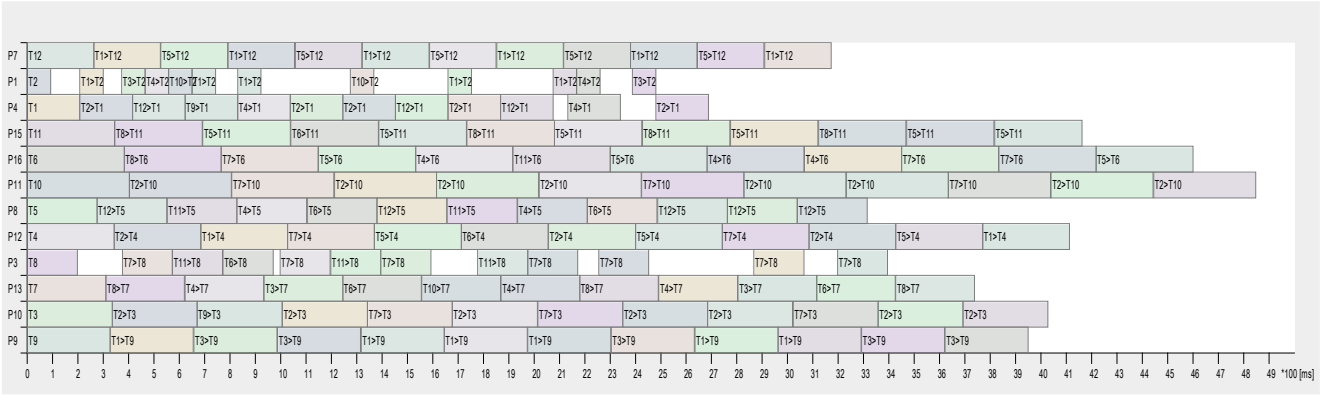


Fig 7. Gantt chart for target scheduling of processes

TABLE III TIME AND COST OF TASKS FOR GREEDY AND ITERATIVE IMPROVEMENT ALGORITHM

Lp.	Number of tasks	Greedy algorithm		Iterative improvement algorithm	
		Time [ms]	Cost [$/h]	Time [ms]	Cost [$/h]
1	36	4658.3333	0.01600	4658.3333	0.01600
2	49	4800.8333	0.02200	4884.3333	0.02155
3	64	4986.6667	0.05035	4346.6667	0.02480
4	81	4946.0000	0.05600	4938.0000	0.03010
5	100	4872.8265	0.10625	4921.0000	0.04550
6	121	4868.2500	0.13320	4714.2235	0.05655
7	144	4907.0000	0.13945	4987.0000	0.08075
AVERAGE COST		0,07475		0,03951	

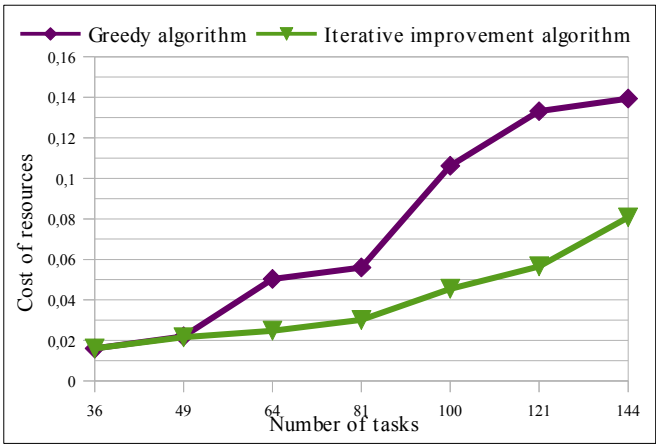


Fig 8. Dependence between the number of requests and the cost of IaaS services for greedy and iterative improvement algorithm.

[6] R. Huang, H. Casanova, A. A. Chien, "Automatic resource specification generation for resource selection" *ACM/IEEE Conference on Supercomputing*, November 2007, Reno, pp 1–11, http://dx.doi.org/10.1145/1362622.1362638.

[7] E. Deelman, G. Singh, M. Livny, B. Berriman, J. Good, "The cost of doing science on the cloud: the montage example" *ACM/IEEE Conference on High Performance Computing, Networking, Storage and Analysis*, November 2008, Austin, pp 1–12, http://dx.doi.org/10.1109/SC.2008.5217932.

[8] L. Mengkun, C. Ming, X. Jun, "Cloud Computing: A Synthesis Models for Resource Service Management" *2010 Second International Conference on Communication Systems, Networks and Applications (ICCSNA 2010),* vol.2, June 2010, Hong Kong, pp. 208-211, http://dx.doi.org/10.1109/ICCSNA.2010.5588886.

[9] Y. Zhi, Y. Changqin, L. Yan, "A Cost-based Resource Scheduling Paradigm in Cloud Computing" *12th International Conference on Parallel and Distributed Computing, Applications and Technologies*, October 2011, Washington, pp. 417-422, http://dx.doi.org/10.1109/PDCAT.2011.1.

[10] W. Ybin, T. Ling, "Research on Cloud Design Resources Scheduling Based on Genetic Algorithm" *International Conference on Systems and Informatics (ICSAI 2012)*, May 2012, Yantai, pp. 2651-2656, http://dx.doi.org/10.1109/ICSAI.2012.6223598.

[11] F. M. Aymerich, G. Fenu, S. Surcis, "A real time financial system based on grid and cloud computing" *ACM symposium on Applied Computing*, March 2009, New York, pp 1219–1220, http://dx.doi.org/10.1145/1529282.1529555.

[12] S. Liu, G. Quan, S. Ren, "On-Line Scheduling of Real-Time Services for Cloud Computing" *World Congress on Services*, July 2010, Miami, pp 459–464, http://dx.doi.org/10.1109/SERVICES.2010.109.

[13] W. Tsai, Q. Shao, X. Sun, J. Elston, "Real-Time Service-Oriented Cloud Computing" *World Congress on Services*, July 2010, Miami, pp 473–478, http://dx.doi.org/10.1109/SERVICES.2010.127.

[14] K. H. Kim, A. Beloglazov, R. Buyya, "Power-aware provisioning of cloud resources for realtime services" *International Workshop on Middleware for Grids, Clouds and e-Science*, November 2009, New York, pp.1–6, http://dx.doi.org/10.1145/1657120.1657121.

[15] K. Kumar, J. Feng, Y. Nimmagadda, Y. Lu, "Resource Allocation for Real-Time Tasks using Cloud Computing" *International Conference on Computer Communications and Networks (ICCCN)*, July 2011 Maui, pp. 1-7, http://dx.doi.org/10.1109/ICCCN.2011.6006077.

[16] G. C. Buttazzo, "Hard real-time computing systems: predictable scheduling algorithms and applications". Vol. 24. Springer, 2011, pp. 1 – 22, http://dx.doi.org/10.1007/978-1-4614-0676-1.

[17] G. Tel, "Introduction to Distributed Algorithms" Cambridge University Press, 2nd edition, 2001.

[18] E. J. H. Chang, "Echo Algorithms: Depth Parallel Operations on General Graphs" *IEEE Transactions on Software Engineering*, July 1982, pp. 391 – 401, http://dx.doi.org/10.1109/TSE.1982.235573.

[19] S. Bąk, R. Czarnecki, S. Deniziak "Synthesis of real-time cloud applications for Internet of things" *Turkish Journal of Electrical*

Engineering and Computer Sciences, to be published, http://dx.doi.org/10.3906/elk-1302-178.

[20] S. Bąk, R. Czarnecki, S. Deniziak "Synthesis of Real-Time Applications for Internet of Things" *Lecture Notes in Computer Science vol. 7719*, 2013 pp. 37-51, http://dx.doi.org/10.1007/978-3-642-37015-1_4.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 753–760
DOI: 10.15439/2014F331
ACSIS, Vol. 2

Performance Analysis of SaaS Ticket Management Systems

Pano Gushev
Innovation Dooel
1000 Skopje, Macedonia
pano.gushev@innovation.com.mk

Sasko Ristov, Marjan Gusev
Ss. Cyril and Methodius University, FCSE
1000 Skopje, Macedonia
{sashko.ristov, marjan.gushev}@finki.ukim.mk

***Abstract*—Cloud architecture has the ability of sharing hardware resources and services among multiple tenants. In this paper we measure the performance for the multi-VM (multiple virtual machines) cloud architecture and compare it with the single-VM architecture. Renting resources on a cloud usually comes with a variety of options, such as use of more and smaller virtual machines or use of less and bigger virtual machines. The objective of this research is to find out which scenario gives better performance for the same price of rented resources. This will be done by comparing the following attributes: Average response time, Pages per second, Average page time, Requests per second, CPU time. We setup a hypothesis that the multi-VM approach would be better, and the best architecture is the one offering the highest number of small virtual machines, predicting that the computational demands will spread to different virtual machines in a balanced manner. The results confirm the hypothesis and lead to a recommendation for an optimal architecture of a cloud based solution for a common transactional web solution.**

***Index Terms*—Cloud; Web Services; Windows Azure.**

I. Introduction

CLOUD computing has a growing trend in the past few years and companies increasingly understand its benefits [1]. Gartner [2] determined that the size of the cloud market is $150 Billion by 2013 and predicts that virtualised server workloads will reach a high of 60% in 2014. This trend pushes companies to migrate their services and applications in the cloud.

Benchmarking the cloud and multi-tier applications that are hosted within will help the developers to determine the most appropriate architecture, services, and settings [3]. In this paper we present results of the performance analysis of different approaches in building possible cloud architectures and solutions of the Ticket Management System (TMS).

The objective of this research is to find out which scenario performs the best for the same price of rented resources. We assume that a single processor with four cores has approximately the same price with 4 CPUs with one core, 8 GB RAM has roughly the same price with two RAMs of 4 GB etc.

Traditional distributed approach claims that the best performance is achieved when the workload is distributed to a huge number of processing units executing balanced tasks. However, it is very hard to organise the tasks with perfect load balancing in a distributed system. The cloud, on the other hand, initiates new challenges, such as predicting server CPU utilisation and resource under-provisioning. In addition, scalability and elasticity are main concerns when the customers desire to deploy their solutions on the cloud. Client expectations are very important, due to the recent huge offer of cloud providers. Summarising the above, a typical customer challenge is selection which architecture performs the best, renting more and less powerful Virtual Machines (VMs) or a smaller number of more processing powerful VMs.

We have set a hypothesis that the best performance will be achieved if the processing is distributed to more less powerful VMs. The testing environment is deploying the ticket management SaaS solution on various Windows Azure configurations with approximately the same cost.

The rest of the paper is organised as follows. In Section II, we discuss the related work and in Section III describe the testing environment, ticket management web service and test data. Section IV presents the results and Section V compares the results and evaluates the performance. The final Section VI is devoted to conclusion and future work.

II. Related work

This section presents the recent research in the area of cloud performance benchmarking, resource allocation and multi-tier transactional cloud web application.

A. Cloud performance discrepancy

Renting the same amount of IaaS (Infrastructure as a Service) resources on the Azure cloud [4] costs the same, regardless whether they are all provisioned in a single huge virtual machine or into several smaller virtual machines. This pricing scheme also follows the other most common public clouds [5], [6].

An orchestration of hardware resources on the cloud impacts the performance of the hosted cloud application or web service. Allocating the resources among many concurrent instances of virtual machines is better than using VM with multiprocessors using parallelization [7].

Gusev et al. [8] determined a region where response time is up to 10 times better if web services are spread among several small VMs, rather than in a single VM. This phenomenon motivated us to analyse how orchestration impacts the performance of the real cloud N-tier application in Azure.

Cloud multitenant virtual environment also provides discrepant performance during the time and therefore, a performance isolation is required [9]. Koh et al. reported that a VM

978-83-60810-58-3/$25.00

provides different performance on the same physical machine during the time, if it is instantiated among the other active VMs [10]. Jayasinghe et al. [11] reported that a configuration that provides the best performance in one cloud can provide the worst performance in another. Iakymchuk proposed a method to improve the performance with underutilization of physical resources by adding more nodes [12].

B. Windows Azure performance

Many authors have conducted research on Azure's performance. It is most suitable for application with small amount of communication [13] and it does not work well for tightly-coupled applications compared to Amazon EC2 or cluster [14]. Hill et al. [15] provided a performance analysis of VMs, storage services and SQL services in Azure. Azure also can decrease the costs and time for deployment, as well as support efficient TCP data transfers [16].

C. Multi-tier performance benchmarking

Today's web applications are usually 3 or 4-tier applications, commonly identified as multi-tier (n-tier) applications. Each tier is usually hosted on a separate machine or VM. These architectures are usually interesting due to several bottlenecks that can appear, which will decrease rapidly their performance.

Several authors propose benchmark tools and methodologies to simulate the multi-tier cloud web applications to analyse their performance and determine their deficiencies.

Turner et al. [17] propose a C-MART benchmark, which emulates a modern web application hosted in a cloud environment. Rubis [18] is another benchmark application that simulates an auction site and evaluates design patterns and the performance scalability of application servers. However, it cannot be used as a modern application benchmark because there are many flaws for Web 2.0 applications [19]. Also it requires a lot of effort to install, which outweighs its usefulness [20]. TPC-W is also a transactional web benchmark, which emulates an online bookstore [21]. But, it also has deficiencies [3]. In this paper, we use a real cloud SaaS application hosted in Azure to determine the scalability and user load impact on different customers' and cloud service provider's parameters.

Apart from benchmarking tools, several methodologies for benchmarking the N-tier applications are proposed and evaluated with them. Wang et al. [22] proposed a method for bottleneck detection, which correlates throughput and load. Chen et al. [23] proposed a predictive performance model that analyses cloud based N-tier web applications and determines appropriate resource allocation to each tier of an application.

III. Methodology

This section describes the architecture of the system and the environment used for testing, including the testing tools, hardware architecture, test cases and implementation procedure.

A. TMS Architecture

As a testbed-application we use our TMS SaaS solution [24] as a scalable and elastic multi-VM cloud solution. This benchmark has all features to evaluate performances of a typical transactional based SaaS solution, such as, simple operations for each transaction, and variable and huge number of users that generate different workload.

The benchmark TMS solution consists of three code modules and one optional module [24]. Core modules are: System management module (SMM), Company management module (CMM) and Core functions module (CFM) and optional is Additional functional module (AFM).

The core of this TMS cloud solution is the SMM module, shared for all companies. All resource provisioning in the cloud are managed by this module. It is always active to provide company subscription management, authentication, authorisation etc.

CMM is a company specific module, personalised and customised by TMS users. Each company offers this module as a service to its end customers. This module is responsible for all general configurations within a company.

Another company specific module is the CFM module, responsible for all essential processes to realise ticket management (bug reporting) within a company. It enables a connection between companies (providers) and their customers. This module provides functionalities such as: ticket creation, responds by a resolver, customer approvals of the responses, creation and execution of test-cases etc.

The only optional module is the AFM module, dedicated for add-on functions, such as enabling a possibility to create and execute test cases, especially useful for bug reporting systems.

All modules in this TMS solution are divided in two groups: static and dynamic modules. Static modules are always active modules, unlike dynamic modules that are activated only when some company needs them. From the modules we mentioned above SMM is a static module and the other three are dynamic modules.

This system integrates a very important Broker module in order to communicate with different company specific modules and acts as a role of system service orchestrator. Fig. 1 [24] presents three agents that are a part of the Broker module.

Let us explain shortly how this system organisation works. As we mentioned before SMM is the main module of the system and it is always active to provide features for "pay-by-use" cloud concept, including authentication, authorisation, accounting with management of the company, its' contracts and subscriptions etc. This module is realised by two agents:

- *Admin agent*, which provides the accounting features for the company.
- *Infrastructure agent*, which instantiates and closes various instances, providing the optimal resource utilisation and reducing the overall cost for renting the hardware resources.

The dynamic modules of this system have the opportunity to be hosted on the same machine or on different virtual machines, which is essential for our testing of the elasticity, because we want to analyse and compare the performances obtained in a single-VM and multi-VM environments.

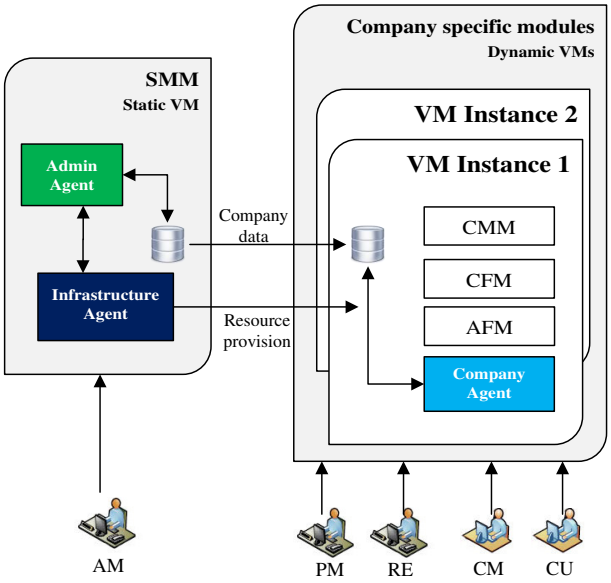


Fig. 1: Architecture of benchmark TMS SaaS solution

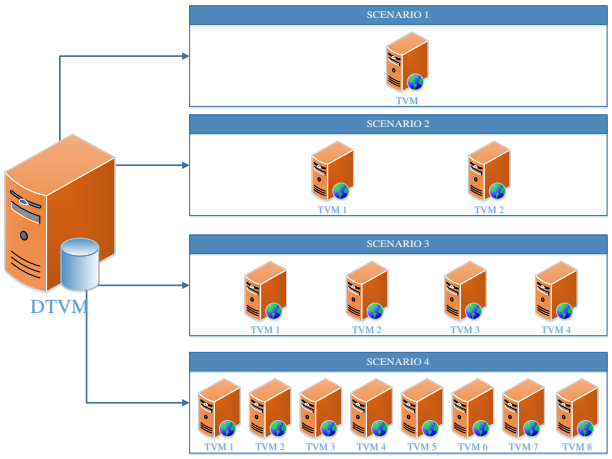


Fig. 2: Architecture of testing environment and scenarios.

TABLE I: Test case specification

Scenario	VM Type	Cores	RAM	Resources
All	DTVM	8	14 GB	8 Cores 14 GB
Scenario 1	1 TVM	8	14 GB	8 Cores 14 GB
Scenario 2	2 TVMs	4	7 GB	8 Cores 14 GB
Scenario 3	4 TVMs	2	3.5 GB	8 Cores 14 GB
Scenario 4	8 TVMs	1	1.75 GB	8 Cores 14 GB

Each dynamic VM hosts two core software modules: the CMM and CFM modules, independently for each company. Additionally dynamic part of the AFM module is hosted on each dynamic VM.

The Company agent is hosted on the same VM, to enable a communication with SMM and data synchronisation between the SMM and VM.

B. Testing environment

The testing environment consists of VMs and Windows Azure Cloud [4]. Windows Azure provides on-demand infrastructure that scales and adapts to the user ever changing business needs. With Windows Azure, the user can spin up new Windows Server and Linux VMs relatively fast and customise the environment. To create fast robust deployment, Windows Azure allows using custom images or building on the pre-configured images.

To realise the experiments we have created architecture for four different scenarios that use one *Database and Testing VM* (DTVM) and a set of *transactional VMs* (TVMs). All of them are deployed on Windows Azure as presented on Fig. 2.

The testing machine is used to run the tests for all experiments. The DTVM consists of AMD Opteron 4171 HE 2.10 GHz CPU with 8 cores and 14GB RAM. It hosts the SQL database for all scenarios. Windows Server 2012 Datacenter is running on the DTVM. The installed testing tools are selected from the Visual Studio 2012. The experiment results are measured with the *Web Performance and Load Test Project*, as a powerful tool for this kind of research.

The main function of the DTVM is to run all the tests for predefined scenarios. The first scenario is predefined as a single-VM, while the rest of them concern the multi-VM approach. The TVMs are configured to have equal cost in each scenario using resources on a machine with AMD Opteron 4171 HE 2.10 GHz CPU with 8 cores and 14GB RAM. Therefore, the first scenario is defined by one TVM that uses 8 CPU cores and has 14 GB of RAM. Two TVMs are used in the second scenario, each of them with 4 CPU cores and 7 GB of RAM. The third scenario uses 4 TVMs with 2 CPU cores and 3.5 GB of RAM, while the fourth scenario 8 TVMs with 1 CPU core and 1.75 GB of RAM.

All scenarios use VMs with same type of CPU and RAM. Windows Server 2012 Datacenter is preinstalled on TVMs including IIS and Application roles. Detailed features of the DTVM and TVMs in all scenarios are shown in Table I.

Experiments are tested by simulating two types of network: Cable DSL 1.5Mbps and Cable DSL 768Mbps. All experiments are realised with the use of a web browsers. To enable platform independence we have tested the experiments by equally using all of the following web browsers: Internet Explorer versions 9 and 10, Firefox, Chrome and Safari.

C. Web Service Description

TMS [25] is a cloud solution realised as a transactional web based software, where a user can access a dynamically created web page, browse the site, list page details and update a selected information, as most of typical web solutions.

The objective of this research is to find the best architecture when analysing the performance for the same price of rented resources. Our hypothesis is based on an assumption that multi-VM approach would be better and the best performance will be obtained renting the largest number of small VMs

for different companies. Note that the database is deployed on a separate VM in all experiments and the testing mostly relates to selecting a proper configuration for a web server that realises functions of web browsing and database update.

We define two test cases to analyse the performance. The first test case is based only on a web service providing a functionality of browsing through the application and the second uses update functionality with data store in database.

In the first test case, the user logs into the system, opens a page with presented tickets and selects a ticket to display relevant information. The user can browse the site and list the content of a given page. In the second test case, despite the previous steps realised for browsing and listing the required information, the user responds on the information by entering an additional information and changing the ticket status. It finishes with database update.

These web services are hosted by all VMs defined by each scenario. Note that the DTVM does not deploy these web services, but hosts the testing tools and the database required by the web services. The 4 scenarios realise both the single-VM and multi-VM approaches using the same amount of available resources. Hosting the web service and database on separate VMs may reduce the overall performance, because of communication needed between the VMs. But in case of large user loads all requests for a web service and interaction with the database will be distributed between VMs. Database operations are more expensive than the communication in Windows Azure, where all VMs are connected in virtual network with stable network connection between them.

D. Testing procedure

Visual Studio 2012 Web Performance and Load Test Project is a tool that offers many testing options. We decided to use the test "Tests based of the number of virtual users" for the purpose of this research providing relevant information about throughput and achieved speed.

We choose Visual studio because this tool provides a lot of possibilities for parametrisation. With Visual Studio 2012 we can choose different types of tests, depending of the number of the users or number of tests. Also we can choose the type of networks, type of browsers that we use for testing. Tests can be executed on different ways depending of number of tests or time span. A very important issue is that the results from tests are specified with sufficient details.

Visual studio 2013 provides a new option to run tests in Windows Azure without new configuration. Administrators do not have to deploy any VM or to configure different services. Through Load Test Web Service, Visual Studio 2013 loads the tests on the cloud. Behind this service is a pool of test agents that is used to run the tests. All results from a test, along with other performance counters are available.

Four experiments are defined, each for a test case defined by a corresponding scenario:

- *Experiment 1* - a single-VM with a DTVM and 1 TVM
- *Experiment 2* - a multi-VM with a DTVM and 2 TVMs
- *Experiment 3* - a multi-VM with a DTVM and 4 TVMs
- *Experiment 4* - a multi-VM with a DTVM and 8 TVMs

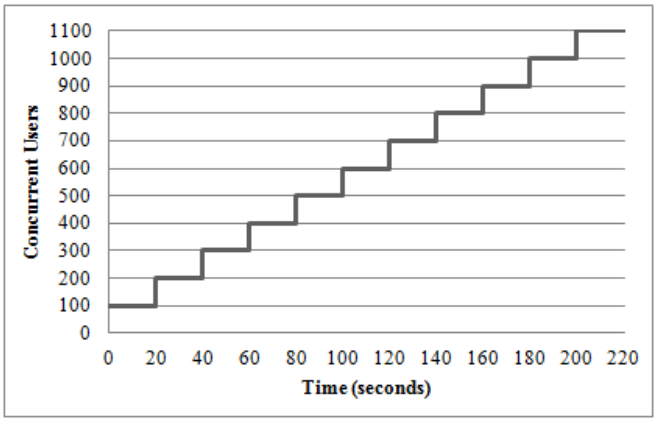


Fig. 3: Test plan for each experiment

All experiments use two VMs with 8 cores and 14GB RAM each. The first experiment simulates a single-VM environment, while the remaining a multi-VM environment with different configuration of TVMs enabled by renting the same cloud resources. Experiments perform different tests based on different loads defined by a various number of virtual users of the system per test.

The selected test runs every test for 3 minutes and 40 seconds. Certain number of users is defined per each testing interval to simulate increased load, as shown in Fig. 3. Our workload generation for multi-tier web application in the cloud is similar to other similar approaches [26]. The initial load configuration starts from 100 concurrent users and this is increased by a step of 100 users every 20 seconds, ending with a maximum of 1100 users.

All test case scenarios are executed for each experiment. To ensure good results, the tests started only after a 2 minutes warm up period of Visual Studio.

E. Test data

The selected testing tool measures and calculates a lot of parameters. For the purpose of our research, the following five different parameters are analyzed:

- T_r - Average response time,
- P - Pages per second,
- T_p - Average page time,
- R - Requests per second, and
- CPU_U - CPU utilisation.

The values of each of these criteria are based on the selected load defined by the number of users.

The performance factors for measured times are calculated as reciprocal value and indicated as throughput and speed, by the following measures:

- T - Overall throughput calculated as $T = 1/T_r$,
- S - CPU speed calculated as $S = 1/CPU_U$.

Each experiment (for a different scenario) results with different performance factors. Let $i, j \in \{1, 2, 3, 4\}$ identify

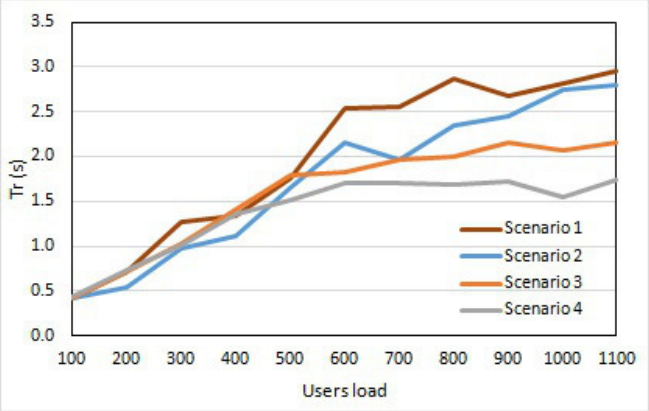


Fig. 4: Average response time as a function of user load

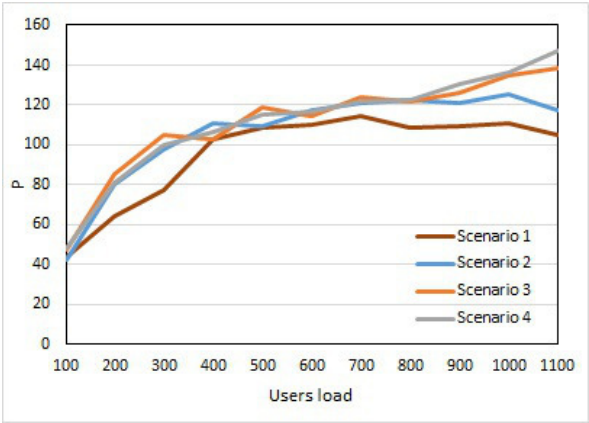


Fig. 5: Pages per second as a function of user load

two experiments. Their relative index as presented in (1) gives the relative comparison value for performance factor F, which stands instead of T, P, R, or S, corresponding to overall throughput, page throughput, request throughput and achieved processor speed. The higher the relative performance factor is, the better performance.

$$F_{ij} = \frac{F_i - F_j}{F_j} \tag{1}$$

In addition, when compared to Experiment 1 we obtain relative performance increase of multi-VM vs single-VM approach for the same number of users.

IV. Experiments and Results

This section presents the results achieved for each performance criteria: average response time, pages per second, average page time, handled requests per second and CPU utilisation. Parameters are analysed as a function of user load.

The brown line in each graph presents the results for scenario 1, the blue line scenario 2, the orange line scenario 3 and the grey line scenario 4.

A. Average Response Time T_r

Response Time refers to the average time that is necessary to receive the entire response to a request, starting from the moment when a request is sent to the web server. The difference between the Response Time and Transaction Time is in think time that occurs during a transaction. Transaction Time includes think time, but Response Time does not [27].

Fig. 4 presents the average response time results for all experiments. The average response time proportionally depends on the load increase defined by the number of concurrent users. Two different regions are observed with different behaviors. In the left region ($N \leq 500$) all four scenarios provide similar average response times. However, Scenario 1 is the worst, while Scenario 2 is slightly better than others for all test cases in the region. Scenario 1 also provides the worst performance in the right region ($N > 500$), but now Scenario 4 has the smallest average response time. All curves in the left region have a trend for linear increment until the value $1.5s$, which saturates in the right region in the range of $[1.5s, 3s]$.

B. Average pages per second P

Average pages per second refers to the number of pages that are sent per second during the load test run [27]. Fig. 5 displays the results for the performance factor P defined by achieved average pages per second.

Single-VM (Scenario 1) also shows the lowest performance for this parameter for each user load. Multi-VM scenarios provide similar performance for smaller load ($N \leq 500$), but the results are similar as the previous parameter for greater load. That is, using greater number of smaller TVMs is better than having smaller number of greater TVMs.

All curves in the left region also have a trend for linear increment starting from 40 to 110, which saturates in the right region in the range of $[100, 150]$.

C. Average page time T_p

Average page time refers to the average response time for all requests for a single web page [27]. The results of the average page time are shown in Fig. 6. Also in this case, the results show that the multi-VM approaches perform better than the single-VM scenario. Scenario 4 performs the best for a large number of concurrent users higher than 700, while Scenario 2 for smaller loads.

The curves follow the similar trend as the previous two parameters, in the regions divided with $N = 500$ concurrent requests. The average page time increases starting from 0.5 to 2.5s and saturates in the range [2.0s, 3.5s] for all scenarios.

D. Handled requests per second R

Requests per second consist of details for individual requests that are sent during a load test. This includes all HTTP and dependent requests such as images. *Request per second* refers to the rate per second of a request during the load test run [27].

Results for the overall handled requests per second are presented in Fig. 7. The obtained behavior and performance

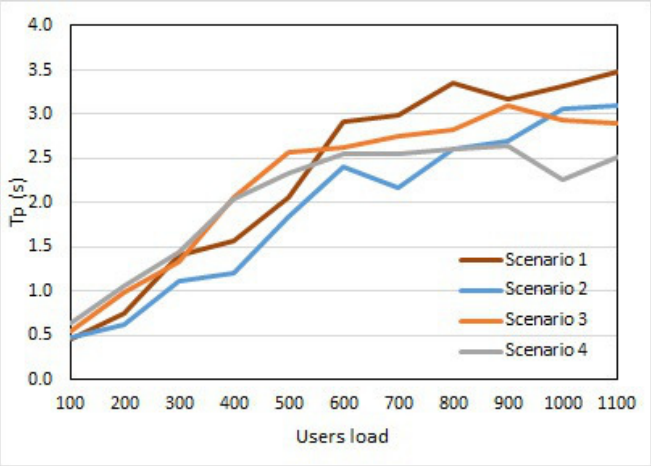


Fig. 6: Average page time as a function of user load

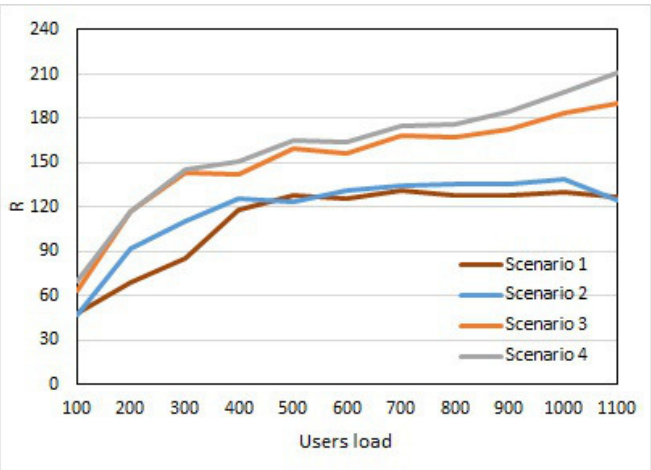


Fig. 7: Requests per second as a function of user load

trend is similar to the cases with previous analysed parameters. That is, the scenarios with greater number of smaller TVMs (scenarios 4 and 3) are better than having smaller number of greater TVMs (scenarios 2 and 1).

We observe small discrepancy compared to other parameters. That is, the curves start to saturate for different user load. Scenarios 1 and 2 increase the values for the handled requests starting from 50 to 120, while scenarios 3 and 4 provide increase from 70 to 140.

Scenarios 3 and 4 slightly saturates for $N > 300$ (they continue to increase, but with smaller intensity) in the range $[140, 210]$, while scenarios 1 and 2 saturate for $N > 400$ and their curves are almost constant in the right region around the range $[120, 140]$.

E. CPU Utilization R

Processor time denotes the percentage of time that a VM instance charges against the processor user time for individual request issued during a load test. The theoretical maximum for this parameter is $number\_of\_processors \cdot 100$.

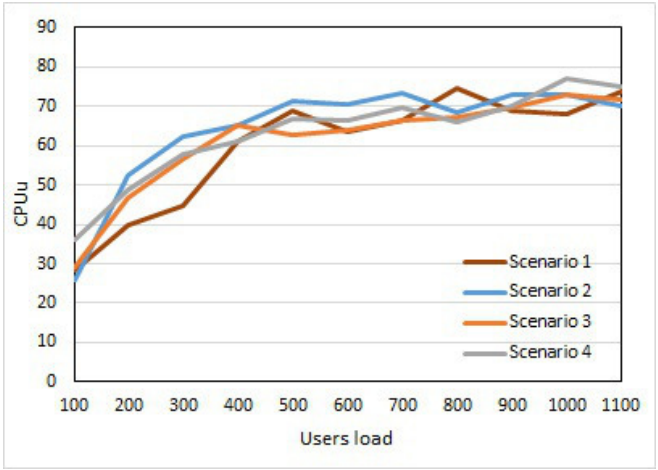


Fig. 8: CPU utilization as a function of user load

CPU utilisation is very important parameter for cloud service provider because it is directly connected with cost for power supply and cooling due to increased heating. The results for this parameter are shown in Fig. 8. Although all scenarios are very similar, lower CPU utilisation is noticed for the single-VM environment, while Scenario 2 mostly utilizes the CPU.

Similar two regions are observed, which are divided with $N = 400$. The curves increase in the left region until 65%, and then the CPU utilization saturates in the range of $[65\%, 70\%]$.

V. Discussion

This section compares the results achieved from the four cloud approaches and discussed relevant explanations for analyzed behavior and performance trends.

A. Parameter correlations

This section presents the correlations between the response time and page throughput and between the response time and CPU utilization.

1) Response time vs page throughput: The parameters T_r, P, T_p, and T_R are correlated among each other, as shown in figures 4, 5, 6, and 7. The results show that user loads for $N = 400$ or $N = 500$ are borders of two different regions. The parameters behave the same in a single region, i.e., increasing trend in the left region (smaller load) and saturating trend in the right region (heavy load).

2) Response time vs CPU utilization: Figures 9 and 10 present the overall throughput and CPU speed. We can conclude that both figures are very similar, which confirms that both parameters are very correlated as a function of user load.

B. Scenario comparisons

In most cases, especially for increased loads, we concluded that Scenario 4 performs the best compared to the others.

The load defined by the number of concurrent users has noticeable impact on the results. All performance factors lead

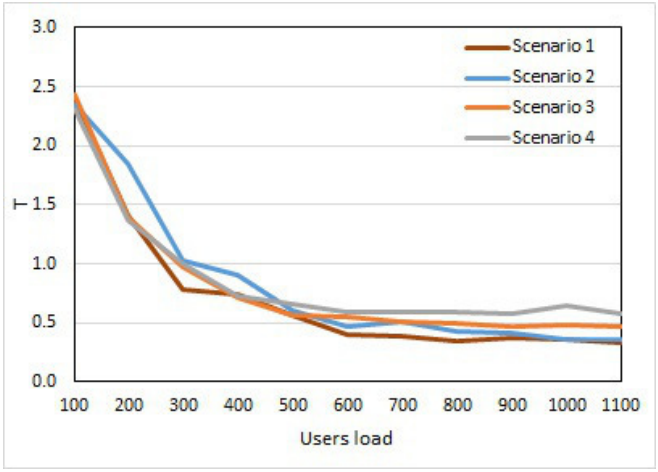


Fig. 9: Overall throughput T as a function of user load

Fig. 11: Relative throughput T_{12}, T_{13} and T_{14}.

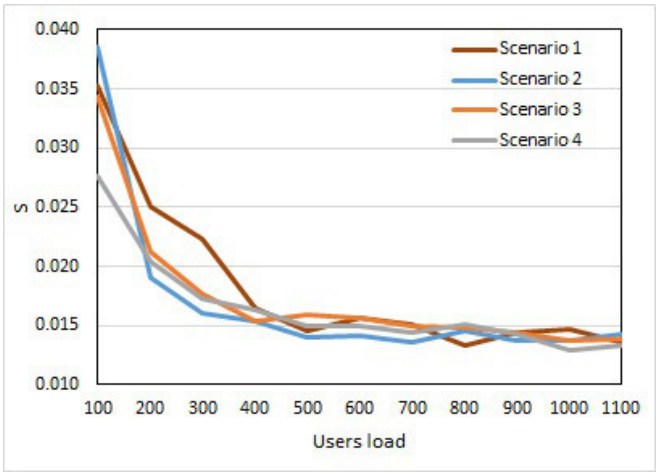


Fig. 10: Achieved CPU speed S as a function of user load

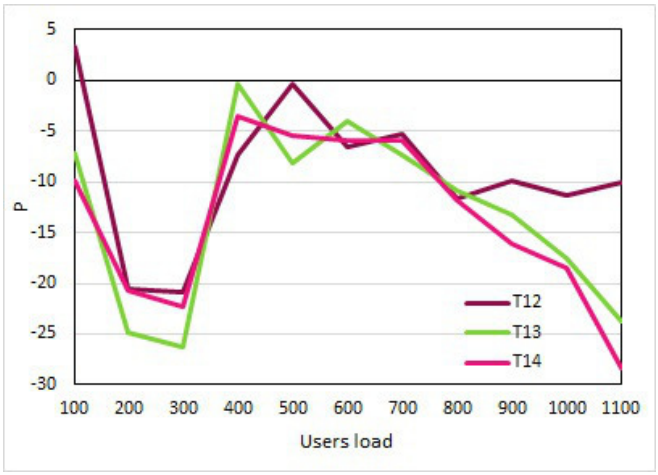


Fig. 12: Relative pages per second R_{12}, R_{13} and R_{14}.

to a conclusion that our hypothesis is proved for increased loads higher than 400 users.

Figures 11, 12 and 13 present the relative values of parameters for the three multi-VM scenarios compared to the single-VM scenario. They confirm the previous conclusions that the scenarios with greater number of smaller VMs is better than smaller number of bigger VMs as TVM.

VI. Conclusion

Performance is critical when choosing appropriate configuration for rented cloud resources. Cloud providers offer a variety of options, so the users get confused when selecting the optimal configuration. This research brings conclusion about a typical behavior in transactional web solutions for expected small and medium number of concurrent users. The ticket management solution is a transactional-based dynamic web site where a lot of information fluctuates among users, and information update is a frequent task. It uses a lot of users which access a dynamic web site.

Fig. 13: Relative page time between multi- and single-VM

The presented architecture for the cloud solution is based on separating the database layer form the business logic and presentation layers on different VMs. This solution uses a database allocated to a single and powerful VM, and distributing the transactional business logic and presentation layer to a set of other VMs. An intended user may choose whether to host a new powerful VM or to distribute the load to more less powerful VMs using the same cloud resources.

We have performed load testing with increasing number of users, testing the performance of 4 different configurations that use the same rented resources. We were motivated by the challenge to find the most optimal configuration that performs the best. This research proves our hypothesis that using a higher number of small VMs performs better than other configurations, confirming the traditional distributed concept to distribute tasks on more balanced processing units.

This result suggests an architecture for cloud based solutions that performs the best for transactional web sites, which do not perform complex computations and are mainly based on page browsing and database update operations. The recommendation is to separate the database layer from other layers, allocating the database layer to a powerful VM and distributing the business logic and presentation layers to more less powerful VMs.

REFERENCES

[1] A. Murua, I. Gonzalez, and E. Gomez-Martinez, "Cloud-based assistive technology services," in *Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on*, Sept 2011, pp. 985–989.

[2] L. Toomey. 5 cloud computing statistics you may find surprising. Cloudspectator. [Online]. Available: http://cloudcomputingtopics.com/2011/11/5-cloud-computing-statistics-you-may-find-surprising/

[3] C. Binnig, D. Kossmann, T. Kraska, and S. Loesing, "How is the weather tomorrow?: Towards a benchmark for the cloud," in *Proceedings of the Second International Workshop on Testing Database Systems*, ser. DBTest '09. ACM, 2009. doi: 10.1145/1594156.1594168 pp. 9:1–9:6. [Online]. Available: http://doi.acm.org/10.1145/1594156.1594168

[4] Microsoft. Windows azure. Microsoft. [Online]. Available: http://www.windowsazure.com

[5] Google, "Compute Engine," 2013. [Online]. Available: http://cloud.google.com/pricing/

[6] Amazon, "EC2," 2013. [Online]. Available: http://aws.amazon.com/ec2/

[7] M. Gusev and S. Ristov, "The optimal resource allocation among virtual machines in cloud computing," in *Proceedings of The 3rd International Conference on Cloud Computing, GRIDs, and Virtualization (CLOUD COMPUTING 2012)*, 2012, pp. 36–42.

[8] M. Gusev, S. Ristov, G. Velkoski, and M. Simjanoska, "Optimal resource allocation to host web services in cloud," in *Proceedings of the 2013 IEEE Sixth International Conference on Cloud Computing*, ser. CLOUD '13, USA, June 2013. doi: 10.1109/CLOUD.2013.103 pp. 948–949. [Online]. Available: http://dx.doi.org/10.1109/CLOUD.2013.103

[9] W. Wang, X. Huang, X. Qin, W. Zhang, J. Wei, and H. Zhong, "Application-level cpu consumption estimation: Towards performance isolation of multi-tenancy web applications," in *Cloud Computing (CLOUD), 2012 IEEE 5th Int. Conf. on*, 2012. doi: 10.1109/CLOUD.2012.81 pp. 439–446. [Online]. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6253536

[10] Y. Koh, R. Knauerhase, P. Brett, M. Bowman, Z. Wen, and C. Pu, "An Analysis of Performance Interference Effects in Virtual Environments," in *Performance Analysis of Systems Software, 2007. ISPASS 2007. IEEE International Symposium on*, april 2007, pp. 200 –209.

[11] D. Jayasinghe, S. Malkowski, Q. Wang, J. Li, P. Xiong, and C. Pu, "Variations in performance and scalability when migrating n-tier applications to different clouds," in *Cloud Computing (CLOUD), 2011 IEEE Int. Conf. on*, 2011. doi: 10.1109/CLOUD.2011.43 pp. 73–80. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6008695&isnumber=6008659

[12] R. Iakymchuk, J. Napper, and P. Bientinesi, "Improving high-performance computations on clouds through resource underutilization," in *Proceedings of the 2011 ACM Symposium on Applied Computing*, ser. SAC '11. ACM, 2011. doi: 10.1145/1982185.1982217 pp. 119–126. [Online]. Available: http://doi.acm.org/10.1145/1982185.1982217

[13] E. Roloff, F. Birck, M. Diener, A. Carissimi, and P. Navaux, "Evaluating high performance computing on the windows azure platform," in *Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on*, June 2012. doi: 10.1109/CLOUD.2012.47. ISSN 2159-6182 pp. 803–810. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6253582&isnumber=6253471

[14] V. Subramanian, H. Ma, L. Wang, E.-J. Lee, and P. Chen, "Rapid 3D Seismic Source Inversion Using Windows Azure and Amazon EC2," in *Proceedings of IEEE*, ser. SERVICES '11, 2011, pp. 602–606.

[15] Z. Hill, J. Li, M. Mao, A. Ruiz-Alvarez, and M. Humphrey, "Early observations on the performance of windows azure," in *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, ser. HPDC '10. ACM, 2010. doi: 10.1145/1851476.1851532. ISBN 978-1-60558-942-8 pp. 367–376. [Online]. Available: http://doi.acm.org/10.1145/1851476.1851532

[16] R. Tudoran, A. Costan, G. Antoniu, and L. Bougé, "A performance evaluation of Azure and Nimbus clouds for scientific applications," in *Proc. of the 2nd Int. Workshop on Cloud Computing Platforms*, ser. CloudCP '12. ACM, 2012. doi: 10.1145/2168697.2168701 pp. 4:1–4:6. [Online]. Available: http://doi.acm.org/10.1145/2168697.2168701

[17] A. Turner, A. Fox, J. Payne, and H. Kim, "C-mart: Benchmarking the cloud," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 24, no. 6, pp. 1256–1266, June 2013. doi: 10.1109/TPDS.2012.335. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6381404&isnumber=6507529

[18] "Rubis," 2014. [Online]. Available: http://rubis.ow2.org/

[19] E. Cecchet, V. Udayabhanu, T. Wood, and P. Shenoy, "Benchlab: An open testbed for realistic benchmarking of web applications," in *Proc. of the 2Nd USENIX Conf. on Web Application Development*, ser. WebApps'11, 2011, pp. 4–4. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002168.2002172

[20] B. Pugh and J. Spacco, "Rubis revisited: Why j2ee benchmarking is hard," in *Companion to the 19th Annual ACM SIGPLAN Conf. on Object-oriented Programming Systems, Languages, and Applications (OOPSLA '04)*. ACM, 2004. doi: 10.1145/1028664.1028751 pp. 204–205. [Online]. Available: http://doi.acm.org/10.1145/1028664.1028751

[21] "Tpc benchmarkw(web commerce) specification," 2002. [Online]. Available: http://www.tpc.org/tpcw/spec/tpcw_v1.8.pdf

[22] Q. Wang, Y. Kanemasa, J. Li, D. Jayasinghe, T. Shimizu, M. Matsubara, M. Kawaba, and C. Pu, "Detecting transient bottlenecks in n-tier applications through fine-grained analysis," in *Distributed Computing Systems (ICDCS), 2013 IEEE 33rd International Conference on*, July 2013. doi: 10.1109/ICDCS.2013.17 pp. 31–40. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6681573&isnumber=6681559

[23] X. Chen, H. Chen, Q. Zheng, W. Wang, and G. Liu, "Characterizing web application performance for maximizing service providers' profits in clouds," in *Cloud and Service Computing (CSC), 2011 International Conference on*, Dec 2011, pp. 191–198.

[24] M. Gusev, S. Ristov, and P. Gushev, "Developing a ticket management saas solution," in *MIPRO, 2014 Proceedings of the 37th International Convention, IEEE Conference Publications*, Croatia, 2014, pp. 328–333.

[25] P. Gushev, A. Guseva, S. Ristov, and M. Gusev, "Cloud solutions for bug reporting," in *XLVIII Int. Scientific Conf. on Information, Comm. and Energy Systems and Technologies*, 2013, pp. 227–230.

[26] W. Iqbal, M. Dailey, and D. Carrera, "Sla-driven dynamic resource management for multi-tier web applications in a cloud," in *Cluster, Cloud and Grid Computing (CCGrid), 2010 10th IEEE/ACM International Conference on*, May 2010. doi: 10.1109/CCGRID.2010.59 pp. 832–837. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5493374&isnumber=5493340

[27] Microsoft. Analyzing load test results and errors in the tables view of the load test analyzer. Microsoft. [Online]. Available: http://msdn.microsoft.com/en-us/library/ms404656.aspx#analyzingloadtestresultserrorstablesviewtherequeststable

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 761–768
DOI: 10.15439/2014F311
ACSIS, Vol. 2

Creating portable TOSCA archive for iKnow University Management System

Magdalena Kostoska, Ivan Chorbev and Marjan Gusev
University Ss. Cyril and Methodius, Faculty of Computer Sciences and Engineering,
1000 Skopje, Macedonia
{magdalena.kostoska, ivan.chorbev, marjan.gushev}@finki.ukim.mk

***Abstract*—Portability of developed solutions is a huge challenge for both the customers and cloud providers. A typical customer would like to realize a flexible software solution which can be easily deployed on a selected infrastructure or transferred on a cloud environment. Finally, the idea of a software environment independent of a specific cloud provider enables preferable conditions for customers.**

These portability issues are analyzed for a case study of a custom University Management System - iKnow, used by the University Ss. Cyril and Methodius in Macedonia. This paper presents implementation of a Topology and Orchestration Specification for Cloud Applications (TOSCA) specification to enable a software solution to be deployed on a flexible and portable environment.

The motivation to consider the implementation of such a platform is initiated due to the rising number of students that use the iKnow system, reaching the peak during the enrollment periods. A possible solution is to consider the possibility of moving the iKnow system (or its part) to a cloud environment and balance the workload using flexible resources. We conducted research to describe the iKnow system using the TOSCA specification and to enable a rather easy deployment and maintenance.

This paper presents the challenges and issues regarding the TOSCA specification while implemented on the iKnow system.

I. Introduction

IKNOW is the University Management Information system that provides electronic services for both university management and students [1]. It is deployed in most of the universities in Macedonia, including the University Ss Cyril and Methodius, where this research is conducted.

Our analysis shows peak usage of the iKnow system during the enrollment period. More than 7 000 applicants use it to register, fill applications, overview ranked lists, submit appeals and perform similar activities in the applying process. From the other side, more than 400 administration users within committees of 27 different faculties work on data checking, evaluation, ranking and similar processes. At the same time more than 25 000 enrolled students use the iKnow system for course enrollment and exam application. Along with the students, this system is also frequently used by the administrative staff of the faculties.

Heavy workload only in a short time period is a problem that motivates us to find an alternative solution to increase the bandwidth and processing power in the enrollment period, without buying additional hardware. Although, the natural solution for our needs is the cloud, there is one more demand, the University Management does not prefer ”lock in” to a specific cloud provider and would like a flexible platform that can be easily ported from one cloud provider to another or return to the existing University server.

A solution to these portability demands initiated our research on how to create easily portable, reusable and maintainable system specification, that we can use on any cloud provider, flexible enough to choose the provider offering the desired highest performances. We propose the usage of Topology and Orchestration Specification for Cloud Applications (TOSCA), a new standard for description of applications and services [2][3].

The main goal of the realized research was to explore the possibility of using TOSCA to describe the iKnow system and to address all the challenges of moving this system to the cloud, or changing the cloud provider. As a result, the research identifies TOSCA features and drawbacks of such an implementation.

The rest of the paper is organized as follows: Related work about cloud portability and TOSCA analysis is presented in Section II. Background about the TOSCA specification and the iKnow system (architecture and current deployment) is given in Sections III and IV respectively. The TOSCA model of the iKnow system is presented in Section V. Section VI describes the challenges and issues we have met during this research. Finally, in Section VII we evaluate the significance of each of the found challenges and we conclude with future work.

II. Related Work

The main problem in Cloud computing is the lack of unified standards [4]. Several standards are on the way of their development, approval of the community and wider adoption [5].

Latest research considering the portability of cloud services in each of the layers of the service stack, tackle the portability problem from three different perspectives. One perspective is the usage of standards to accomplish this task. In this direction they try to identify current challenges, efforts and options [6], [7] or to extend open standards for this purpose [8]. The second perspective searches for other proprietary solutions like the abstraction-driven approach where abstract languages are used to specify solutions [9] or to create a storage abstraction layer (CSAL)[10]. The third perspective consists of other approaches that focus on exploiting the semantic technology to create semantic interoperability framework [11],

978-83-60810-58-3/$25.00

to automatically analyze cloud vendor APIs [12] or to use mOSAIC for this intention [4], [13].

OASIS (Organization for the Advancement of Structured Information Standards) introduces a new cloud standard called TOSCA (Topology and Orchestration Specification for Cloud Applications). The TOSCA technical committee has published the first version of a standard for "*interoperable description of application and infrastructure cloud services, the relationships between parts of the service, and the operational behavior of these services (e.g., deploy, patch, shutdown)*" [2]. This initiated a spread of research activities in different directions, such as IBM efforts to implement this standard [14], [15], creation of environment for execution [16], [17], modeling tools [18], [19] and attempt to create this type of specification [20].

III. TOSCA

This section presents a brief overview of the language, the structure and the usage of TOSCA.

Topology and Orchestration Specification for Cloud Applications (TOSCA) is developed by OASIS, a non-profit, international consortium. The main goal of the OASIS TOSCA Technical Committee is to enhance the portability of cloud applications and services by enabling interoperable description of application and infrastructure cloud services, the relationships between parts of the service, and the operational behavior of these services [2].

This specification utilizes open standards to describe a service and explain how to manage it independently from the supplier and the cloud. The current release of this specification is Version 1.0 and defines services by using a Service Template document. The TOSCA language introduces a grammar for describing service templates by means of Topology Templates and plans. TOSCA utilizes XML Schema 1.0 and WSDL 1.1. It also contains non-normative references to BPEL 2.0, BPMN 2.0, OVF and XPATH 1.0.

The specification defines a meta model for defining the structure of an IT service and its management. The core of TOSCA is the Service Template, depicted in Fig. 1. It consists of two logical parts: service topology description and management plans. The Topology Template defines the structure of a service. Plans define the process models that are used to create, terminate and manage a service.

The structure of the services is defined by the Topology Template. This template consists of Node Templates and Relationship Templates. The Node Template describes the required components and their properties, operations and interfaces, while the Relationship Template describes connectivity between components (Node Templates) by stating the direction of the relationship (source and target) and can have additional parameters. The Node Type and Relationship Type templates formally describe nodes and relationships (in terms of properties, interfaces etc.), while the templates are created and set concrete values based on these types definitions. The purpose of Plans is to interpret the templates and execute appropriate actions. They are defined as process models and they rely on BPMN (primarily) and BPEL. In addition, this specification allows nesting (ex. One Service Template to be part of another).

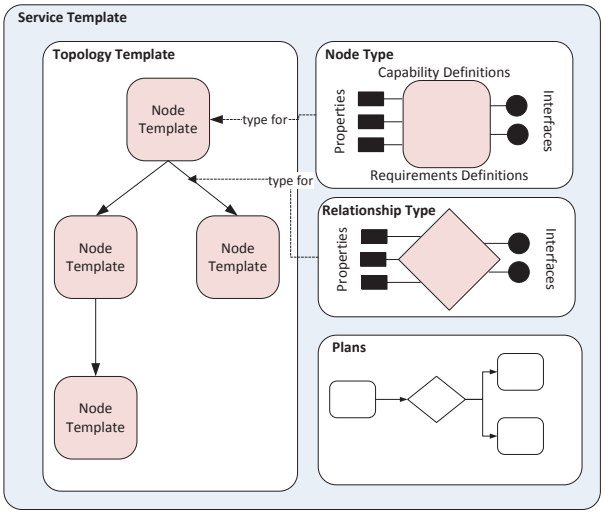


Fig. 1: TOSCA Service Template (adapted from [2])

In order to allow automatic deployment or build of services TOSCA specification also defines Artifacts' installable or executional types and objects like scripts, libraries, installation binaries, installation images and other objects.

IV. iKnow University Management System

The architecture and current deployment of the iKnow University Management System is presented in this section.

The main goal of the iKnow system is to provide electronic services and to store and exchange electronic information among students, professors, administration, university management and the Ministry of Education.

A. Architecture

The system consists of two main components, each consisting of modules: the new students Enrollment Student Services (ESS) and the Core Student Services (CSS) [1].

The Enrollment component consists of the faculty enrollment functionalities like enrollment wizard, manual entry of candidate's data and candidate's data processing, ranking and results publishing.

The Core student services component includes modules for administration, management of study programs and schedules, student activities, personal identification and access control, electronic payment and migration of legacy data.

The main architecture used to construct the system is NTier, depicted in Fig. 2. The two main components (Enrollment and Core) are almost independent of one another, except for several business processes (like transferring enrolled students into the core module). Although they use the same database server, physically they use two databases (enrollment and core

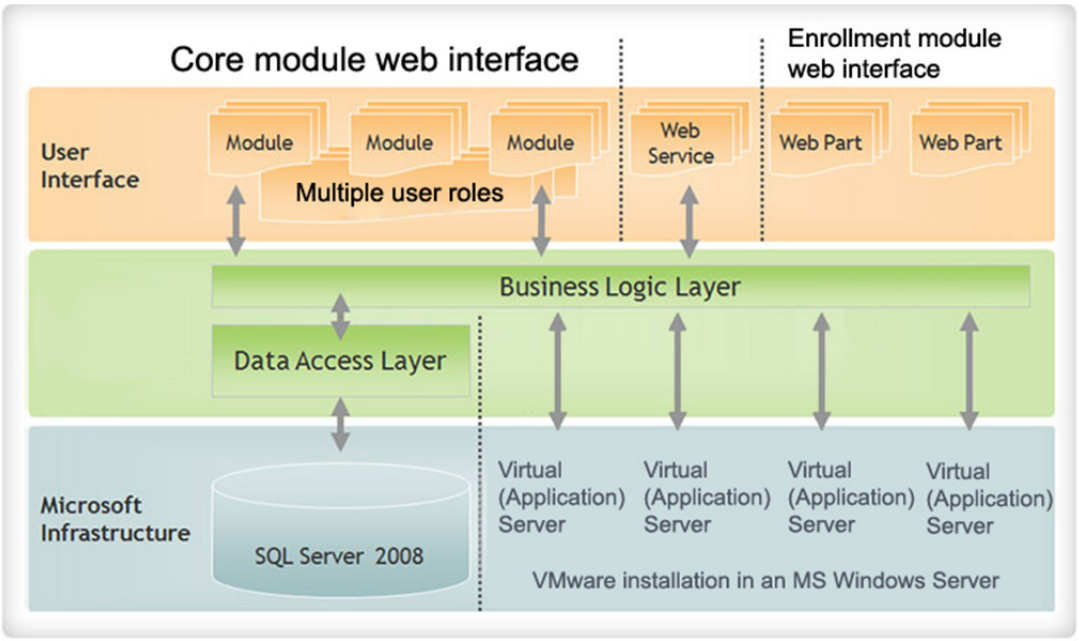


Fig. 2: iKnow architecture [1]

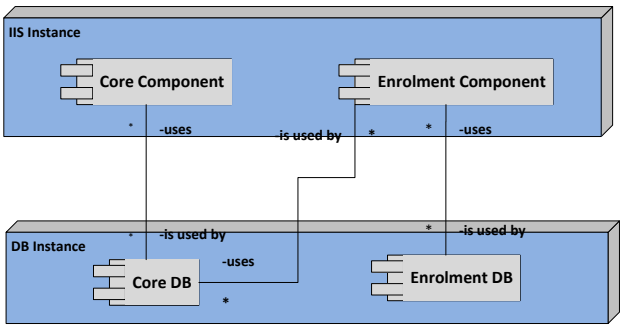


Fig. 3: Modules and database interactions

database). Fig. 3 shows the interactions among modules and databases.

The current solution consists of several project and both modules use some of them, like UniSS.DataModel, a Class library project containing the Entity framework model of the data as well as the audit log component. The UniSS.Repositories project is a Class library project storing the classes for implementation of the Repository template. The UniSS.Logic project contains the implementation of the MVP template, while UniSS.Forms is a Web Forms application project in which all forms are separated according to their functionalities.

The Model-View-Presenter (MVP) is used for all the modules [1].

B. *Current Deployment*

The current deployment includes two physical servers with implemented virtualization approach and an independent storage system. The first server is used as SQL server with Microsoft SQL Server installed and the second is used as an application server. The application server is upgraded by installing 4 virtual Windows servers using VMware. Additional Load Balancer balances the workload of each instance of the application. Both servers have an Intel Xeon e5630 2.56 GHz CPU and 16GB of RAM, 1 TB storage. The system is supported by additional backup internet link in case of failures of the primary link. In addition, a third fully synchronized server is installed on remote location and can take over in case of failure of the primary servers. All communications are encrypted using SSL protocol and a digital certificate.

V. iKnow TOSCA specification and artifacts

The TOSCA specification for the iKnow University Management System will be presented by specification of corresponding nodes, relationships, artifacts and plans.

A. *Nodes*

The first step is to identify the logical set of component services that the application is both composed of, as well as those services that the application relies upon for deployment, installation, execution and other lifecycle operations [21]. Each of these basic components, their interfaces and properties are described as TOSCA *Node Types*. TOSCA recommends creation of three node types: base, specific and custom. Fig. 4

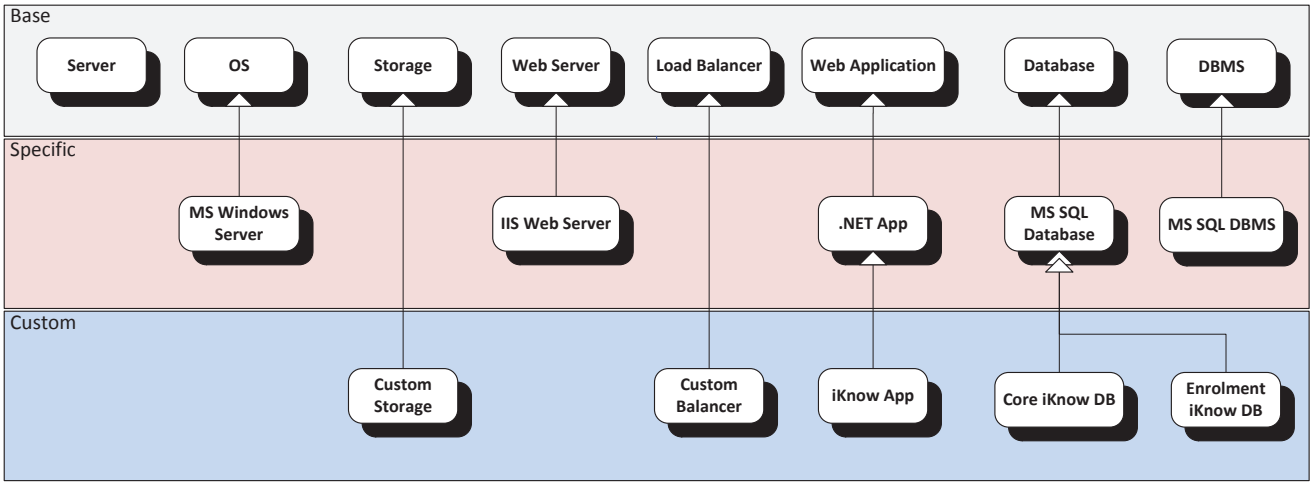


Fig. 4: Node Type Modeling and Node Type Inheritance Hierarchy for iKnow

illustrates all node types and inheritance hierarchy for the iKnow system.

The *Base Node Types* describe the basic set of components services needed for the proper functioning of the application. All Base Node Types are directly derived from a generic TOSCA root node type. Listing 1 shows part of the definition, due to the large specification. For the same reason the name spaces are removed.

```
<NodeType name="Server">
      <DerivedFrom typeRef="RootNodeType"/>
    <RequirementDefinitions>
      <RequirementDefinition lowerBound="0"
          name="container"
requirementType="ServerContainerRequirement"
upperBound="1"/>
    </RequirementDefinitions>
      ...
</NodeType>
<NodeType name="WebServer">
      <DerivedFrom typeRef="RootNodeType"/>
      <Operation name="install"/>
      <Operation name="configure"/>
      <Operation name="start"/>
      <Operation name="stop"/>
      <Operation name="uninstall"/>
      ...
</NodeType>
<NodeType name="LoadBalancer">
      <DerivedFrom typeRef="RootNodeType"/>
         <Operation name="install"/>
      <Operation name="configure"/>
      <Operation name="start"/>
      <Operation name="stop"/>
      <Operation name="uninstall"/>
      ...
</NodeType>
<NodeType name="WebApplication">
      <DerivedFrom typeRef="RootNodeType"/>
        <Interfaces>
     <Interface name="lifecycle">
      <Operation name="install"/>
      <Operation name="configure"/>
      <Operation name="uninstall"/>
     </Interface>
   </Interfaces>
      ...
</NodeType>
<NodeType name="Database">
      <DerivedFrom typeRef="RootNodeType"/>
      <Interfaces>
     <Interface name="lifecycle">
      <Operation name="install"/>
      <Operation name="uninstall"/>
      <Operation name="start"/>
      <Operation name="stop"/>
     </Interface>
   </Interfaces>
      ...
</NodeType>
<NodeType name="DBMS">
      <DerivedFrom typeRef="RootNodeType"/>
      <Interfaces>
     <Interface name="lifecycle">
      <Operation name="install"/>
      <Operation name="configure"/>
      <Operation name="start"/>
      <Operation name="stop"/>
      <Operation name="uninstall"/>
     </Interface>
   </Interfaces>
      ...
</NodeType>
```

Listing 1: Definitions of Basic Node Types

The next step is to extend the base types to represent specific instances (to define *Specific Node Types*). They are derived directly from the base types. In our use case, we define five

specific node types: MS Windows Server, IIS Web Server, .NET Application, MS SQL Database and MS SQL Database Management Service. Listing 2 shows the definition for IIS Web Server.

```
<NodeType name="IISWebServer">
   <DerivedFrom typeRef="WebServer"/>
   <Interfaces>
     <Interface name="lifecycle">
       <Operation name="install"/>
         <InputParameters>
              <InputParameter name="httpport"
                  type="integer"/>
              <InputParameter name="username"
                  type="string"/>
              <InputParameter name="password"
                  type="string"/>
       ...
     </Interface>
   </Interfaces>
</NodeType>
```

Listing 2: Definition of IIS Web Server Specific Node Type

Furthermore we can extend the specific node types to *Custom Node Types*, which are the custom build elements of the solution. In the iKnow system use case we define five custom node types for the storage, the balancer, the iKnow application as well as the two databases (Core and Enrolment). Listing 3 shows the definition for the Core database.

```
<NodeType name="CoreiKnowDB">
   <DerivedFrom typeRef="MSSQLDB"/>
   <Interfaces>
     <Interface name="lifecycle">
       <Operation name="install"/>
       <Operation name="start"/>
       <Operation name="stop"/>
       <Operation name="uninstall"/>
       <Operation name="backup"/>
     </Interface>
   </Interfaces>
</NodeType>
```

Listing 3: Definitions of Core DB custom node types

The final step in the node definition is to create *Node Templates*, based on the defined node types, which can be instantiated with specific properties. Node types only describe properties, requirements and capabilities, while Node Templates provide specific property settings to be applied when the specification is deployed on the cloud provider. Listing 4 shows the definition for the IIS Web Server Node Template.

B. Relationships

TOSCA uses *Relationship Types* to describe how these TOSCA node relate to each other in the cloud deployment. The same as node, the relationship type can be base, specific and custom. Some of the relationship types are shown in Listing 5.

Similar to Node Templates, TOSCA *Relationship Templates* are describing the logical relationships and other dependencies between the application's node templates and are needed for deployment. It describes the source and target Node Templates. Listing 6 describes the relationship between iKnow Core Database and MS SQL by using the Requirement and Capability properties defined in this node templates.

```
<NodeTemplate id="IISWebServer"
      name="IIS Web Server"
      type="IISWebServer">
   <Properties>
      <httpport>80</httpport>
      <username>sa</username>
      <password>sa</password>
   </Properties>
</NodeTemplate>
```

Listing 4: Definitions of IIS Web Server Node Template

```
<RelationshipType name="DeployedOn">
   <DerivedFrom
       typeRef="RootRelationshipType"/>
   <ValidSource typeRef="WebApplication"/>
   <ValidTarget typeRef="WebServer"/>
</RelationshipType>

<RelationshipType name="ConnectsTo">
   <DerivedFrom
       typeRef="RootRelationshipType"/>
   <ValidSource typeRef="WebApplication"/>
   <ValidTarget typeRef="Database"/>
</RelationshipType>

<RelationshipType name="HostedOn">
   <DerivedFrom
       typeRef="RootRelationshipType"/>
   <ValidSource typeRef="Database"/>
   <ValidTarget typeRef="DBMS"/>
</RelationshipType>

<RelationshipType
    name="CoreDBHostedOnMSSQLDBMS">
   <DerivedFrom typeRef="HostedOn"/>
   <SourceInterfaces>
     <Interface name="HostedOn">
       <Operation name="hostOn"/>
     </Interface>
   </SourceInterfaces>
   <ValidSource typeRef="CoreiKnowDB"/>
   <ValidTarget typeRef="MSSQLDBMS"/>
</RelationshipType>

<RelationshipType
    name="iKnowAppDeployedOnIISWebServer">
   <DerivedFrom typeRef="DeployedOn"/>
   <SourceInterfaces>
     <Interface name="DeployedOn">
       <Operation name="deployOn"/>
     </Interface>
   </SourceInterfaces>
   <ValidSource typeRef="iKnowApp"/>
   <ValidTarget typeRef="IISWebServer"/>
</RelationshipType>
```

Listing 5: Sample of iKnow Relationship Types

```
<RelationshipTemplate
    id="CoreDB_HostedOn_MSSQLDBMS"
    name="hosted on"
        type="CoreDBHostedOnMSSQLDBMS">
    <SourceElement ref="CoreDB_container"/>
    <TargetElement ref="MSSQLDBMS_databases"/>
</RelationshipTemplate>

<NodeTemplate id="CoreDatabase"
      name="iKnowCoreDB"
      type="CoreDB">
   <Properties>... </Properties>
   <Requirements>
    <Requirement id="CoreDB_container"
        name="container"
              type="MSSQLDBContainerReq"/>
   </Requirements>
   <Capabilities>...</Capabilities>
</NodeTemplate>

<NodeTemplate id="MSSql" name="MS_SQL"
    type="MySQL">
   <Properties>... </Properties>
   <Requirements>... </Requirements>
   <Capabilities>
    <Capability id="MSSql_databases"
        name="databases"
             type="MSSQLDBContainerCap"/>
   </Capabilities>
  </NodeTemplate>
```

Listing 6: Sample Relationship Template for Node Templates

C. Artifacts

Artifacts are needed in order to deploy and install the cloud application. The may vary from scripts, files to packages and virtual images.

```
<ArtifactType name="FileArtifact">
   <DerivedFrom typeRef="RootArtifactType"/>
</ArtifactType>

<ArtifactType name="ScriptArtifact">
   <DerivedFrom typeRef="RootArtifactType"/>
   <PropertiesDefinition
       element="ScriptArtifactProperties"/>
</ArtifactType>

<ArtifactType name="ArchiveArtifact">
  <DerivedFrom typeRef="RootArtifactType"/>
   <PropertiesDefinition
       element="ArchiveArtifactProperties"/>
</ArtifactType>

<ArtifactType name="PackageArtifact">
   <DerivedFrom typeRef="RootArtifactType"/>
   <PropertiesDefinition
       element="PackageArtifactProperties"/>
</ArtifactType>
```

Listing 7: TOSCA Primer Artifacts Types used for iKnow[21]

The installation of the iKnow system on the cloud environment requires careful setting of the environment, artifact to set the service components (like the required web application and database servers) and packages and file to deploy and host the custom solutions. All of the element must be orchestrated and set in a predefined order.

The basic artifacts types for the iKnow use case are: file, script, package and archive artifacts. Since the TOSCA primer already defines these types, we use the same. Listing 7 shows the basic artifacts defined by TOSCA primer [21].

The archive artifacts are used for packaging a collection of files and can contain additional metadata. In the iKnow use case these types are used for packaging the storage files. The package artifacts contain collections of software application or service files and in our case are used for the application deployment code. Listing 8 shows the created archive artifact for this intention. The script packages are used for deployment, configuration and similar activities.

```
<ArtifactTemplate id="iKnowApp"
   name="iKnowApplication-archive"
       type="ArchiveArtifact">
   <Properties>
    <ArchiveArtifactProperties>
     <ArchiveInformation
         archiveReference="files/iKnowApp.zip"
         archiveType="zip"/>
    </ArchiveArtifactProperties>
   </Properties>
   <ArtifactReferences>
    <ArtifactReference
        reference="files/iKnowApp.zip"/>
   </ArtifactReferences>
</ArtifactTemplate>
```

Listing 8: Archive Artifact used for iKnow application

D. Plans

Plans are defined as process models. TOSCA relies on BPMN or BPEL languages to define a plan. There is no official support for creation of management plans. For that reason we have created diagrams that can be transformed into one of the offered languages. At this moment the plans for iKnow system include deployment and server instantiation management in case of increased workload of the system. Fig. 5 depicts the deployment topology plan this application.

VI. Discussion

TOSCA is a recent proposal for standard and it is still in its early stage of adoption, presented by the current version 1.0. There are no commercial environments that enable usage of this specification, besides the several efforts made in this area for open-source solutions [17], [16]. Although it is promising, there is no guarantee of a wider acceptance by major cloud providers and software developers; and further standard commercialization.

Another possible approach is the usage of BPMN to define plans that will manage the life cycle of the application, especially for more complex management. BPMN is a general-purpose language and does not offer support for creation of

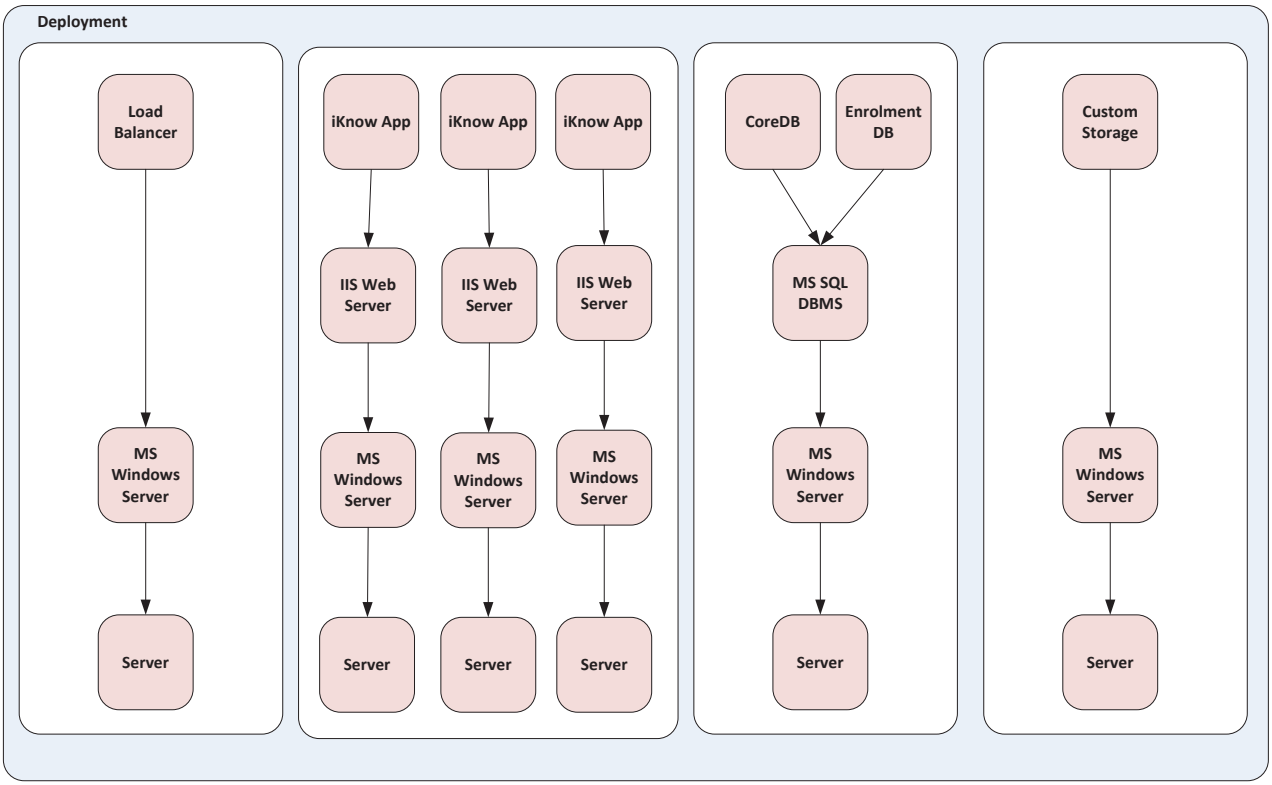


Fig. 5: Deployment Topology for iKnow

management plans. An effort in this direction is in process [22], but without commercial environment to support this effort, the results are not visible yet. This is the reason why the plans for the iKnow use case are still work in progress.

On the other hand TOSCA offers flexible definition of a software application. For smaller applications it offers straightforward way of definition, but for more complex applications, the specification may become very large and complex. In this paper we have analyzed the iKnow case study, as a mid-size project, where the specification grows fast. Usage of modeling tools like Winery [18] can ease this process. The greatest benefit is the promise of re-using the same specification by different cloud providers and environments. Another benefit, in this moment, is the formalized specification, which can be transformed in future and used for other purposes.

To create a TOSCA specification for a given application is a huge challenge at this time, since there is only one published primer [21] and it does not offer a lot of details. Creation of some parts of the iKnow TOSCA specification was based on assumptions considering the absence of concrete environment to test, like proper configuring of the environment with scripts, ability to convert the plans for installation and instantiation management into BMPL, etc.

An important concern is the usage of commercial solutions (in the case of iKnow it is based on Microsoft solutions, such as MS Server, MS SQL Database etc.). Due to the closed environment it is not always possible to properly configure all the necessary data with scripts or APIs. Other main concern is the security definition and settings. This is still a crucial work-in-progress area in the cloud.

Initial steps towards specifying an automated general procedure to extend the TOSCA specification and enable portability of SaaS applications has been reported by the authors in [23]. Automated portability was analysed in [24]. We have also demonstrated how a general procedure can be applied on a simple SOA transactional-based service example [25].

VII. Conclusion

In this paper we have presented a TOSCA specification of the iKnow University Management System, which is so far a unique example and case study to apply this new standard for university management software solutions. The proposed specification and challenges we have faced are not specific only for the given software solution, but can be rather used as a case study to ease the process of TOSCA specification of similar solutions.

Besides the benefits that TOSCA offers, we have pointed current drawbacks of this specification. This work contributes to support the future progress of TOSCA as a developing

standard with bright future in the area of cloud provision of interoperable services.

As future work we plan to continue working and expanding this specification with plans and to help creating an environment where this specification can be tested. We also plan to redefine and rework the iKnow system to create two independent modules (Core and Enrollment) so they can be independently transferred to separate platforms.

ACKNOWLEDGMENT

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius" University, Skopje, Macedonia through the project XaaSin-Cloud (Everything-as-a-Service in Cloud). All testing was performed on facilities at the Laboratory of Innovative Internet Technologies.

The iKnow project has been developed as a part of the Tempus JPGR 511342 Project.

REFERENCES

[1] I. Chorbev, M. Gusev, D. Gjorgjevikj, and A. Madevska-Bogdanova, "Architecture of an electronic student services system and its implementation," in *ICT Innovations 2012, Web proceedings*, S. Markovski and M. Gusev, Eds., 2012. ISBN 1857-7288 pp. 381 – 400.

[2] Organization for the Advancement of Structured Information Standards. (2013, Mar.) Topology and orchestration specification for cloud applications version 1.0. [Online]. Available: http://docs.oasis-open.org/tosca/TOSCA/v1.0/cs01/TOSCA-v1.0-cs01.html

[3] T. Binz, G. Breiter, F. Leymann, and T. Spatzier, "Portable cloud services using tosca." *IEEE Internet Computing*, vol. 16, no. 3, pp. 80–85, 2012. doi: 10.1109/MIC.2012.43

[4] F. Moscato, R. Aversa, B. Di Martino, T. Fortis, and V. Munteanu, "An analysis of mosaic ontology for cloud resources annotation," in *Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on*. IEEE, 2011, pp. 973–980.

[5] S. Labes, J. Repschlager, R. Zarnekow, A. Stanik, and O. Kao, "Standardization approaches within cloud computing: evaluation of infrastructure as a service architecture," in *Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on*. IEEE, 2012, pp. 923–930.

[6] G. A. Lewis, "Role of standards in cloud-computing interoperability," in *System Sciences (HICSS), 2013 46th Hawaii International Conference on*. IEEE, 2013. doi: 10.1109/HICSS.2013.470 pp. 1652–1661.

[7] Z. Zhang, C. Wu, and D. W. Cheung, "A survey on cloud interoperability: taxonomies, standards, and practice," *ACM SIGMETRICS Performance Evaluation Review*, vol. 40, no. 4, pp. 13–22, 2013.

[8] F. Galán, A. Sampaio, L. Rodero-Merino, I. Loy, V. Gil, and L. M. Vaquero, "Service specification in cloud environments based on extensions to open standards," in *Proceedings of the Fourth International ICST Conference on COMmunication System softWAre and middlewaRE*. ACM, 2009. doi: 10.1145/1621890.1621915 pp. 1–12.

[9] A. Ranabahu, E. Maximilien, A. Sheth, and K. Thirunarayan, "Application portability in cloud computing: An abstraction driven perspective," *IEEE Transactions on Services Computing*, vol. PP, no. 99, p. 1, 2013. doi: 10.1109/TSC.2013.25

[10] Z. Hill and M. Humphrey, "CSAL: A cloud storage abstraction layer to enable portable cloud applications," in *Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on*. IEEE, 2010. doi: 10.1109/CloudCom.2010.88 pp. 504–511.

[11] N. Loutas, E. Kamateri, and K. Tarabanis, "A semantic interoperability framework for cloud platform as a service," in *Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on*. IEEE, 2011, pp. 280–287.

[12] G. Cretella and B. Di Martino, "Towards automatic analysis of cloud vendors apis for supporting cloud application portability," in *Complex, Intelligent and Software Intensive Systems (CISIS), 2012 Sixth International Conference on*. IEEE, 2012. doi: 10.1109/CISIS.2012.162 pp. 61–67.

[13] D. Petcu, "How to build a reliable mOSAIC of multiple cloud services," in *Proceedings of the 1st European Workshop on Dependable Cloud Computing*, ser. EWDCC '12. ACM, 2012. doi: 10.1145/2365316.2365320. ISBN 978-1-4503-1149-6 pp. 4:1–4:2.

[14] G. Breiter, M. Behrendt, M. Gupta, S. Moser, R. Schulze, I. Sippli, and T. Spatzier, "Software defined environments based on TOSCA in IBM cloud implementations," *IBM Journal of Research and Development*, vol. 58, no. 2, pp. 1–10, 2014. doi: 10.1147/JRD.2014.2304772

[15] S. Durairajan and P. Sundararajan, "Portable service management deployment over cloud platforms to support production workloads," in *Cloud Computing in Emerging Markets (CCEM), 2013 IEEE International Conference on*, Oct 2013. doi: 10.1109/CCEM.2013.6684438 pp. 1–7.

[16] T. Binz, U. Breitenbücher, F. Haupt, O. Kopp, F. Leymann, A. Nowak, and S. Wagner, "OpenTOSCA – a runtime for TOSCA-based cloud applications," in *11th International Conference on Service-Oriented Computing*, ser. LNCS, vol. 8274. Springer, 2013. doi: 10.1007/978-3-642-45005-1_62 pp. 692–695.

[17] M. Kostoska, M. Gusev, and S. Ristov, "A new cloud services portability platform," *Procedia Engineering*, vol. 69, no. 0, pp. 1268 – 1275, 2014. doi: 10.1016/j.proeng.2014.03.118

[18] O. Kopp, T. Binz, U. Breitenbücher, and F. Leymann, "Winery – modeling tool for TOSCA-based cloud applications," in *11th International Conference on Service-Oriented Computing*, ser. LNCS, vol. 8274. Springer, 2013, pp. 700–704.

[19] U. Breitenbücher, T. Binz, O. Kopp, F. Leymann, and D. Schumm, "Vino4TOSCA: A visual notation for application topologies based on TOSCA," in *OTM 2012, Part I*, ser. LNCS, vol. 7565. Springer-Verlag, 2012. doi: 10.1007/978-3-642-33606-5_25 pp. 416–424.

[20] F. Li, M. Vogler, M. Claessens, and S. Dustdar, "Towards automated IoT application deployment by a cloud-based approach," in *Service-Oriented Computing and Applications (SOCA), 6th IEEE International Conference on*, 2013, pp. 61–68.

[21] Organization for the Advancement of Structured Information Standards. (2013, Jan.) Topology and orchestration specification for cloud applications (TOSCA) primer version 1.0. [Online]. Available: http://docs.oasis-open.org/tosca/tosca-primer/v1.0/tosca-primer-v1.0.html

[22] O. Kopp, T. Binz, U. Breitenbücher, and F. Leymann, "BPMN4TOSCA: A domain-specific language to model management plans for composite applications," in *Business Process Model and Notation*, ser. Lecture Notes in Business Information Processing, vol. 125. Springer, 2012. doi: 10.1007/978-3-642-33155-8_4. ISBN 978-3-642-33154-1 pp. 38 – 52.

[23] M. Kostoska, M. Gusev, and S. Ristov, "P-TOSCA portability model for PaaS hosted applications," University Ss Cyril and Methodius, Faculty of Computer Sciences and Engineering, Skopje, Macedonia, Tech. Rep. LiiT 22, 2014.

[24] M. Gusev, M. Kostoska, and S. Ristov, "P-TOSCA automated application portability," University Ss Cyril and Methodius, Faculty of Computer Sciences and Engineering, Skopje, Macedonia, Tech. Rep. LiiT 55, 2014.

[25] S. Ristov, M. Kostoska, and M. Gusev, "P-TOSCA portability demo case," in *2014 IEEE 3rd Int. Conf. on Cloud Networking (IEEE CLOUDNET)*, 2014.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 769–774
DOI: 10.15439/2014F366
ACSIS, Vol. 2

Performance Analysis of Distributed Internet System Models using QPN Simulation

Tomasz Rak
Rzeszow University of Technology,
Department of Computer and Control Engineering
al. Powstancow Warszawy 12, 35-959 Rzeszow, Poland
Email: trak@kia.prz.edu.pl

***Abstract*—The paper presents creating web system models. The aim of the work was to develop models of the distributed Internet system that allow the performance evaluation. From many possible methods we have selected Queueing Petri Nets consisted of two classes of formal models (Queuing Nets and Petri Nets). In the paper web systems are modeled by Queueing Petri Nets tool. The paper includes the selected results of models simulation. Our approach predicts the performance of distributed Internet system.**

I. Introduction

THE Internet system consists of a set of distributed nodes to provide up-to-date data in set time frames. Groups of nodes (clusters) are organized in layers conducting predefined services (e.g. WWW service).

Nowadays, Internet systems modeling and design develop in two ways. On the one hand, formal models which can be used to analyze performance parameters are proposed [1], [2], [3], [4], [5]. To describe Internet systems such formal methods like Queuing Nets and Petri Nets are used. Sometimes elements of the control theory are used to manage the movement of packages on web servers [6]. Experiments are the second way [7]. Applying experiments and models greatly influences the validity of the systems being developed. The convergence of simulation results with the real systems results confirms correctness of the modeling methods. The following mathematical models are used to describe Internet systems:

- analytical models: obtained on the basis of systems observation for the queuing systems with significant assumptions regarding the requests arrival and service process,
- simulation models of qualitative and quantitative analysis: Time Coloured Petri Nets, Queuing Petri Nets [8] or generalized queuing models based on the queuing theory (e.g. CSIM software libraries [9], [10]).

Our earlier works [9], [11] are based on Queuing Nets (QN) and Timed Coloured Petri Nets (TCPN). A distributed Internet system model, initially described in compliance with Queuing Net rules, is mapped onto Timed Coloured Petri Net structure by means of queueing system templates. We have used two types of formal models that have been exploited in the industry. In our elaboration we created separate system models using Queuing Nets and Petri Nets, which allow the performance analysis. We used experiments to check real distributed Internet system parameters. We verified some constructed models with the real experimental environment as a benchmark (Performance Engineering analysis). The validation results show that the model is able to predict the performance with error about 20[%] [9].

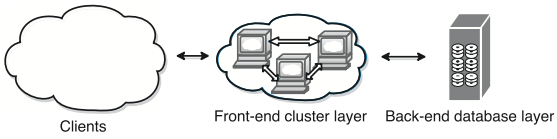


Fig. 1. Distributed Internet system architecture

Distributed Internet systems analysis based on Quality of Service metrics: performance (utilization, throughput, response time), availability, reliability. In these studies the performance is measured in terms of the mean response time of business transactions.

The remaining work is organized as follows. Section II presents distributed Internet system architecture. In the next section, we describe used formal methods. Section IV presents models and simulation analysis. The final section contains concluding remarks.

II. Distributed Internet System Architecture

Distributed Internet system architecture is made up of several layers. In our approach the presented architecture has been simplified to two layers (Fig. 1) based on [8] results:

- Front-end layer is based on the presentation and processing mechanisms.
- Back-end layer contains one or - in the case of replication [9] - several databases. This layer keeps the system data.

Architecture composed of these layers is used for e-business systems. The presented double-layer system architecture realizes Internet system functions. Access to the system is realized through transactions. Proposed in the paper our approach may be treated as an extension and continuation of solutions presented in [9].

In the paper we consider one class of Internet systems. In these systems a started transaction may be cancelled as a result of a system offer change. Not all transactions will be successfully finished. We shall consider cases in which the number of requests per second is hundreds or thousands.

978-83-60810-58-3/$25.00

Such a situation may cause the rejection of a large number of requests, due to timeliness loss. Therefore, partial processing of unrealized requests, also increases the response time for requests processed correctly. Transactions realization related to the system offer must take into account the results of previous transactions associated with this offer. Such systems are known as Interactive Internet Systems with Dynamically Changing Offers [9]. The presented systems class is interesting from the practical point of view. A stock exchange system (e-trading), where transactions are carried out on-line, could be their representative. The study considers the class of interactive Internet systems, for which the rate of offer change is equal to clients' time of interaction with the system. It is assumed that the offers are submitted by a seller or a broker. The offer change may cause that a transaction started earlier, is interrupted and unfinished. It is also assumed that transactions are realized immediately and they apply a common set of resources (such as sale of goods). It is assumed that the buyer can buy a collection of shares in a single transaction. The following features distinguish the described systems from others:

- short response time required - a necessity to transfer results to a client in a short time,
- sequence processing - a large part of transactions requires sequential processing,
- peak of intensity processing - processing a large number of client requests at the same time.

The Internet systems have different response time requirements. The response time for different Internet system types divided into three main groups. The major group, from the viewpoint of our study, is a group of systems, for which the required response time is the shortest. On-line auctions, sports betting, on-line ticketing and e-trading are distributed Internet system examples [9]. Within the Internet system classes we can distinguish a class for which the service is heavily dependent on the offers' time variability.

The characteristic feature of many Internet systems is a large number of customers using the Internet services at the same time. In the case of the described systems class, customers are often focused on one event related to the same system offer (the same database resources). Based on these futures we used e-trading system as a benchmark with two-layered architecture (cluster in front-end layer and one database instance in back-end layer).

III. Mathematical Models

In our solution we propose a very popular formal method - Queueing Petri Net (QPN). This method is based on Queueing Nets and Petri Nets.

A. Queuing Theory

Queuing Theory deals with modeling and optimizing different types of service units. Queueing Net usually consists of a set of connected queuing systems. The various queue systems represent computer components. Queueing Nets are very popular for the quantitative analysis. To analyze any queue system it is necessary to determine:

- arrival process,
- service distribution,
- service discipline,
- scheduling strategies.

B. Petri Nets

Petri Nets are used to specify and analyze the concurrency in systems. The system dynamics is described by the rules of tokens flow. The net scheme can be subjected to a formal analysis in order to carry out a qualitative analysis, based on determining its logical validity. Petri Nets are referred to as the connection between engineering description and theoretical approach. Petri Nets are well-known models used to describe and analyze the service units. Petri Net cannot be used for a quantitative analysis due to lack of time aspects. Some Petri Nets, such as Stochastic Petri Nets or Time Coloured Petri Nets, try to meet the requirements of quantitative analysis. The studies focus on incoming load measuring, e.g. measure of the response time or presentation of an overall modeling plan.

C. Queueing Petri Nets

In our solution we propose Queueing Petri Net formalism [12]. There is a very popular formal method of functional and performance modeling (performance analysis). These nets provide sufficient power to express modeling and analyzing of complex on-line systems. The choice of Queueing Petri Net was caused by a possibility of obtaining the different character information. The main idea of Queueing Petri Net is to add queueing and timing aspects to the net places.

Queuing Nets - quantitative analysis - have a queue and scheduling discipline and are suitable for modeling competition of equipment. Petri Nets - qualitative analysis - have tokens representing the tasks and are suitable for modeling software. Queueing Petri Nets have the advantages of Queuing Nets (e.g., evaluation of the system performance, the network efficiency) and Petri Nets (e.g., logical assessment of the system correctness).

Queueing Petri Net consists of queueing places (resource or state) which contain two components: a queue and a depository for tokens that completed their service at a queue. Input transitions are fired and then tokens are inserted into a queueing place according to the queue's scheduling strategy. Tokens are entered into the queueing place in the same way as in other Petri nets. After service, the tokens are not available for output transactions. They are immediately moved to a depository, where they become available for output transitions. Queueing places can have variable scheduling strategies and service distributions or impose a scheduling discipline on arrival tokens without a delay. [8]

The response time for analysis was chosen from many Performance Engineering parameters. The response time is a sum of residences and queues time and service demand.

TABLE I
PARAMETERS OF EXPERIMENTAL ENVIRONMENT

	Parameter	Value
Software	Application server threads pool per node	30
	Database server connections pool per node	40
Client workload	Number of requests per second	5-20
	Number of clients	220
	Experiment time [s]	300

IV. PERFORMACE ANALYSIS

Queueing Petri Net models are used to predict the distributed Internet system performance.

A. Experiments

First we present the results of our experimental analysis. The goal is to check the service demand parameters for front-end and back-end nodes.

Deployment details are as follows: Gbit LAN network and three front-end nodes and one back-end node. Software environment is based on Linux and consists of: workload generator, load balancer (Apache Tomcat Connector), application server (GlassFish) and database server (Oracle). All important configuration parameters were described in the table (Table I).

Modern distributed Internet systems are usually built on middleware platforms such as J2EE. We use DayTrader performance benchmark which is available as an open source application. Overall, the DayTrader application is primarily used for performance research on a wide range of software components and platforms. Experimental system helps to identify configuration parameters. DayTrader is a suite of workloads that allows performance analysis of J2EE application server. DayTrader is a benchmark application built around the paradigm of an online stock trading system. It drives a trade scenario that allows to monitor the stock portfolio, inquire about stock quotes, buy or sell stock. The load generator is implemented using multi-threaded Java application connected to DayTrader benchmark. By client business transactions we mean the stock-broker operations: Buy Quote, Sell Quote, Update Profile, Show Quote, Get Home, Get Portfolio, Show Account and Login/Logout (Table II). Each business transaction emulates a specyfic class of client session.

Experiments (one node in front-end and one node in back-end layer) have shown that the mean number of requests per second (DayTrader was able to complete) for front-end layer is about 1300. The figure (Fig. 2) shows among others the mean number of requests per second (DayTrader was able to complete) for front-end layer (maximum 1309 requests per second for 220 clients and 15 requests per second workload). Respectively the mean measured number of requests per second for back-end layer is about 7500 requests per second.

Starting the server cluster in the front-end layer requires a mechanism that would allow an equable distribution of load. It must also be a gateway that transfers requests and responses between a user and an application. In such a

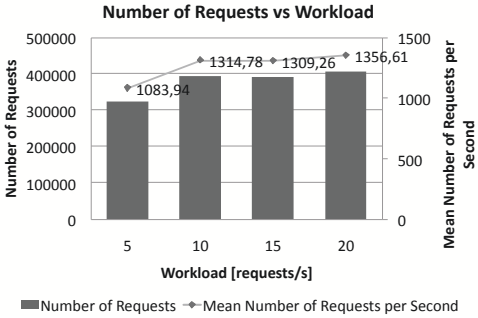


Fig. 2. Number of requests vs load (number of requests and mean number of requests per second)

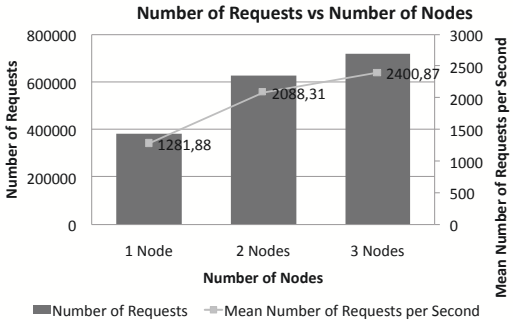


Fig. 3. Number of requests vs number of nodes (number of requests and mean number of requests per second)

scenario, only a gateway is visible from the outside and - on the basis of the request - it determines which part of the system (application server), and how, will be used to perform the request. Built-in load balancer is not available in the free version of the GlassFish server. Apache Tomcat Connector (mod_jk) has been used as the load balancer. Exampled client Uniform Resource Identifier query (Table II): http://[DayTraderApp]/daytrader/app?action=query.

Also cluster (three nodes in front-end and one node in back-end layer) experiments (Fig. 3) have shown that the mean number of requests per second (DayTrader was able to complete) for the front-end layer is about 2400 (for three front-end nodes, 220 clients and 15 requests per second workload). The mean measured number of requests per second for the back-end layer is the same as earlier.

One of the most important requests - Buy Quote (Requests class, which has a bigger impact on the behavior of the system (Fig. 4)) is used in simulations, because we have one class of requests in simulations. Simulations are only an approximation of reality. Buy Quote is only a few percent of all requests (Table III), because the experimental system is based on the real system workload.

B. Models

Multiple front-end nodes and one back-end node are the main configuration scenario. The Queueing Petri Net models (Fig. 5) are used to predict the system performance. We use the Queueing Petri net Modeling Environment [8] tool. Queueing

TABLE II
VALUE OF ACTION PARAMETER IN UNIFORM RESOURCE IDENTIFIER ADDRESS

Query	Transaction	Parameters	Description
buy (GET)	Buy Quote	$symbol$ – stocks symbols; $quantity$ – number	Buy and return the number of specified stocks
$sell$ (GET)	Sell Quote	$holdingId$ – stocks ID, which will be sold	Sell indicated stocks
$update\_profile$ (GET)	Update Profile	$password$ and $cpassword$ – new password; $fullname$ – name and surname; $address$ – address; $creditcard$ – credit card number; $email$ – email address	Update the logged-in user profile
$quotes$ (GET)	Show Quotes	$symbols$ – comma-separated stocks to display	Display information about the required stocks
$home$ (GET)	Get Home	–	Generates a logged-in user's homepage
$portfolio$ (GET)	Get Portfolio	–	Display a list of stocks held by the user
$account$ (GET)	Show Account	–	Display the logged-in user profile
$login$ (POST)	–	$uuid$ – user ID; $password$ – user password	Log the user in the system (session is created on the server side and its identifier returned in cookie)
$logout$ (GET)	–	–	Close the user session

TABLE III
PERCENTAGE OF QUERIES

Query	[%]
buy	5
$sell$	5
$update\_profile$	4
$quotes$	40
$home$	20
$portfolio$	12
$account$	10
$login/logout$	4

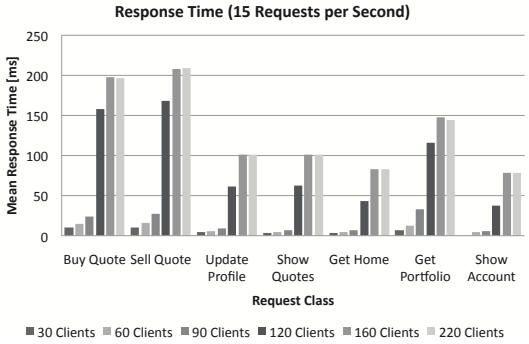


Fig. 4. Exampled real system response time for 15 requests per second workload (one node in front-end layer and one node in back-end layer)

Petri net Modeling Environment is an open-source tool for stochastic modeling and analysis based on the Queueing Petri Net modeling formalism was used in many works [13], [8], [2], [3], [4]. Total response time (Eq. 1) is a sum of all individual response times of queues and depositories in a simulation model without the client queue response time (client think time).

$$R = R_{QPN\_PLACES_{(QUEUE)}} + R_{QPN\_PLACES_{(DEPOSITORY)}} + R_{PLACES_{(QUEUE)}} + R_{QPN\_CLIENTS\_PLACE_{(DEPOSITORY)}} \quad (1)$$

Servers of the front-end layer are modeled using the Processor Sharing (PS) queuing systems (FE_CPU places). The back-end server is modeled by First In First Out (FIFO) queue (BE_I/O place). $PLACES$ (Eq. 1) represent the places (FE and BE) used to stop incoming requests when they await application server threads and database server connections respectively. Clients think time is modeled by Infinite Server (IS) scheduling strategy ($CLIENTS$ place). Application server threads and database server connections are modeled respectively by $THREADS$ and $CONNECTIONS$ places (Fig. 5).

Software and client workload parameters are the same as in the experiment environment. Service in all queueing places is modeled by an exponential distribution (λ parameter). Service demands in layers are based on experimental results in Sect. IV-A: $d_{FE\_CPU} = 0,714$ [ms] and $d_{BE\_I/O} = 0,133$ [ms]. Initial marking for places corresponds to the input parameters of the cluster experiment: number of clients (number of tokens in $CLIENTS$ place), application server threads pool (number of tokens in $THREADS$ place), database server connections pool (number of tokens in $CONNECTIONS$ place). In these models we have three types[1] of tokens: requests, application server threads and connections to the database server. The process of requests arrival to the system is modeled by the exponential distribution with the λ parameter (client think time) corresponding to the number of client requests per second.

C. Simulation Results

Many simulations were performed for various input parameters (Table IV).

The number of clients was increasing in accordance with the values presented in the table (Table IV). We used scenarios in witch we have a single requests class, the Buy Quote transactions. The first scenario involves a certain number of clients, a variable number of nodes and a variable number of requests per second for the entire system. The second scenario involves the response time of the entire system (Sys), the response time of the front-end layer and back-end layer (FE+BE) and the response time of the back-end layer (BE). In both scenarios the number of application server nodes is 1, 3, 6 and 9. The distributed Internet system model is used to predict the performance of the system for the scenarios mentioned above.

[1]A color specifying the types of tokens that can be resided in the place.

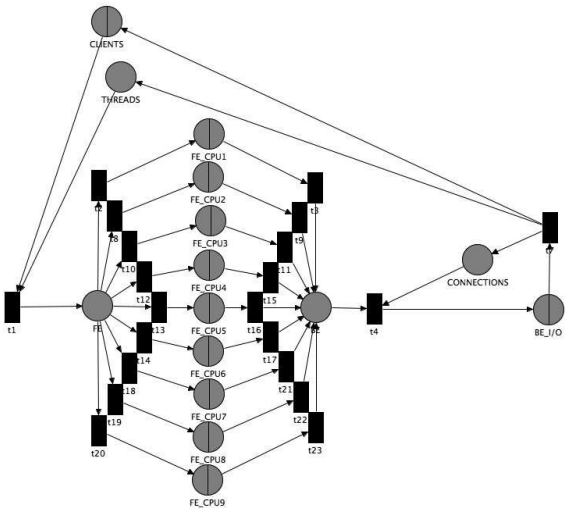


Fig. 5. Model of Internet system with front-end cluster (example for 9 nodes)

TABLE IV
PARAMETERS OF SIMULATIONS (ONE CLASS OF REQUESTS CORRESPONDS WITH BUY QUOTE REQUESTS)

	Parameter	Value
QPME	FE queue[a]	FE_CPUn
	BE queue	BE_I/O
Software[b]	THREADS place	30[c]
	CONNECTIONS place	40[d]
Client workload	λ	0,015[e]
	CLIENT place	30; 120; 210; 300; 390; 480
	Simulation time [s]	300

[a] n - number of front-end nodes
[b] Initial marking per node
[c] 30 threads for one front-end node, 60 threads for two front-end nodes, etc.
[d] 40 connections for one front-end node, 80 connections for two front-end nodes, etc.
[e] Client think time equals 66,67 [ms]

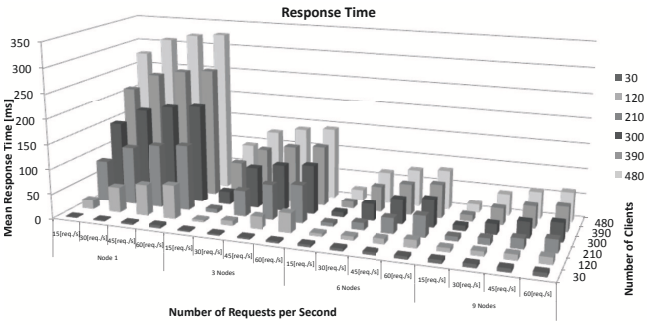


Fig. 6. Mean response time simulation results (system, fron-end and back-end layer, back-end layer) for different number of nodes (1, 3, 6 and 9), requests per second (15, 30, 45 and 60) and clients (30, 120, 210, 300, 390 and 480)

The figure (Fig. 6) reports the analysis results for all scenarios. In all cases, the model predictions are understandable. We investigate the behavior of the system as the workload intensively increases. As a result, the response time of transactions is improved for cases with a higher number of front-end nodes. As we can see increasing the number of nodes while simultaneously increasing number of application server threads and connections to the database is a good solution.

The overall response time decreases while the number of nodes increases (the change of requests per second (15, 30, 45, 60)). The response time of one front-end node architecture for all cases is the biggest. The response time difference between the 6 and 9 nodes is much smaller than that between 1 and 3 nodes in the front-end layer. When more nodes in the front-end layer are added the analysis of their impact on other elements of the system should be preluded.

In the second scenario the changes of the number of requests per second (1 and 3 nodes) do not have an impact on the back-end layer response time (BE). In the next cases with a higher number of nodes in the front-end layer (6 and 9) we can observe an increasing response time for the back-end layer. It can be seen already at 30 and more requests per second (Fig. 7). Overall system response time increases with increasing workload, even with a larger number of nodes (Fig. 7).

V. CONCLUSIONS

We can not always add new devices to improve performance, because the initial cost and maintenance will become too large. Also not every system can or should be virtualized or put in the cloud computing. Because the overall system capacity is unknown we propose the combination of benchmarking and modeling solution. Our earlier works propose Performance Engineering frameworks to evaluate performance during the different phases of their life cycle. Our present approach predicts performance for the distributed Internet system. The benchmark used in our work has got realistic workload.

We analyze the response time characteristics of different configurations. We develop a framework that helps to identify

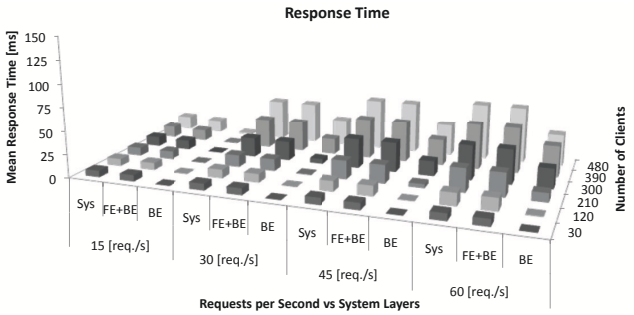


Fig. 7. Mean response time simulation results (system, fron-end and back-end layer, back-end layer) for different number of requests per second and clients (9 nodes)

performance requirements. The study demonstrates the modeling power and shows how the discussed models can be used to represent the system behavior. We used Queueing Petri Net models to predict the system performance for several different workloads and configuration scenarios. We used simulations because available analysis techniques are useless. It was not possible to predict the system performance under a large workload and a large number of nodes.

A number of different models of realistic size and complexity were considered. The benchmark was run for 300 seconds per test and each test was repeated 10 times to improve the reliability of results. The QPN model was simulated using the method of non-overlapping batch means method to estimate steady state mean token residence times. The average predicted response times are within the 95[%] confidence interval of the measured average response times. For all the simulations the confidence intervals were sufficiently small for the results to be reliable. Our analysis showed that the data reported by SimQPN is very stable.

The convergence of simulation results with the real systems results confirms the correctness of the modeling methods and their theoretical values. The validation results show the main advantage of this model (Table V). The relative error is lower than 15[%]. QPN model is a better than QN or TCPN models.

TABLE V
MODELING RESPONSE TIME ERROR FOR SCENARIO WITH 300 CLIENTS (EXAMPLE FOR 60 [REQUESTS/S])

Number of nodes	Model [ms]	Measured [ms]	Error [%]
1	198,48	229,65	13,5
3	99,47	114,96	13,4

Energy consumption in information and communications technology is growing annually by 4[%] despite efficiency gains in technology. It is therefore important to study ways of reducing energy consumption [14]. Power consumption depends on the load and on the number of running nodes in the cluster-based Web system. We shall study the compromise between a perceived average response time and energy consumption (practical value).

The future research will focus on verification of the system behavior in the case of a higher number of requests classes used in simulations. We shall provide a larger-scale analysis using hundreds of nodes.

REFERENCES

[1] X. Chen, C. Ho, R. Osman, P. Harrison, and W. Knottenbelt, "Understanding, modelling and improving the performance of web applications in multi-core virtualised environments," in *Proc. of the 5th ACM/SPEC International Conference on Performance Engineering*. ACM, 2014. doi: 10.1145/2568088.2568102 pp. 197–207. [Online]. Available: http://dx.doi.org/10.1145/2568088.2568102

[2] S. Kounev, C. Rathfelder, and B. Klatt, "Modeling of event-based communication in component-based architectures: state-of-the-art and future directions," *Electron. Notes Theor. Comput. Sci.*, pp. 3–9, 2013. doi: 10.1016/j.entcs.2013.04.002. [Online]. Available: http://dx.doi.org/10.1016/j.entcs.2013.04.002

[3] H. Koziolek, "Performance evaluation of component-based software systems: A survey," *Perform. Eval.*, pp. 634–658, 2010. doi: 10.1016/j.peva.2009.07.007. [Online]. Available: http://dx.doi.org/10.1016/j.peva.2009.07.007

[4] P. Meier, S. Kounev, and H. Koziolek, "Automated transformation of component-based software architecture models to queueing petri nets," in *International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*. IEEE, 2011. doi: 10.1109/MASCOTS.2011.23 pp. 339–348. [Online]. Available: http://dx.doi.org/10.1109/MASCOTS.2011.23

[5] C. Rathfelder, D. Evans, and S. Kounev, "Predictive modelling of peer-to-peer event-driven communication in component-based systems," in *Computer Performance Engineering*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2010, vol. 6342, pp. 219–235. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-15784-4_15

[6] S. Samolej and T. Szmuc, "Htcpns-based modelling and evaluation of dynamic computer cluster reconfiguration," in *Advances in Software Engineering Techniques*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, vol. 7054, pp. 97–108. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-28038-2_8

[7] K. Zatwarnicki, "Operation of cluster-based web system guaranteeing web page response time," in *Computational Collective Intelligence. Technologies and Applications*. Springer Berlin Heidelberg, 2013, vol. 8083, pp. 477–486.

[8] S. Kounev, *Performance engineering of distributed component-base systems: benchmarking, modeling and performance prediction*. Shaker, 2006. [Online]. Available: http://books.google.pl/books?id=OQGeAAAACAAJ

[9] T. Rak and J. Werewka, "Performance analysis of interactive internet systems for a class of systems with dynamically changing offers," in *Advances in Software Engineering Techniques*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, vol. 7054, pp. 109–123. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-28038-2_9

[10] K. Zatwarnicki, "Identification of web server," in *Computer Networks*, ser. Communications in Computer and Information Science. Springer Berlin Heidelberg, 2011, vol. 160, pp. 45–54. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-21771-5_6

[11] T. Rak and S. Samolej, "Distributed internet systems modeling using tcpns," in *International Multiconference on Computer Science and Information Technology*, 2008. doi: 10.1109/IMCSIT.2008.4747298 pp. 559–566. [Online]. Available: http://dx.doi.org/10.1109/IMCSIT.2008.4747298

[12] F. Bause, "Queueing petri nets - a formalism for the combined qualitative and quantitative analysis of systems," in *Proc. of the 5th International Workshop on Petri nets and Performance Models*. IEEE Computer Society, 1993. doi: 10.1109/PNPM.1993.393439 pp. 14–23. [Online]. Available: http://dx.doi.org/10.1109/PNPM.1993.393439

[13] D. Coulden, R. Osman, and W. Knottenbelt, "Performance modelling of database contention using queueing petri nets," in *Proc. of the 4th ACM/SPEC International Conference on Performance Engineering*. ACM, 2013. doi: 10.1145/2479871.2479919 pp. 331–334. [Online]. Available: http://dx.doi.org/10.1145/2479871.2479919

[14] E. Gelenbe and R. Lent, "Trade-offs between energy and quality of service," in *Sustainable Internet and ICT for Sustainability*. IEEE, 2012, pp. 1–5.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 775–780
DOI: 10.15439/2014F454
ACSIS, Vol. 2

Implementation of a Network Based Cloud Load Balancer

Sasko Ristov, Marjan Gusev, Kiril Cvetkov
University Ss Cyril and Methodius, FCSE,
Rugjer Boshkovic 16,
Skopje, Macedonia
Email:{sashko.ristov, marjan.gushev}@finki.ukim.mk, kiril_cvetkov@yahoo.com

Goran Velkoski
Innovation LLC,
Vostanichka 118,
Skopje, Macedonia
goran.velkoski@innovation.com.mk

***Abstract*—Cloud service providers offer their customers to rent or release hardware resources (CPU, RAM, HDD), which are isolated in virtual machine instances, on demand. Increased load on customer applications or web services require more resources than a physical server can supply, which enforces the cloud provider to implement some load balancing technique in order to scatter the load among several virtual or physical servers. Many load balancers exist, both centralized and distributed, with various techniques. In this paper we present a new solution for a low level load balancer (L3B), working on a network level of OSI model. When a network packet arrives, its header is altered in order to forward to some end-point server. After the server replies, the packet's header is also changed using the previously stored mapping and forwarded to the client. Unfortunately, the results of the experiments showed that this implementation did not provide the expected results, i.e., to achieve linear speedup when more server nodes are added.**

***Index Terms*—Distributed Computing; HPC; Performance; Web Services.**

I. Introduction

THE dynamic way of living enforces the companies to be more adaptive to changes. A lot of new companies are emerging and growing fast, thus taking the market share ruled of the leading companies. Therefore, services in each company must be prepared for dynamic changes and cope with increasing or decreasing the load [1]. If a company wants to save costs and buy a smaller amount of resources, then those resources cannot handle the load peaks. Nevertheless, buying huge amount of resources will increase the costs and the resources will be underutilized most of the time [2].

A possible solution is to migrate the services in commercial clouds. This process requires a dynamic strategy for a given company, to enable efficient acquiring or releasing the cloud resources. Additionally, increasing the resources requires a sophisticated load balancing strategy to maximize the effective and efficient usage and utilization of the rented resources. The final effect is to maximize the cost, performance and utilization ratio.

Recently, we have proposed an architecture for a Low Level Load Balancer (L3B) [3]. This architecture provides a scalable cloud environment, which can distribute server load among several active virtual machines that are integrated over the communication link [4].

In this paper, we present a new load balancer, realized on a network level. It dynamically balances the incoming network packets among all active virtual machines and returns the responses to the clients that send the requests. This balancer adds a small latency, which does not impact the total response time. The balancer also increases the security of the services since the client cannot see the internal cloud network.

The rest of the paper is organized as follows. Section II presents the related work in the area of load balancing. In Section III, we briefly describe the architecture of L3B. The development and the performance of the L3B balancer implementing the proposed architecture is presented in Section IV. Finally, Section V concludes the work and presents future work.

II. Related Work

Load balancing is an inherited feature from grids onto clouds. It may have various uses in cloud computing, such as:

- *Failover*, as continuation of a service after the failure of a cloud resource;
- *Energy conservation and resource consumption* kept to a minimum; and
- *Scalability*, as a feature of cloud computing requirements.

Rimal et al. [5] give an overview of load balancing techniques used by various cloud providers and solutions, including Amazon, Google, SalesForce, Azure, Eucalyptus, OpenNebula, etc. Most of these techniques are implementations of a conventional Round Robin schema, weighted selection mechanisms, HAProxy, Sticky session, SSL Least Connect, Source address, cluster server load equalization, high performance protocols over EC2, hardware load balancing, cloud controllers, etc.

A lot of load balancing techniques have existed long time before the introduction the cloud computing paradigm and the virtualization technique, which can be grouped in three groups [6]:

- Session switching at the application layer;
- Packet-switching mode at the network layer; or
- Processor load balancing mode.

978-83-60810-58-3/$25.00

Heinzl and Metz [7] classified the load balancers in two groups: Hardware and software, and commercial and open source.

In this paper we propose a network based load balancer, realized by packet-switching mode at the network layer of OSI model.

However, load balancing is not the only sufficient essential part for realization of the cloud. Resource brokering is also important, which is known as elastic load balancer (ELB), such as Amazon's ELB. Hu-Sheng et al. [8] proposed TeraScaler ELB, which is able to dynamically adjust the processing capacity of back-end server cluster with the applied load.

In our previous research, we proposed an architecture of a L3B balancer [3]. We have developed the load balancer and in this paper we present the architecture and design. A set of experiments is also conducted in order to prove the performance of our solution.

Load balancing in the cloud was analyzed by several authors, with corresponding surveys about performance of various load balancing algorithms [9], [10].

Load balancing techniques can be either centralized on a specific server or distributed by content aware policy. The centralized load balancing techniques are controlled by a single central node and all the other nodes communicate with this node, such as the Central Load Balancing Policy for Virtual Machines (CLBVM) solution proposed by Bhadani and Chaudhary [11], or the CLBDM solution proposed by Radojevic and Zagar [12], which takes into consideration other parameters as server load and application performance on top of the Round Robin Algorithm.

Centralized load balancing approaches in cloud computing do not offer full scalability features due to design limitations and communication overhead [13].

III. Proposed Architecture

This section describes the architectural design of the new proposed L3B balancer.

A. Overall model

The overall system architecture is presented in Figure 1, presenting the position of all items and their external interconnections. L3B is placed in front of a pool of virtual machine instances on the cloud, communicating with the Internet clients from one side and with physical or virtualized servers from the other side. The main function is to realize the load balancing of the clients' requests for services on various servers.

The latency that the L3B adds to the client requests should be compensated with the reduced response time by the servers due to reduced number of requests.

The internal architecture of the L3B balancer [3] consists of two modules: *Resource Management Module (RMM)* and *Packet Management Module (PMM)*, as presented in Figure 2 [3]. The purpose of these modules is to realize accounting functions, which is essential for load balancing.

The main objective of the RMM module is to manage the provision of cloud resources with dynamic accounting of the current load and utilization of active virtual machine instances. Whenever the accounting shows a need for more resources, the RMM module communicates with the cloud controller, by sending a command to initiate creation of a new virtual machine instance. If there is an information of underutilization, a corresponding command is sent by the RMM module to shut an existing virtual machine instance.

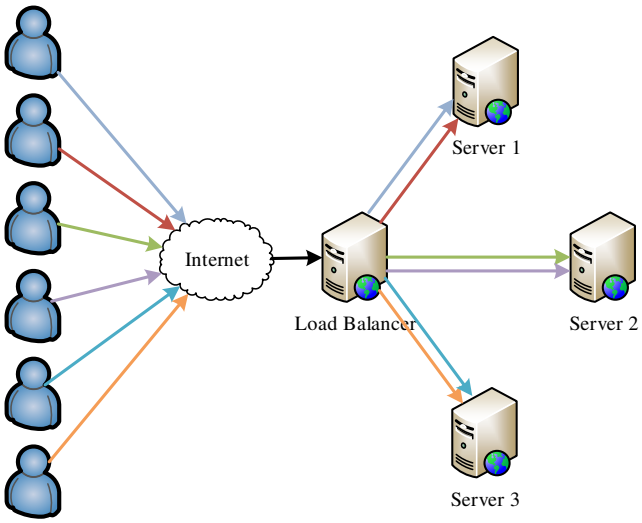


Fig. 1. L3B System architecture

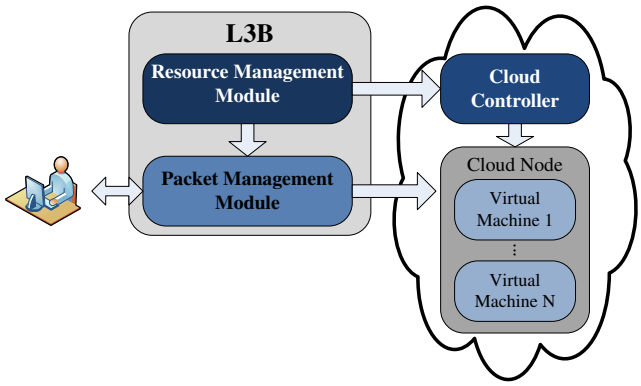


Fig. 2. Modules in the internal L3B architecture

Besides the connection to the cloud controller, the RMM also communicates with the PMM module. Particularly, the RMM module sends an information to the PMM module about active virtual machine instances to enable quality information about the work balance and to enable conditions for the PMM to realize the load balacing. The PMM module's main task is to redirect the input packets to some of the active virtual machine instances and then to forward the responses to the client that has sent the request.

This paper presents an extension of this idea, with details on the development of the PMM module, since we are interested

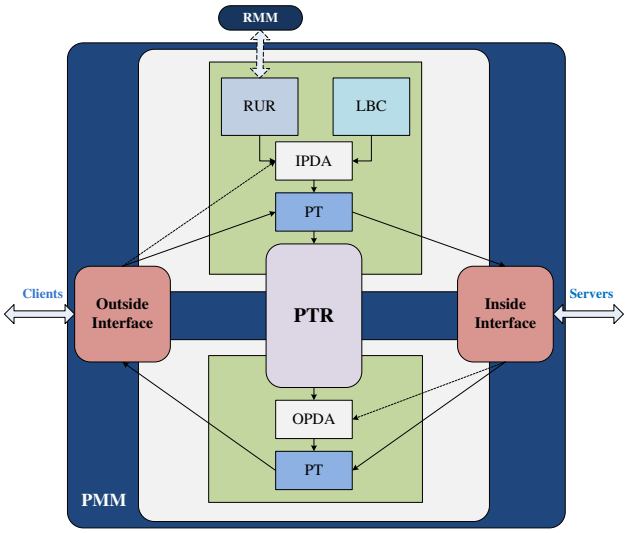


Fig. 3. Design of the PMM module

in presenting a proof that it will be efficient for load balancing of client requests among several active machines (or virtual machine instances in the cloud). The development of the RMM module is out of scope of this paper since it directly depends on the cloud controller and its APIs and interfaces.

B. Design of the PMM module

The PMM module is the core of the L3B balancer. The main functionality is to enable intelligent management of all the inconsistent traffic coming from the clients.

The design of PMM module is presented in Figure 3. It is used to manage the client's requests, i.e. to forward particular incoming packet to a particular virtual machine instance, enabling an environment to balance the load of all virtual machine instances. When the target virtual machine responds back, PMM forwards the response packet to the client that has sent the corresponding request.

The PMM module consists of two interfaces to establish a communication to:

- the *clients*, presented by the client machines that sends requests and
- *virtual machine instances*, presented by the Cloud Nodes.

Figure 3 shows the two internal modules for communication with the external parts:

- *L3B Outside Interface (OI)* that communicates with the clients, and
- *L3B Inside Interface (II)*, responsible to communicate with the virtual machine instances.

The main part of the L3B, realized as a core processing unit is the

- *Packet Translation Repository (PTR)* responsible for receiving the packets in both directions, their processing and forwarding to the destination.

The other internal PMM modules, which are designed to establish communication in direction from outside LAN to inside LAN, (direction from OI to II) and information exchange with the RMM module are the following:

- *Input Packet Decision Agent (IPDA)*, a part that realizes the most important agent, which implements the intelligent algorithm to derive smart decisions in assigning the requests to appropriate resources,
- *Resource Utilization Repository (RUR)* is a part that stores information about utilization of the physical resources, by sensing the active virtual machines and communicating with the RMM module;
- *Load Balancing Configuration (LBC)* is dedicated to store the information about configuration for the load balancing;
- *Packet Translation (PT)* agent is the part that realizes the low level networking on IP level in direction from OI to II.

The PMM module contains also parts responsible for low level network communication in direction from inside LAN to outside LAN (direction from II to OI), as follows:

- *Output Packet Decision Agent (OPDA)* is the agent responsible to make decisions and assign responses to the requestors, realized by appropriate communication with PT and PTR.
- Inverted *Packet Translation (PT)* agent realizes the low level networking on IP level in direction from II to OI.

The OI arbitrates among the clients and the PMM inner agents. Its function is based on a decision making and triggering some actions.

A typical scenario, when a new packet arrives, the OI instantly sends information to the IPDA. The IPDA agent uses a sophisticated intelligent algorithm to realize smart decisions needed to assign the requests to appropriate resources. The decision can be made only by using relevant information by the RUR and LBC about current utilization of the physical resources and load balancing configuration. Finally, after IPDA has analyzed the real time information of the hardware utilization, it determines which virtual machine can handle the request in order to preserve sustainable performance. The realized decision is sent to the PT agent, which receives the packet from OI and uses it to proceed with the IP header translation using the NAT/PAT (Network Address Translation / Port Address Translation). This functionality enables an environment to translate internal and external network addresses for each packet. These translations are then stored in the PTR and finally the corresponding transformed packet is forwarded to the II part, as inside interface to forward it to the target virtual machine instance.

As soon as the packet reaches the II part, it is forwarded to the corresponding cloud node and the target virtual machine instance. The packet has modified header to be transmitted in the network. The information stored in the header can be efficiently used for packet flow in the opposite direction. In this case the II receives the packet with responses from the

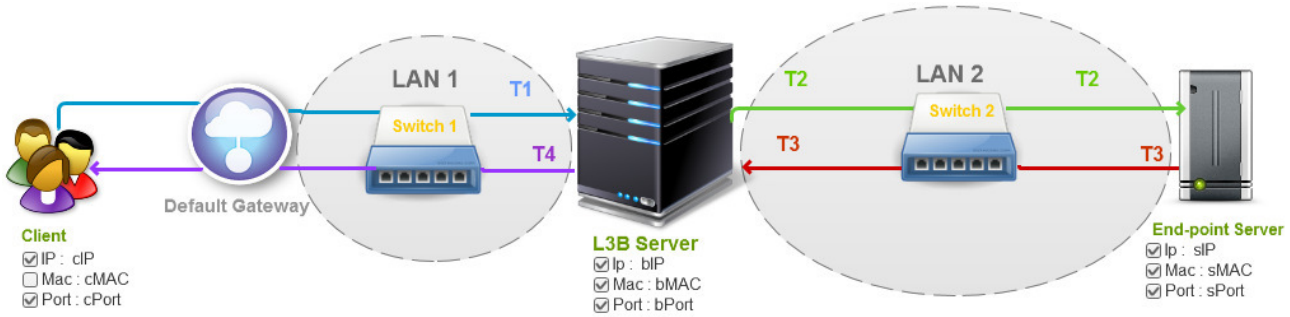


Fig. 4. The L3B implementation

virtual machine. The packet encapsulated in data link frame contains the corresponding MAC addresses of the source and target. Now the packet is sent to OPDA, which communicates to the PTR to receive detailed PT instructions. Note that there is an inverted PT in this section, which repeats the inverse NAT/PAT procedure to translate the external IP address into the original destination IP address. Then the packet is sent to the OI to forward the response to the client.

IV. The Implementation of L3B

This section presents the implementation of L3B according to the architecture described in Section III. The results of the experiments to determine its impact to the end-service's performance are also presented.

A. L3B functional requirements

The efficient and effective load balancer should comply with the following functional requirements:

- To balance all incoming packets that are sent by the clients to the active end-point servers;
- Must have a configuration file where a specific port can be configured where the clients should sent their requests;
- Must have a configuration file to specify the IP addresses of active end-point servers where the incoming packets will be balanced; and
- Introducing L3B between the client and more end-point servers must reduce the overall response time of the client requests.

B. The Architecture

The L3B is implemented in JAVA to be platform independent. This feature compensates the small JAVA deficiency, manifested as a reduced performance. Recently, this is not emphasized a lot, since modern Java virtual machines' performance do not fall behind other solutions such as C or C++ [14].

The architecture of the L3B implementation is depicted in Figure 4. Only one end-point server is presented in order to simplify the explanation. The participants in this communication scenario are the client, L3B server and end-point server (virtual machine hosted on the cloud). There are two LANs identified in the figure, LAN1 as a network segment between the default gateway and the L3B server, and LAN2, as a network segment between the L3B server and the end-point server (virtual machine hosted on the cloud). We assume that these LANs are supported by corresponding fast switches. In addition, this organization allows the end-point servers and L3B to be in the same subnet to decrease network latency. Considering the L3B modules, actually OI communicates with default gateway in LAN1 and II communicates with virtual machines hosted in LAN2. The participant' parameters (IP address, Mac address, Port) are presented below each participant, as presented in Figure 4.

TABLE I
Header values of a request in different states in the L3B implementation

Header Value	T_1	T_2	T_3	T_4
Source IP	cIP	bIP	sIP	bIP
Destination IP	bIP	sIP	bIP	cIP
Source port	cPort	obPort	sPort	bPort
Destination Port	bPort	sPort	obPort	cPort
Source MAC	Def. Gateway	bMac	sMac	bMac
Destination MAC	bMac	sMac	bMac	Def. Gateway

Lets discuss a typical scenario in this L3B architecture and implementation. When a client sends a request, it is routed through Internet until the L3B default gateway, identified as T_1 state in LAN1. Now, the incoming packet is received by the OI module in the L3B balancer. This packet is processed, as discussed in Section III.

The packet header processing, transforms its Ethernet, IP and TCP headers. More details for the scenario presented in Figure 4 are specified in Table I. Besides transformation, the old and new headers are stored in corresponding repository (PTR). As explained in the previous Section, IPDA is the agent that makes the decision to select the end-point server as a destination where the packet should be sent. In the figure, this is labeled as T_2 in the LAN2.

When the end-point server processing is finished, a response packet starts to be sent in the opposite direction in the analyzed architecture. The end-point server replies back to the L3B, by sending the packet in the LAN2, labeled with T_3 in the figure. Similarly to the previous explanation, II unit from the L3B balancer receives the packet, and OPDA uses the

information stored in PTR repository to make a decision about the header translation. Then the inverse PT agent transforms the corresponding Ethernet, MAC and IP addresses, according to the decision made by OPDA, and updates the configuration in PTR. After header translation, PT sends the packet with changed headers to the OI unit, which forwards it to the relevant client, which originated the counterpart request.

C. Implementation challenges

Our implementation of the architecture uses jNetPcap library [15] to sniff the network traffic and alter the incoming packets. Using the jNetPcap library made some problems during the packet forwarding.

The packet sniffer Wireshark [16] showed that a regular packet arrived at the server T_2, with correct values for appropriate fields in Ethernet and IP header. However, the server did not respond back to the L3B balancer and T_3 was not initiated.

The solution to this problem, acquired more experiments and programming. We redeveloped the L3B to open a new socket to the end point server for each packet. However, although this change solved the problem and the packet was sent back to the L3B balancer, and then from L3B to the client, the results were not promising. The following two sections present the testing methodology and the results of the experiment to determine the L3B performance.

D. Testing methodology

Four workstations with the same hardware resources and platform are used as a testing environment. Each workstation has Intel(R) Xeon(R) CPU X5647 @ 2.93GHz with 8GB RAM memory, installed with Ubuntu 12.04. The client is on the same LAN as the L3B workstation and two end-point workstations to reduce the network latency.

A client requests a packet with constant web page content, which size is $56KB$. The different number of concurrent requests are then initiated to create various loads. We choose this page size in order to be smaller than the IP packet limit of 64KB. Usually the web requests and responses are smaller than $64KB$.

The concurrent requests are simulated with SOAPUI. Each test case consists of sending N concurrent requests per second, such that N varies in each test case starting from $N = 5$ until $N = 100$ requests, by increasing N with step 5. Each test case lasts 60 seconds.

E. Performance analysis of the L3B implementation

The defined test cases were executed in order to check the impact of the L3B balancer to the end-point web server. Both experiment scenarios are examined, with and without L3B.

Figure 5 presents the results of the experiments. We observe that the last functional requirement is not satisfied for neither test case, meaning that the most important feature is not realized. In reality it shows that introducing the L3B balancer made the response time even worse, compared to the scenario without L3B.

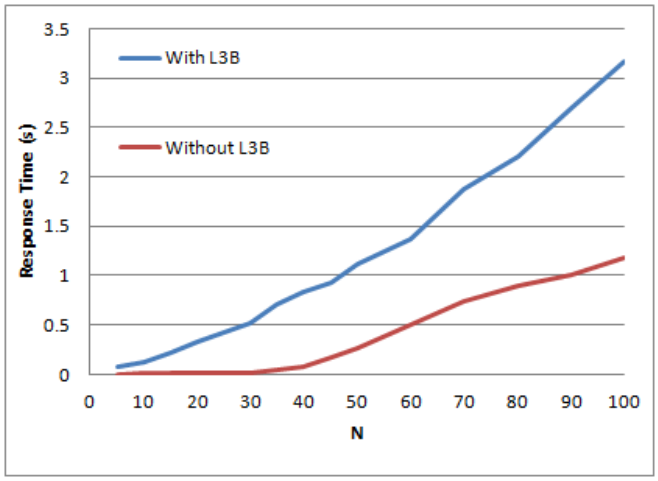


Fig. 5. Results of the experiments for the L3B implementation

The results show that although opening a new socket for each incoming packet was a solution to reply the server, it is performs worse and overall it is a bad solution.

This motivated us to analyze various proposals how to make the L3B more efficient. The main idea was to try to open one or several sockets from the L3B server to the servers in order to reduce the overall response time.

V. Conclusion and Future Work

In this paper we have presented a new approach for a load balancer, that is, the low level load balancer that works on a network layer of OSI model. The balancer is developed according to the L3B architecture [3], but its implementation have even reduced the performance when two end-point servers are used compared to the case when only one end-point server is used without the L3B balancer.

Since this implementation of the L3B balancer did not yield the expected results, we will proceed to improve the L3B architecture and implementation. Creating a new socket for each arrival packet reduced the performance of the L3B balancer and the architecture will be improved by creating a virtual client.

References

[1] A. Murua, I. Gonzalez, and E. Gomez-Martinez, "Cloud-based assistive technology services," in *Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on*, Sept 2011, pp. 985–989.

[2] S. Ristov, M. Gusev, G. Armenski, K. Bozinoski, and G. Velkoski, "Architecture and organization of e-assessment cloud solution," in *Global Engineering Education Conference (EDUCON), 2013 IEEE*, March 2013. doi: 10.1109/EduCon.2013.6530189. ISSN 2165-9559 pp. 736–743, best paper award. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6530189&isnumber=6530074

[3] M. Simjanoska, S. Ristov, G. Velkoski, and M. Gusev, "L3b: Low level load balancer in the cloud," in *EUROCON, 2013 IEEE*, Zagreb, Croatia, 2013. doi: 10.1109/EUROCON.2013.6624994 pp. 250–257. [Online]. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6624994

[4] L. Schubert, M. Assel, and S. Wesner, "Resource fabrics: The next level of grids and clouds," in *Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on*, Oct 2010. ISSN 2157-5525 pp. 677–684.

[5] B. Rimal, E. Choi, and I. Lumb, "A taxonomy and survey of cloud computing systems," in *INC, IMS and IDC, 2009. NCM '09. Fifth International Joint Conference on*, Aug 2009. doi: 10.1109/NCM.2009.218 pp. 44–51. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5331755&isnumber=5331299

[6] B. Radojevic and M. Zagar, "Analysis of issues with load balancing algorithms in hosted (cloud) environments," in *MIPRO, 2011 Proceedings of the 34th International Convention*, May 2011, pp. 416–420. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5967092&isnumber=5967009

[7] S. Heinzl and C. Metz, "Toward a cloud-ready dynamic load balancer based on the apache web server," in *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2013 IEEE 22nd International Workshop on*, June 2013. doi: 10.1109/WETICE.2013.63. ISSN 1524-4547 pp. 342–345. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6570639&isnumber=6570561

[8] H.-S. Wu, C.-J. Wang, and J.-Y. Xie, "Terascaler elb-an algorithm of prediction-based elastic load balancing resource management in cloud computing," in *Advanced Information Networking and Applications Workshops (WAINA), 2013 27th International Conference on*, March 2013. doi: 10.1109/WAINA.2013.79 pp. 649–654. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6550470&isnumber=6550285

[9] K. Nuaimi, N. Mohamed, M. Nuaimi, and J. Al-Jaroodi, "A survey of load balancing in cloud computing: Challenges and algorithms," in *Network Cloud Computing and Applications (NCCA), 2012 Second Symposium on*, Dec 2012. doi: 10.1109/NCCA.2012.29 pp. 137–142. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6472470&isnumber=6472451

[10] N. J. Kansal and I. Chana, "Cloud load balancing techniques: A step towards green computing," *IJCSI International Journal of Computer Science Issues*, vol. 9, no. 1, pp. 238–246, 2012.

[11] A. Bhadani and S. Chaudhary, "Performance evaluation of web servers using central load balancing policy over virtual machines on cloud," in *Proceedings of the Third Annual ACM Bangalore Conference*, ser. COMPUTE '10. ACM, 2010. doi: 10.1145/1754288.1754304. ISBN 978-1-4503-0001-8 pp. 16:1–16:4. [Online]. Available: http://doi.acm.org/10.1145/1754288.1754304

[12] B. Radojevic and M. Zagar, "Analysis of issues with load balancing algorithms in hosted (cloud) environments," in *MIPRO, 2011 Proceedings of the 34th International Convention*, May 2011, pp. 416–420. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5967092&isnumber=5967009

[13] D. Ardagna, S. Casolari, and B. Panicucci, "Flexible distributed capacity allocation and load redirect algorithms for cloud systems," in *Cloud Computing (CLOUD), 2011 IEEE International Conference on*, July 2011. doi: 10.1109/CLOUD.2011.32. ISSN 2159-6182 pp. 163–170. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6008706&isnumber=6008659

[14] D. A. Patterson and J. L. Hennessy, *Computer Organization and Design, Fourth Edition: The Hardware/Software Interface*. Morgan Kaufmann, 2009. ISBN 978-0-12-374493-7

[15] S. Technologies, "jnetpcap," 2014. [Online]. Available: http://jnetpcap.com/

[16] "Wireshark," 2014. [Online]. Available: http://www.wireshark.org/

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 781–790
DOI: 10.15439/2014F337
ACSIS, Vol. 2

Supporting job-level secure access to GPGPU resources on existing grid infrastructures

John Walsh, Jonathan Dukes
School of Computer Science and Statistics,
Trinity College Dublin,
Email: John.Walsh@scss.tcd.ie, Jonathan.Dukes@scss.tcd.ie

***Abstract*—Grids provide secure, utility-like access to a wide variety of large-scale, distributed computational and storage resources. In particular, the European Grid Infrastructure (EGI) and Open Science Grid (OSG) have excelled in processing vast workloads of independent jobs for the research community.**

Researchers demand increasingly faster processing speeds to solve increasingly larger and more complex problems. To meet this need, attention has shifted over the past decade away from single-core processing models towards the use of multi-core, many-core and massively parallel computational accelerators. The increasing availability and use of General Purpose Graphic Processing Units (GPGPUs) are an example of this.

This paper addresses many of the challenges that exist in the integration of resources such as GPGPUs into Grid infrastructures. Specifically, solutions are proposed for discovering and describing GPGPU Grid resources, specifying multi-GPGPU job requirements, performing multi-GPGPU allocation to jobs, dynamically updating publicly-readable GPGPU usage information and enforcing GPGPU access control to prevent distinct jobs from inadvertently accessing the same device. The proposed solution is fully compatible with widely-used and accepted standards and middleware including the GLUE 2.0 schema and EGI Unified Middleware Distribution. A prototype implementation is also described.

I. Introduction

GRID Computing [1] developed out of the need for geographically distributed scientific communities to co-operate in order to investigate scientific problems with increasing computational complexity. The most famous example of such a community is the Particle Physics community investigating the existence of the Higgs Boson using the Large Hadron Collider (LHC). Not only do such large-scale problems require a given scientific community to share vast amounts of data among its (distributed) users, it may also be unfeasible to to process this data at a single location (known as a "Site" or "Resource Centre") – hence distributing the data processing tasks and storage to multiple locations is often required. Such grid-systems draw upon distributed computing, resource discovery and sharing, distributed data-management, authentication and authorisation, role-based access control and process accounting.

Grids developed around a "single program/single CPU" execution model. However, since 2005 the exponential growth of CPU speed and processing power has plateaued [2], and this has generated some questions about the future of computational-based scientific research using this "single program/single CPU" approach. *Parallel Processing* taking advantage of emerging multi-CPU cores (multicore) or many processing cores (many-core) on General Purpose Graphic Processing Units (GPGPUs) or Intel's Xeon Phi Computational Accelerator is regarded as a "white-knight" like solution to this problem. Indeed, the trend towards the extensive usage of GPGPUs and Intel's Xeon Phi, commonly known as "Computational Accelerators" (CAs) [3], in High Performance Computing environments can be seen in the twice-yearly "Top 500 Supercomputer" lists [4].

Support for grid-based parallel applications using Message Passing Interface (MPI) [5] has been available for a number of years [6]. No such support currently exists for the emerging CA-based parallel-processing architectures despite indications that many grid resource-centres and users were planning to incorporate CAs into their future work-plans [7] [8].

The core objective of the work presented in this paper is to address the challanges of integrating GPGPU resources into grid infrastructure in such a way that there are no changes (or minimal changes) to how the user works. In addition, there should there be no significant changes to the grid infrastructure itself. The approach taken to solve this integration problem is to first consider the more general problem of integrating any new resource. A set of *Grid Resource Integration Principles* is developed (Section III) that take these constraints into account.

In this paper: Section II reviews some of the key concepts in Grid Computing, including service-discovery (the Grid Information System) and the job submission lifecycle. Section III introduces the multi-layer abstract achitecture that seperates GPGPU job resource requests from the GPGPU allocation and access protection layers. Section IV presents a realisation of this model in the form of a prototype execution-model that extends the architecture and capabilities of a grid based on the popular Unified Middleware Distribution (UMD). This extension provides new services that address four key grid components, namely: (i) GPGPU service discovery; (ii) multi-GPGPU resource allocation through a grid job description language and batch system integration; (iii) dynamic updating of publicly-readable status information describing the GPGPU resource usage that complies with current standards; and finally, (iv) per-job access controls that prevent distinct jobs from inadvertently accessing the same GPGPUs. Section V looks at related work, and when applicable, discusses how and why the approach taken here differs. Finally, Section VI

978-83-60810-58-3/$25.00

reviews the current implementation and the scope for future work in this area.

II. Grid Computing

The terminology "Grid" and "Grid Computing" has a wide and varied interpretation [9]. With this in mind, there is a need to clarify these terms in the context of this work and thus define the scope of the work. This section introduces key concepts in Grid Computing – specifically Grids such as the European Grid Infrastructure (EGI) and Open Science Grid (OSG) that primarily provide computing and storage resources to researchers. A high-level overview of the lifecycle of a user's job, from job submission through to execution using a Grid resource, will be presented. Grids are, by nature, large-scale, distributed infrastructures and accessing Grid resources relies on maintaining a uniform, structured, global view of the Grid. An overview of the systems that maintain this information will also be presented.

Key Concepts

A Grid is a distributed collection of computational and storage resources where (i) each resource is controlled and managed solely and independently by its owner or *resource-provider* (for example a University, research centre, company or private individual) and (ii) each resource-provider has some level of control over how the resource is accessed and used. This definition is sufficiently general to include both large-scale Grids, such as EGI or OSG, and also "compute-cycle" volunteer donation systems such as BOINC [10].

The work described in this paper is concerned with large-scale Grids, such as EGI or OSG, that are composed of multiple *resource-centres*, each operated by their owner or *resource-provider*. Each resource-centre provides one or more *Compute Elements* (CEs). These are services that provide access to computational resources such as a cluster of worker nodes accessed through a batch system. A *Grid Information System* is used to publish the capability of each resource-centre, in the form of a description of the resources that it provides, the current utilisation and availability of those resources and a description of the mechanisms for accessing them. To make use of the computational resources provided by a Grid, users submit *jobs* that are described using a formal *job description language* (JDL) (or *resource-specification language* (RSL)). In simple terms, a job consists of an executable program, a specification of the software environment in which the program must run, any input parameters and data required by the program and the files that will contain the outputs from the job. Users are grouped into *Virtual Organisations* based, for example, on their research area and these groupings are used to control access to – and account for the usage of – grid resources.

Grid users can request that their jobs be executed on a specific resource, based on *a priori* knowledge of the capabilities of a resource-centre. However, ideally Grids such as EGI and OSG will allow users to submit a job and let the Grid "decide" where that job should execute. To facilitate this, as well as capturing basic information about the job (e.g. executable program, input parameters and data), a JDL description of the job can also describe the job's resource requirements (e.g. number of CPU cores, minimum CPU specification, minimum memory requirement, software environment). In this scenario, instead of submitting a job directly to a resource-centre, the job is submitted to a Grid *Workload Management System* (WMS). The WMS acts as a broker, using the information published about each resource-centre through the Grid Information System, together with the description of the job, to find all resources that match the job's requirements. Furthermore, the broker can select one of the matched resources (e.g. using pre-defined policies or heuristics) and orchestrate the execution of the job on the chosen resource.

The Grid Job Life-cycle

The usual starting point when submitting a job to the grid is the User Interface (UI). This is a service node that contains the necessary command tools to interact with other grid services. The UI is configured to interact with one or more Workload Management Systems (WMS).

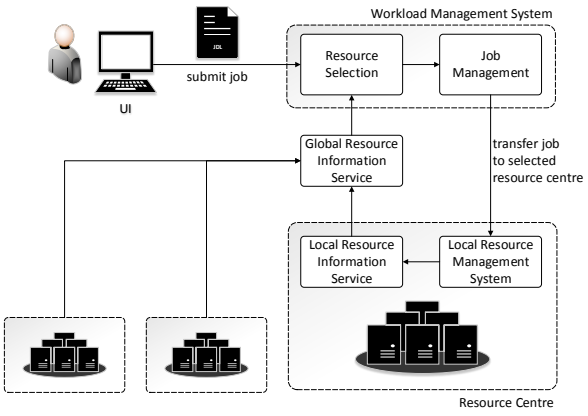


Fig. 1: The Job Submission Chain

Figure 1 illustrates the life-cycle of a single grid job, from the submission of the job in JDL on the UI, through its orchestration by a WMS, to its eventual execution at a resource-centre. A high-level outline of the flow of a grid job through the WMS includes the following stages:

1) When the job is submitted to the WMS, a copy of the JDL file and any input files specified in it are copied to the Workload Management System (WMS).
2) The WMS will determine all locations where the job can possibly run. This is the "match-making" process. The potential locations are then *ranked* in preference. The user can also influence the rank ordering by specifying a *RANK* expression in the JDL.
3) Once a target CE at a resource-centre is chosen, the WMS will engage with the CE.
4) The CE builds a *JobWrapper*. This executable is built taking into account the local batch system, also known

as a *Local Resource Management System* (LRMS), on the CE. Some of the roles of the JobWrapper are to build a job submission script for the batch system, transfer all input files from the WMS, and then submit the job to the batch system.

5) The batch system is responsible for scheduling the execution of the user's job on one or more *CPU-Cores* on a *worker node* (WN).
6) After execution, the output files specified by the JDL description are transferred back to the WMS and can be retrieved by the user.

The match-making process requires knowledge of the overall state of the distributed set of resources. This state information is managed by a "Grid Information System".

Grid Information System

The Grid Information System is a critical component required for discovering services, determining the status of resources and selecting resources for submitted jobs. This system is composed of an information model (a *schema*) for describing entities (such as computational resources and available software) and the relationship between those entities, as well as a "presentation layer" that publishes this information as a frequently-updated queryable service (a *realisation*).

The *Grid Laboratory Uniform Environment* schema (*GLUE schema*) was developed as an Open Grid Forum (OGF) "reference standard" for multi-disciplinary Grids. There are currently two major (incompatible) versions of the GLUE schema in common use – GLUE 1.3 [11] and GLUE 2.0 [12]. Grids such as EGI are currently transitioning from GLUE 1.3 to GLUE 2.0 by running services that can utilize both standards. GLUE 2.0 offers many improvements over GLUE 1.3, including a richer description of grid entities and their state and the ability to easily publish extra details about grid-entities (using "OtherInfo" attributes). Moreover, GLUE 2.0 was developed to improve inter-grid interoperability [13].

Information in a Grid Information System originates from *Information Providers*. The services that are used to access and manage grid resources (e.g. a batch system for a set of worker nodes) should provide an interface that allows the Information Provider for that resource to determine the properties and current state of the resource, transform this information into an appropriate GLUE entity and publish the information (Figure 2).

Propagating GLUE entities so that they are visible "globally" in a Grid Information System follows a natural hierarchical structure (Figure 3): GLUE entities are generated by Information Providers; the set of Information Providers (info-providers) on a particular node publish their state as a *local resource*; the set of local resources form a *domain*; and the combined set of entities from each domain yield a global view.

The GLUE information "presentation layer" is typically managed by the *BDII*([14], Sec. 3.3.5). This is an implementation of a the hierarchical Grid Information System model with three BDII "types" and a set of "information providers" that generate information about the Grid entities. As per

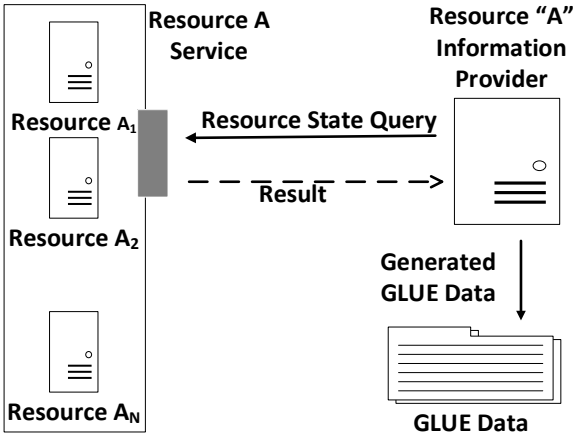


Fig. 2: Abstract Information Provider Model. An Information Provider queries a service that manages a set of Resources of type "A" at a resource centre. The Information Provider transforms the response into one or more GLUE entities.

Figure 3 the Resource-BDII (lowest BDII level) aggregates the state of a grid service node by executing a set of Generic Information Providers (GIP) plugins; the Site-BDII aggregates all the Resource-BDIIs belonging to the given site; and the Top-level BDII aggregates all the Site-BDIIs. Information is "pulled" from the lower to higher levels. In general, all queries about the state of the Grid are made through the Top-Level BDII.

III. Integrating GPGPU resources into Existing Grid Infrastructures

Many grid resource centres have already deployed GPGPU resources [7]. There is, however, no support in place for users to discover these GPGPU resources or submit jobs that specify a dependency on GPGPUs. Currently, users wishing to use these resources rely on *a priori* knowledge of the location and properties of available GPGPUs and the job submission mechanisms required to use them. Furthermore, after inspecting GLUE data published by the Top-Level BDII *lcg-bdii.cern.ch*, it was determined that from a sample of 2887 unique GlueCEUniqueID entities (i.e. interfaces to queues on the grid-connected batch systems), over 58% of these GlueCEUniqueID reported that the Torque/PBS (excluding PBSPro) batch system was used (Table I (a)). In particular, 779 GlueCEUniqueIDs (27%) used Torque/PBS 2.5.7 (Table I (b)). This is indicative of the default UMD Torque/MAUI installation that does not correctly handle generic consumable resources. From this data, it is reasonable to conclude that a significant percentage of grid systems on the EGI grid infrastructure do not support GPGPUs as consumable resources.

This paper addresses the challenge of integrating GPGPUs as first-class grid resources. From a user's perspective, these resources should be easy to discover without relying on *a*

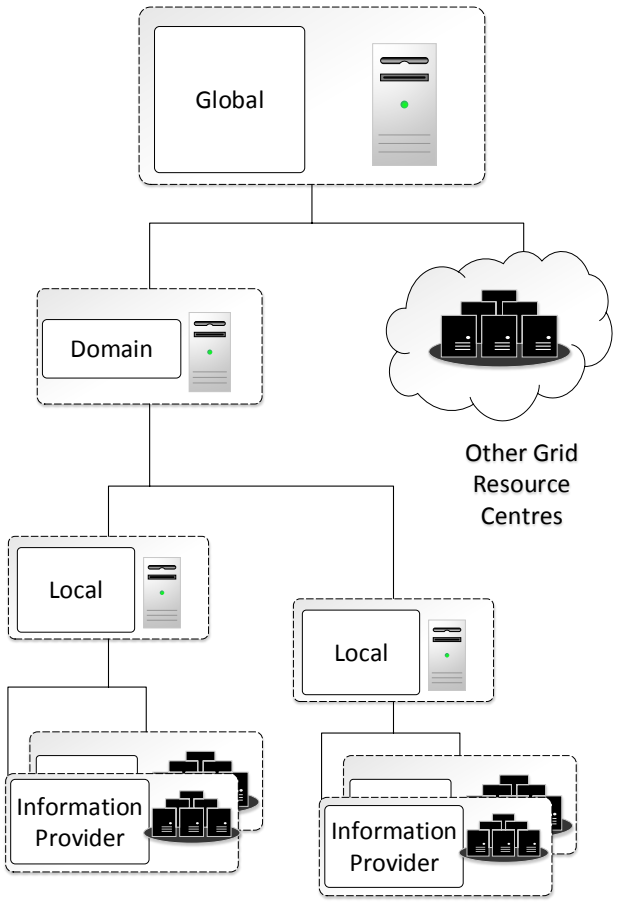


Fig. 3: The Grid Information System. GLUE data is propagated from the lowest level (generated by Information Providers) up to Global level.

TABLE I: Reported LRMS Types

(a)

LRMS Type	Count
Torque	930
PBS	740
LSF	704
Condor	296
GE/SGE	141
SLURM	63
PBSPro	12
Fork	1
Total	2887

(b)

Torque/PBS	Count
2.5.7	779
2.5.X (ex 2.5.7)	706
undefined	84
4.X	45
3.X.	24
2.4.X	8
2.3	24
Total	1670

priori knowledge (*discovery*); the specification, requirements and ranking of these resources should be independent of how the resources are locally managed (*independence*); and the user should have some assurance of exclusive access to the resource (*exclusivity*). Rather than considering GPGPUs specifically, the challenge is viewed as a generic consumable resource access problem. The approach proposed for addressing this challenge is summarised by the following *Grid Resource Integration Principles*:

- **Discovery:** The resource should be published as one or more GLUE entities. The act of publishing the resource into the Grid Information System makes the resource discoverable by users and services. An entity should, where possible, contain attributes that reflect the resources properties, and also where possible, these attributes should be *quantitative*. This allows resources of the same type to be compared and ranked against each other. (For example, the vendor, model and performance characteristics of a GPGPU may be of importance when selecting appropriate resources for a job.)
- **Independence:** There should be a method through which the required resource can be specified using a job description language. The specification should be independent of the way in which the resources are locally managed. (For example, the mechanism used to specify GPGPU requirements to the Torque/MAUI scheduling system differs from that used by SLURM.)
- **Exclusivity:** A resource allocated to a job by a batch system should be available as if it were exclusively allocated to the job. (For example, a GPGPU allocated to a job should not be available or even visible to another job running on the same worker node.)

Case Study – GPGPU integration

As an example of the application of the above *Grid Resource Integration Principles*, the use-case of GPGPUs as grid resources is considered. This use-case is interesting not only because of their parallel-processing capabilities, but because they are also representative of the class of LRMS *Generic Consumable Resources* - a class of non-CPU resources that can be managed by the LRMS.

Discovery: The salient properties that help describe a GPGPU are similar to those used to describe CPUs: vendor, model, memory, speed, benchmarked performance, number of physical GPGPUs per worker node. LRMS properties that can influence WMS orchestration and ranking include, the total number of installed GPGPUs and the number that are currently allocated through the batch system. Users may also be interested in the software required to access the resource, and basic installation details. These properties could be advertised in the Grid Information System as one or more GLUE entities and optionally used as a filter during resource selection.

Independence: The user needs an LRMS-independent way to specify the number of GPGPUs required or the number of GPGPUs that the job needs per-CPU core (this implies a minimum number of GPGPUs per worker node).

Exclusivity: The user needs assurances that two or more jobs concurrently executing on the same worker node cannot use the same GPGPU. In the absence of batch-system support for such exclusivity, an additional mechanism must be provided.

In the next section, a prototype extension of a Grid that integrates GPGPU resources is presented. This prototype introduces a GLUE 2.0 entity and GPGPU Information Providers that provides for the discovery of GPGPU resources; a means to specify GPGPU resource requirements that is independent of the batch system used; and, finally, a new mechanism to ensure that GPGPU resources are allocated exclusively to each job.

IV. Prototype Implementation

As shown in section II, the mechanism for orchestrating the execution of a grid job follows a complex chain of events. As a result, providing access to new grid resources, such as GPGPUs, is non-trivial. In particular, if these resources are to be provided by *existing* grid infrastructures that are *in-production* and in continuous use by an extensive community of users, the challenge becomes more acute. Adding support for new resources cannot be dependent, for example, on architectural changes, the replacement of core services or modifications to the GLUE schema. Instead, the approach taken must integrate with existing infrastructure, while adhering to the principles of *discovery*, *independence* and *exclusivity* (Section III).

The approach taken in this prototype is to provide a set of modular *hooks*, that in principle can be applied to many other resource types, other than GPGPUs. In this section, it will be shown that this approach requires relatively small changes to existing grid middleware.

A. Prototype Infrastructure

The focus of this prototype on the integration of Nvidia GPGPUs using the CUDA runtime and application development framework. (Nvidia GPGPUs are the most widely used GPGPU in High Performance Computing centres.) The prototype was developed using the UMD gLite grid middleware, and consists of a User Interface (UI), a Workload Management System (WMS), a Top-Level BDII (BDII), grid-security infrastructure services, and a Resource Centre using the CREAM CE. The CREAM CE uses Torque 2.5.7 as the batch system server and the MAUI batch scheduler. (MAUI was modified with a publicly available patch to enable Generic Consumable Resources.)

B. Discovery: GPGPU Schema and Information Providers

Section III considered what information about GPGPU resources should be represented through the GLUE-schema. In this prototype, a simple representation of these details is realized by using the GLUE 2.0 *ApplicationEnvironment* entity [12]. This entity is primarily used to describe the properties of software installed on worker nodes. However, a key feature of the ApplicationEnvironment is that it supports non-mandatory attributes that relate to software capacity and utilisation. This feature is used to publish installed GPGPU capacity and utilisation. Furthermore, the definition also allows for other arbitrary information to be published, such as the GPU vendor, model and speed and benchmarked performance.

(An alternative and perhaps superficially more obvious approach would have been to use the GLUE 2.0 *ExecutionEnvironment* entity, which provides for attributes such as CPU vendor, model and speed which may also used to describe GPGPUs. This approach, however, would describe an architecture in which the GPGPUs are independent of – rather than reliant on – the CPUs on the same worker nodes. In contrast, using the ApplicationEnvironment entity to describe GPUs allows the relationship between CPUs and their attached GPUs to be captured by the Grid Information System.)

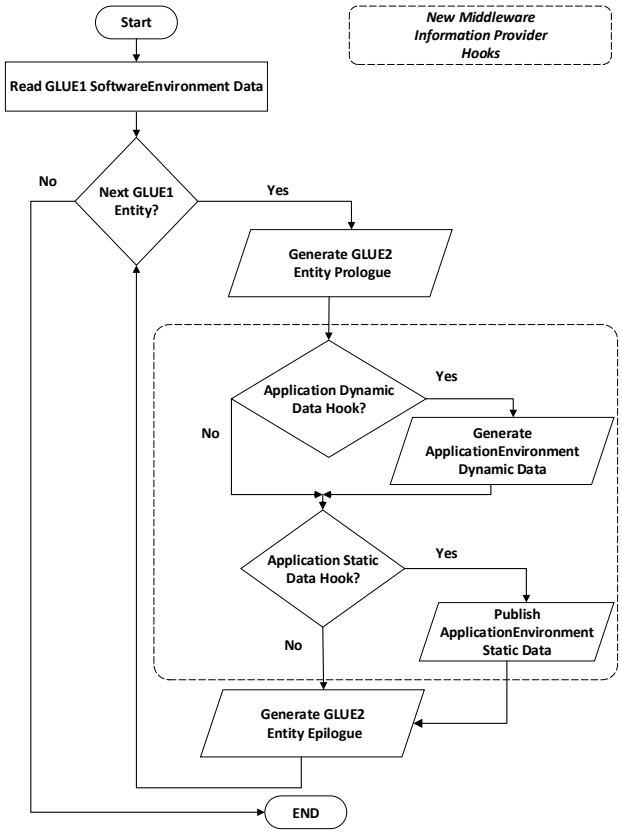


Fig. 4: Workflow of Modified ApplicationEnvironment Information Provider

The current gLite implementation of the GLUE 2.0 ApplicationEnvironment Information Provider already generates ApplicationEnvironment entities by converting GLUE 1.3 SoftwareEnvironment entities into their minimal GLUE 2.0 equivalents. In this prototype, this existing Information Provider is replaced with a new Information Provider that facilitates per-application *hooks*. For each application, separate hooks can be provided to generate static and dynamically changing GLUE data. This is illustrated in Figure 4.

In particular, the prototype uses new CUDA ApplicationEnvironment hooks that publish static data pertaining to the GPGPU hardware and CUDA software properties as well as

dynamically gathering and publishing previously unavailable GPGPU capacity and utilisation data from MAUI. Listing 1 illustrates an example of some of the output generated by the execution of a modified Information Provider where both dynamic and static hooks for the CUDA environment have been added.

Listing 1: An example of the output from the static and dynamic GLUE 2.0 CUDA ApplicationEnvironment hooks

```
...
objectClass: GLUE2ApplicationEnvironment
GLUE2ApplicationEnvironmentMaxJobs: 32
GLUE2ApplicationEnvironmentAppName: CUDA
GLUE2ApplicationEnvironmentFreeJobs: 30
...
GLUE2EntityOtherInfo: ↩
    GPUCUDAComputeCapability=2.1
GLUE2EntityOtherInfo: GPUMainMemorySize=1024
GLUE2EntityOtherInfo: GPUCoresPerMP=48
GLUE2EntityOtherInfo: GPUCores=192
GLUE2EntityOtherInfo: GPUClockSpeed=1660
GLUE2EntityOtherInfo: GPUECCSupport=false
GLUE2EntityOtherInfo: GPUVendor=Nvidia
GLUE2EntityOtherInfo: GPUPerNode=2
```

Table II indicates where the CUDA ApplicationEnvironment hooks provide additional Attribute-Value pairs that are part of the GLUE2 standard (but not generally used), and where the ApplicationEnvironment schema has been further extended (as allowed by the standard) by adding GLUE2 *OtherInfo* values. Furthermore, Table II lists the data-type of each value, the source of the data, and whether the data is generated dynamically or provided statically.

TABLE II: Extended GLUE2 CUDA ApplicationEnvironment

Standard ApplicationEnvironment		**Source**	**Creation**
MaxSlots	Integer	LRMS	Dynamic
MaxJobs	Integer	LRMS	Dynamic
FreeJobs	Integer	LRMS	Dynamic
New EntityOtherInfo Attributues		**Source**	**Creation**
ApplicationArea	String	System	Static
GPUCUDAComputeCapability	Float	GPGPU	Static
GPUMainMemorySize	Integer	GPGPU	Static
GPUMP	Integer	GPGPU	Static
GPUCoresPerMP	Integer	GPGPU	Static
GPUCores	Integer	GPGPU	Static
GPUClockSpeed	Integer	GPGPU	Static
GPUECCSupport	Boolean	GPGPU	Static
GPUVendor	String	GPGPU	Static
GPUModel	String	GPGPU	Static
GPUPerNode	Integer	LRMS	Static

C. Independence: Specifying and Handling GPGPU Job Requirements

Independence implies that the grid user should be able to specify required GPGPU resources within the existing job submission framework in a manner that is independent of any CE batch system implementation. By considering how jobs are orchestrated through a WMS, the changes required to the grid infrastructure to achieve this goal are outlined below.

JDL GPGPU requirements specification: The prototype allows a user to request GPGPUs by adding the following JDL specification:

```
GPUPerNode=X;
```

Here, *X* is the number of GPGPUs to be allocated *per node*. This specification requires no changes to the JDL Language definition syntax. As well as specifying the number of GPUs required, the JDL should also specify the ApplicationEnvironment entity that advertises the availability of the required GPGPU. An example of the complete JDL for a job requiring two GPGPUs and using the CUDA framework is shown in Listing 2.

Listing 2: Example GPGPU Job for gLite based Grids

```
[
Executable = "myScript.sh";
StdOutput = "std.out";
StdError = "std.err";
InputSandbox =  {
                 "GPGPU_acquire_prologue.sh",
                 "GPGPU_job_script.sh",
                 "GPGPU_release_epilogue.sh"
                };
OutputSandbox = { "std.out","std.err"} ;
VirtualOrganisation = "gputestvo";
Requirements = (Member("CUDA",other.↩
    GlueHostApplicationSoftwareRunTimeEnvironment)↩
    );}
GPUPerNode=2;
]
```

Once the job is submitted to the WMS, the set of potential resource-centres is filtered to include only those that advertise that they support the specified ApplicationEnvironment entity (CUDA in the above example). After applying some further job requirements, the WMS will select a matching CE, transfer the job *workload*, and then orchestrate its execution.

Resource Centre Job Execution: The orchestration of the job on the chosen resource centre CE can be classified into three stages: (i) job preparation; (ii) job submission; and (iii) job execution on the selected worker nodes(Figure 5). The *job preparation* stage is used to convert the JDL GPGPU resource specification into an LRMS specification.

During the Job Preparation phase, job input files and executable programs are transferred from the WMS to the CREAM CE. A "JobWrapper" script is created by CREAM. The JobWrapper is executed on the CE, and is responsible, among other things, for constructing a job execution environment appropriate to the LRMS. The JobWrapper script created by CREAM already contains a mechanism to pass or "Forward" job requirements to the LRMS [15]. This prototype exploits this mechanism by implementing a JobWrapper "batch requirements forwarding" script that will parse a copy of the JDL to determine if the $GPUPerNode$ value is defined and, if it is, return a suitable resource request, the format of which will be dependent on the specific batch system.

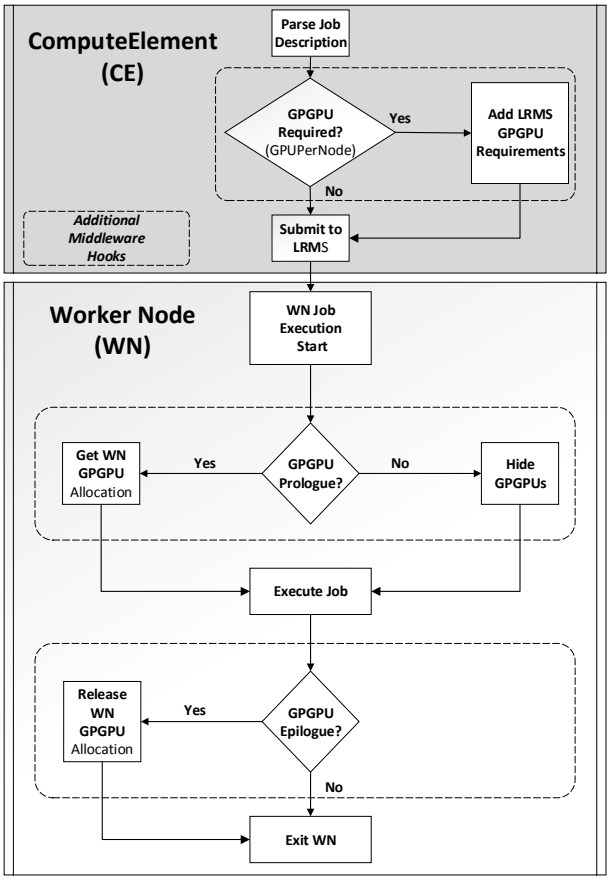


Fig. 5: Batch System GPGPU allocation process.

D. Exclusivity: Restricting Visibility of GPGPU Resources

Multi-core worker nodes allow many independent jobs to execute simultaneously. These independent jobs are protected from each other by executing in their own private disk-area, and by using file-access and process protection enforced by the operating system kernel. GPU (and consequently GPGPU) resources are designed to be accessible by all worker node users. In the case of GPGPUs exposed through a batch system, it is desirable to control access to them. Nvidia supports a number of ways in which access to its GPGPUs can be controlled:

1) the CUDA_VISIBILE_DEVICES environment variable can be used to restrict the visibility of a set of GPGPU devices in a process,
2) the Nvidia *Compute Modes* can permit/deny sharing of a GPGPU between multiple processes.

The handling of GPGPU allocation on common open-source batch systems, such as Torque/MAUI, is still problematic, in particular with mid-range GPGPUs where Nvidia-SMI tools do not report a full complement of utilisation data. A simple prototype service was developed that allows Torque/MAUI batch jobs to request a set of free GPGPUs on the worker node, and to release those GPGPUs back to the service when no longer required. Furthermore, the allocated GPGPUs are not visible to other user jobs on the worker node.

The service, which executes on each worker node, implements a lightweight web-server using the Tornado [16] framework. To improve the security of the service, it runs as an unprivileged user, GPGPU allocation states are maintained in a lightweight persistent database, and the service is internal to the worker node (i.e. it is not available to other nodes).

The server responds to two types of requests: *requestGPUS* and *releaseGPGUs*. These requests are respectively called from the within the *GPGPU_acquire_prologue.sh* and *GPGPU_release_epilogue.sh* Job Hook scripts, specified in the JDL (Listing 2).

Requests to the prototype tornado server must adhere to a strict syntax, and all malformed requests are dropped by the server. In the case the Torque implementation, a request is generated by sending a copy of the "PBS_JOBCOOKIE" and a list of *Universally Unique Identifiers* (UUIDs) - one per requested GPGPU. The PBS_JOBCOOKIE was chosen because, unlike the batch system "jobid", this value is not exposed to other users or processes.

The server manages the allocation of GPGPUs to jobs by using a single database with two tables: *UUID_JOBID* and *GPU_UUID*. The GPU_UUID table associates GPGPU devices with a job-generated UUID. The UUID_JOBID table associates UUIDs to a suitable unique value, such as the $PBS\_JOBCOOKIE$.

1) Initialisation: The GPU_UUID table is initialized at node start-up. A row is added for each physical GPGPU. Similarly, the UUID_JOBID table is created, but remains unpopulated until a GPGPU is requested by a job. Table III shows the initial state of the database tables for a node with two GPGPUs.

TABLE III: Initial State

(a)

GPUID	UUID
0	
1	

(b)

JOBID	UUID

2) Allocation Request: A GPGPU allocation request (Figure 6) should be sent to the server before the job attempts to execute any GPGPU code. When the server receives an allocation request message, a "Request Handler" will attempt to validate it. If the request is valid, then the string is converted into two component values: a JOBID (PBS_JOBCOOKIE) and a list of UUIDs. The request handler then iterates over the list of UUIDS and adds (UUID,JOBID) tuples to the UUID_JOBID table. The handler will also iterate over the list of UUIDs, selecting the first free GPGPU (i.e. rows where the UUID cell is NULL) from the GPUID_UUID table. The selected row is the updated with a new (GPUID,UUID) tuple. Changes to the database are committed. Finally, the server returns a text string to the client in the form of a comma-separated list of integers. This is the list GPGPU devices that

the the job has been allocated. This value is assigned to a *read-only* CUDA_VISIBLE_DEVICE environment variable, thereby ensuring that the user or process can no longer (inadvertently) update its value.

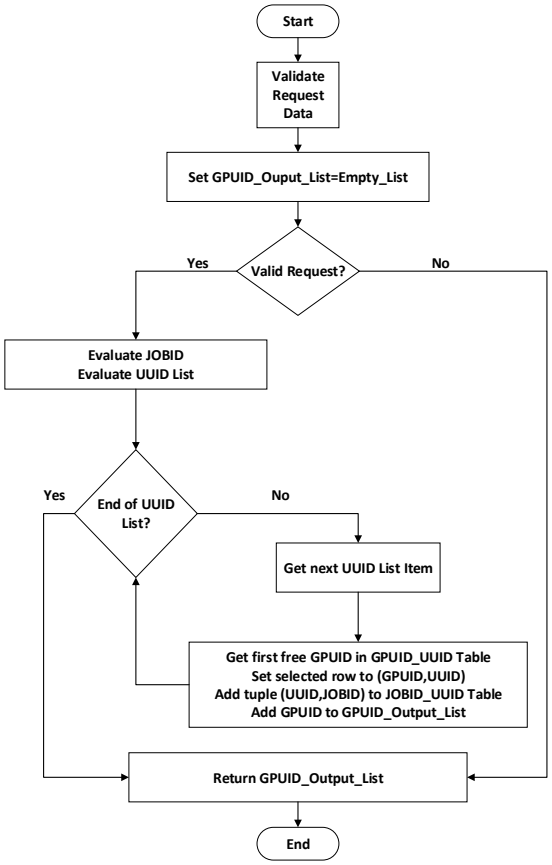


Fig. 6: Worker Node GPGPU allocation subsystem

TABLE IV: Example of a single job running on a node with two GPGPUs

(a) GPGPUs paired to UUIDs

GPUID	UUID
0	$UUID_1$
1	$UUID_2$

(b) $JOBID_1$ linked to $UUID_1$, $UUID_2$

JOBID	UUID
$JOBID_1$	$UUID_1$
$JOBID_1$	$UUID_2$

3) Release Request: A GPGPU "Release" request should be executed during the job's epilogue. This request is complementary to the "Allocation" request. As with the allocation request, the input is validated. If the request is valid, the server's "Request Handler" will remove all tuples matching the UUIDs from the UUID_JOBID table, and all the specified UUIDs from the GPUID_UUID table.

E. Results

Job submission was tested using different values for $GPUPerNode$. Sample output of a job that requested a single GPGPU but executed on a worker node with multiple GPGPUs is listed below (Listing 3):

Listing 3: Example GPGPU Job resricted to a single GPGPU

```
/usr/local/cuda/samples/1_Utilities/deviceQuery/↩
    deviceQuery Starting...
 CUDA Device Query (Runtime API) version (CUDART ↩
     static linking)
......
Detected 1 CUDA Capable device(s)
```

V. Related Work

CUDA_wrapper

CUDA_wrapper [17][18] was developed to help facilitate secure and controlled access to Nvidia-based GPGPU resources on multi-user GPGPU clusters. The wrapper acts by intercepting normal CUDA API fuction calls and overloading them with additional methods that transparently provide additional key-based access-control to the raw GPGPU device.

The approach taken in this paper avoids some drawbacks with the CUDA_wrapper method, namely: (i) Each API interception imposes a latency to handle each CUDA function call; (ii) CUDA_wrapper must be recompiled for each new version of CUDA; (iii) CUDA_wrapper did not work under testing with CUDA 5.5; and, (iv) The wrapper is vendor and LRMS (Torque) specific.

GPU resources on the Grid

There are several examples of current work where GPGPU resources have been partially integrated into Grids. These take the form of both non-virtualised GPGPU resources [19] and virtualised GPGPU resources [20]. In both cases, the grid-users job is given exclusive access to the GPGPU resource. However, both of these implementations lacks resource and service discovery and therefore require *a priori* knowledge of the existance of the resources. Moreover the virtualisation methodology imposes about 5% overhead on GPGPU access and runtimes.

BOINC and Desktop Grids

BOINC has support for multi-disciplinary computational sciences using GPGPU. For example, the *Einstein@Home* project uses GPGPUs to search for weak astrophysical signals from spinning neutron stars (also called pulsars) using data from the LIGO gravitational-wave detectors, the Arecibo radio telescope, and the Fermi gamma-ray satellite [21]. Furthermore, the EDGI Desktop Grid provides a mechanism [22] to bridge between EGI and BOINC-enabled resources. However, such combinations do not address the specific GPGPU service-discovery requirements.

HTCondor GPGPU Support

HTCondor [23] supports advanced GPGPU resource publication, service-discovery, per-job GPGPU match-making, and job-management [24]. Indeed, HTCondor is central to the the UMD WMS match-making service. However, some of the major differences between HTCondor and the work presented

in this paper include: (i) the solution is intended to be compatible with Grids using the OGF GLUE 2.0 information model; and (ii) The approach taken treats GPGPUs as an instance of generic consumable resources.

Application Hook Method

The design and implementation of the presented execution model used in this work follows a pattern similar to that used by MPI-Start [25]. In particular, this is evident in how software hooks are used to build an appropriate site-local application runtime-environment. The major differences in the implementation are: (i) MPI enabled resource centres currently publish information about the local MPI environment as a set of GLUE 1.3 SoftwareEnvironment tags. This information is inefficiently converted into a set of discrete and seemingly unrelated GLUE 2.0 ApplicationEnvironment entities; (ii) The GPGPU implementation attempts to exploit many of the newer features of the GLUE 2.0 ApplicationEnvironment definition. This includes the ability to publish GPGPU resource capacity and utilisation information.

Token based access-control

The use of UUIDs as tokens in the GPGPU allocation and release process is partially based on the use of time-limited UUID tokens in the *Puppet* [26] fabric managent system - a system used to install, configure an maintain machines on a network. In Puppet the token is used as a "shared-key" between the Puppet server and the client machine. The shared-key is used to authorise the download of the client's installation configuration file.

EGI GPGPU Working Group

The European Grid Initiative (EGI) is currently considering the GPGPU resources into grid infrastructures and one of the authors is a member of the EGI GPGPU Working Group [27]. The results presented in this paper represent independent work that may be contributed to this community effort in the future.

EGI Grid Federated Cloud Working Group

The EGI Federated Cloud Working Group uses the GLUE 2.0 ExecutionEnvironment to advertise diverse sets of (virtual) resources [28]. The authors had considered this appraoch, but deemed it to be unsuitable as a way to describe *Consumable Resources*.

VI. Conclusions and Future Work

This work shows how a GLUE 2.0 based multi-disipline scientific grid can be extended to support new models of parallel computing using GPGPUs. A methodology was developed that applied three abstract principles to this resource integration problem, namely: *Discovery*, *Independence* and *Exclusivity*.

The presented prototype is one of the first examples of where a GLUE 2.0 ApplicationEnvironment entity has been extended to include additional attributes that describe the capacity, utilisation and other properties of hardware used directly by the application itself. These attributes can be generated either statically or dynamically. The example use-case demonstrates a CUDA ApplicationEnvironment that is extended to include the total number of installed Nvidia GPGPUs, their current utilisation, and some selected hardware properties. This conforms to the *Discovery* principle.

A method was developed that allows a grid user to specify a GPGPU requirements in the Job Description Language. In particular, the prototype allows the user to specify the number of GPGPUs required per allocated Worker Node. The job requirement is converted into the native LRMS resource specification once the job enters the choosen Resource Centre. The method can also be applied to other resources made available through an LRMS. This is an application of the *Independence* principle.

Ensuring that users have guaranteed and isolated access to GPGPUs in a multi-user system can be difficult. This problem is compounded in the cases where multiple GPGPUs on the same machine can be accessed by multiple users - many batch systems do not indicate what GPGPU has been assigned to each user job. The worker node GPGPU Access Control system discusssed in Section IV-D can assuage this problem. The development of the GPGPU allocation handler for these systems ensures the *Exclusivity* principle.

Work is in progress to support enhanced job specifications, allowing the user to make job placement decisions based on a wider range of published GLUE 2.0 ApplicationEnvironment data. This allows for greater control over the resource discovery and selection process. An example based on published CUDA ApplicationEnvironment (Table II) is illustrated in Listing 4.

Listing 4: Example JDL using CUDA attributes

```
Requirements = GPUVendor=="Nvidia" && (↩
    GPUMainMemorySize >= 512);
```

Although the service demonstrated used Nvidia and CUDA, other environments such as AMD GPGPU and OpenCL can also be trivially accomodated - for example, AMD uses GPU_DEVICE_ORDINAL environment variable to restrict user visibility of AMD GPGPU devices [29]. In addition, this system can work in non-grid Torque/MAUI environments, and requires minor changes to work with other batch systems.

Finally, although the prototype implentation was tested with GPGPU resources, the model can be adapted to cater for wider range of resources, such as Intel's Xeon Phi, FPGAs and Licence controlled software. Said resources are difficult to integrate into grids based on GLUE 1.3. The prototype shows that: (i) there are cases where hardware resources can easily be treated like an ApplicationEnvironment; (ii) arbitrary information can be published about these resources - and this can be generated statically or dynamically; and (iii) job-requirements can be specified in a Job Description Langauge and transformed into batch-system directives without any major changes to the grid middleware.

ACKNOWLEDGMENT

This work was carried out on behalf of the Telecommunications Graduate Initiative (TGI) project. TGI is funded by the Higher Education Authority (HEA) of Ireland under the Programme for Research in Third-Level Institutions (PRTLI) Cycle 5 and co-funded under the European Regional Development Fund (ERDF). The authors also acknowledge the use of additional support and assitance provided by the European Grid Infrastructure. For more information, please reference the EGI - InSPIRE paper (http://go.egi.eu/pdnon).

REFERENCES

[1] I. Foster, C. Kesselman, and S. Tuecke, "The Anatomy of the Grid: Enabling Scalable Virtual Organizations," *Int. J. High Perform. Comput. Appl.*, vol. 15, no. 3, pp. 200–222, Aug. 2001. doi: 10.1177/109434200101500302. [Online]. Available: http://dx.doi.org/10.1177/109434200101500302

[2] S. H. Fuller and L. I. Millett, *The Future of Computing Performance: Game Over or Next Level?* The National Academies Press, 2011. ISBN 9780309159517. [Online]. Available: http://www.nap.edu/openbook.php?record_id=12980

[3] V. Kindratenko, "Computational Accelerator Term Revisted," https://www.ieeetcsc.org/activities/blog/Accelerated_Computing_computational_Accelerator_Term_Revisited, 02 2013.

[4] "Top 500 Supercomputers," http://www.top500.org/.

[5] W. G. et al, "MPI: The Message Passing Interface Standard, Version 2.2," http://www.mpi-forum.org/docs/mpi-2.2/mpi22-report-book.pdf.

[6] J. Walsh, "Message Passing Interface - Current Status and Future Developments, ," EGI Technical Forum, 2010, September 2010.

[7] J. Walsh, B. Coghlan, K. Eigelis, and G. Sipos, "Results from the EGI GPGPU Virtual Team's User and Resource Centre Administrators Surveys," in *Crakow Grid Workshop 2012*, Crakow, Poland, October 2012.

[8] J. Walsh, "Presentation of Results from the EGI GPGPU Virtual Team Surveys," https://indico.egi.eu/indico/materialDisplay.py?contribId=162&sessionId=81&materialId=slides&confId=1019.

[9] H. Stockinger, "Defining the Grid: a snapshot on the current view." *The Journal of Supercomputing*, vol. 42, no. 1, pp. 3–17, 2007. [Online]. Available: http://dblp.uni-trier.de/db/journals/tjs/tjs42.html#Stockinger07

[10] D. P. Anderson, "BOINC: A System for Public-Resource Computing and Storage," in *Proceedings of the 5th IEEE/ACM International Workshop on Grid Computing*, ser. GRID '04. Washington, DC, USA: IEEE Computer Society, 2004. doi: 10.1109/GRID.2004.14. ISBN 0-7695-2256-4 pp. 4–10. [Online]. Available: http://dx.doi.org/10.1109/GRID.2004.14

[11] "OGF GLUE 1.3 Specification," http://forge.gridforum.org/sf/go/doc14185.

[12] "OGF GLUE 2.0 Specification," http://glue20.web.cern.ch/glue20/.

[13] S. Burke, S. Andreozzi, F. Donno, F. Ehm, L. Field, M. Litmaath, and P. Millar, "The Impact and Adoption of GLUE 2.0 in the LCG/EGEE Production Grid," *Journal of Physics: Conference Series*, vol. 219, no. 6, p. 062005, 2010. doi: 10.1088/1742-6596/219/6/062005. [Online]. Available: http://stacks.iop.org/1742-6596/219/i=6/a=062005

[14] S. B. et al., "The gLite 3.2 User Guide," https://edms.cern.ch/file/722398/1.4/gLite-3-UserGuide.pdf, July 2012.

[15] "Forward of Requirements To The Batch System," http://grid.pd.infn.it/cream/field.php?n=Main.ForwardOfRequirementsToTheBatchSystem.

[16] "Tornado HomePage," http://www.tornadoweb.org/, 2014.

[17] "CUDA Wrapper SourceForge Homepage," http://sourceforge.net/projects/cudawrapper/, 2014.

[18] V. V. Kindratenko, J. J. Enos, G. Shi, M. T. Showerman, G. W. Arnold, J. E. Stone, J. C. Phillips, and W.-m. Hwu, "GPU clusters for high-performance computing," 2009. doi: 10.1109/clustr.2009.5289128 pp. 1–8. [Online]. Available: http://dx.doi.org/10.1109/clustr.2009.5289128

[19] S. Toth and M. Ruda, "Practical Experiences with Torque Meta-Scheduling in The Czech National Grid," J. Kitowski, Ed. Krakow, Poland: AGH University of Science and Technology Press, 2012, pp. 33–45.

[20] F. Vella, R. Cefala, A. Costantini, O. Gervasi, and C. Tanci, "GPU Computing in EGI Environment Using a Cloud Approach," in *International Conference on Computational Science and Its Applications (ICCSA) 2011*, 2011. doi: 10.1109/ICCSA.2011.61 pp. 150–155. [Online]. Available: http://dx.doi.org/10.1109/ICCSA.2011.61

[21] J. Breitbart and G. Khanna, "An Exploration of CUDA and CBEA for Einstein@Home," in *Proceedings of the 8th International Conference on Parallel Processing and Applied Mathematics: Part I*, ser. PPAM'09. Berlin, Heidelberg: Springer-Verlag, 2010. ISBN 3-642-14389-X, 978-3-642-14389-2 pp. 486–495. [Online]. Available: http://dl.acm.org/citation.cfm?id=1882792.1882851

[22] A. Visegrádi, J. Kovács, and P. Kacsuk, "Efficient Extension of gLite VOs with BOINC Based Desktop Grids," *Future Gener. Comput. Syst.*, vol. 32, pp. 13–23, Mar. 2014. doi: 10.1016/j.future.2013.10.012. [Online]. Available: http://dx.doi.org/10.1016/j.future.2013.10.012

[23] D. Thain, T. Tannenbaum, and M. Livny, "Distributed Computing in Practice: The Condor Experience: Research Articles," *Concurr. Comput. : Pract. Exper.*, vol. 17, no. 2-4, pp. 323–356, Feb. 2005. doi: 10.1002/cpe.v17:2/4. [Online]. Available: http://dx.doi.org/10.1002/cpe.v17:2/4

[24] "How to Manage GPUs," https://htcondor-wiki.cs.wisc.edu/index.cgi/wiki?p=HowToManageGpus, 2014.

[25] E. F. del Castillo, "Support to MPI Applications on the Grid." *Computing and Informatics*, vol. 31, no. 1, pp. 149–160, 2012. [Online]. Available: http://dblp.uni-trier.de/db/journals/cai/cai31.html#Fernandez-del-Castillo12

[26] "Puppet IT Automation Homepage," http://puppetlabs.com/, 2014.

[27] "EGI GPGPU Working Group HomePage," https://wiki.egi.eu/wiki/GPGPU-WG, 2014.

[28] "EGI Federated Cloud Task Force Homepage," https://wiki.egi.eu/wiki/Fedcloud-tf, 2014.

[29] "AMD APP OpenCL Programming Guide," http://developer.amd.com/wordpress/media/2013/07/AMD_Accelerated_Parallel_Processing_OpenCL_Programming_Guide-rev-2.7.pdf, 2014.

Education, Curricula & Research Methods

ECRM is a FedCSIS conference area aiming at interchange of information, ideas, new viewpoints and research undertakings related to university education and curricula as well as recommended methods of doing research in all computing disciplines, i.e. computer science, computer engineering, software engineering, information technology, and information systems. This area spans typical FedCSIS events (conferences, workshops, etc.) with rigorous paper submissions and review processes as well as panels, PhD and research consortia, summer schools, etc. Events that constitute ECRM are:

- DS-RAIT'14 – 2nd Doctoral Symposium on Recent Advances in Information Technology
- ISEC'14 – 3rd Information Systems Education & Curricula Workshop

3rd Information Systems Education & Curricula Workshop

ISEC goal is to promote the discussion about the convergence between Computer Science and Information Systems topics, so that researchers can present a complete and detailed specification of their educational curricula by means of these two topics. We inspect papers that contribute to the better understanding of emerging and important educational fields of Computer Science (CS) and Information Systems (IS). Authors are invited to submit their papers in English, presenting the results of original research or innovative practical applications in the field.

Regarding the selection process, we plan to perform a triple review process.

Topics

Key issues in this workshop will focus on (but are not limited to):

- Convergence between IS & CS in higher Education
- Specification of IS Education Curricula
- General IS Theory
- IS Scope
- IS Educational Fields
- Student participation in research
- Adaptation to the European Higher Education
- Assessment of students
- Innovative teaching methods
- Training for career and skills development
- Definition of "knowledge" in IS
- Quality and evaluation of teaching
- Social and environmental commitment
- Curriculum organization and curriculum
- The Fundamental Concepts Underpinning IS
- Merge from IS to CS and vice versa in higher Education

Event Chairs

Fardoun, Habib M., King Abdulazziz University, Saudi Arabia

Gallud, José A., University of Castilla-La Mancha, Spain

Tesoriero, Ricardo, University of Castilla-La Mancha, Spain

Program Committee

Abou-Tair, Dhiah el Diehn, German Jordanian University, Jordan

Aknin, Noura, Université Abdelmalek Essaadi, Morocco

Arcega, Francisco, University of Zaragoza

Atzmueller, Martin, Kassel University, Germany

Bento da Silva, Juarez, Universidade Federal de Santa Catarina

Carretero González, Lorenzo, King Abdulaziz University, Spain

Cipres, Antonio Paules, University of Castilla-La Mancha

Collazos Ordoñez, Cesar Alberto, Universidad del Cauca

Corbalan, Montserrat, Technical University of Catalonia (UPC)

De la Guía, Elena, University of Castilla-La Mancha, Spain

Garcia Zubia, Javier, University of Zaragoza

Garrido, Juan Enrique, University of Castilla-La Mancha, Spain

Giménez, Rafael, Barcelona Digital Technology Centre, Spain

Igual, Raul, University of Zaragoza

Kempin, Nils, CGI, Germany

Lambropoulos, Niki, University of Patras, Greece, Greece

Llamas Nistal, Martin, Universidade de Vigo

Majchrzak, Tim A., University of Münster, Germany

Medrano, Carlos, University of Zaragoza

Mystakidis, Stylianos, University of Patras, UOC, UW, Greece

R. Penichet, Víctor M., University of Castilla-La Mancha, Spain

Restivo, Teresa, Universidade do Porto

Romero Lopez, Sebastian, PSJA Southwest Early College High School

Tambo, Erick, United Nations University, Germany

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 795–802
DOI: 10.15439/2014F300
ACSIS, Vol. 2

Flipped Computer Science Classes

R. Robert Gajewski
Warsaw University of Technology
Faculty of Civil Engineering
Al. Armii Ludowej 16, 00-637
Warszawa Poland
Email: rg@il.pw.edu.pl

Marcin Jaczewski
Warsaw University of Technology
Faculty of Civil Engineering
Al. Armii Ludowej 16, 00-637
Warszawa Poland
Email: mjacz@il.pw.edu.pl

***Abstract*—Computer Science (CS) classes have been supported for ten years by multimedia materials in the form of podcasts. Even such materials did not improve quality of learning outcomes. In order to change this the idea of flipped classroom (FC) was used but situation did not change radically, so survey concerning students opinion on FC was conducted. Comparison of results of survey show big differences in attitude to FC and learning between students in Poland and students in United States or Canada. Conclusion of this research is sad—it is very difficult to motivate Polish digital natives to learn..**

I. Introduction

IDEA of inverting education is already nearly fifteen years old. One of the first papers in that field was published in 2000 [1]. This paper describes two parts of subject taught at Miami University while using the inverted classroom concept and analyzes the outcomes. Numerous technologies offered completely new possibilities for students to learn away from the classroom, while school period was used to perform collaborative experiments and worksheets. Authors of the paper concluded that the idea of inverted classroom offers alternatives for various learning styles and report that students favor that strategy. A different outline and evaluation of flipped education within a huge, primarily based on lectures, computer science course was published in 2002 in [2]. In this project new multimedia and video streaming application eTech was employed to change a course. In-class lectures were substituted by recorded lectures and auxiliary materials which could be viewed by students in the Internet independently. This make it possible to utilize the live period in the class for team problem solving facilitated by tutors. Another interesting paper in that field was published one year later in 2003 [3]. Within a series of five experiments hundreds of students from two different universities supervised by three different professors and six different teaching assistants took one semester long course in the field of casual and statistical reasoning in both traditional or online format. Within the frame of this project pre and post test results were compared. Features of the online experience which were helpful and which were not helpful were identified as well as most and least effective student learning strategies. Three years later a paper evaluating a web lecture intervention in a human–computer interaction course was published [4]. By utilizing lectures available in the Web before class more in-class period was used engaging students with hands-on tasks. Class time was spent on learning by doing rather than learning by listening. In 2007 Gannod presented his work in progress on how to use podcasts in an inverted classroom [5]. One year later Helmick presented integrated online courseware for computer science courses [6]. Last but not least in 2008 a paper describing how to use the inverted classroom to teach software engineering was published [7].

The present paper summarizes results of investigation presented in [8] and [9]. Starting from academic year 2012-2013 in some of groups podcasts (mainly software animations in the form of screencasts) were used in different way. Students were asked to watch podcasts at home. During classes students were supposed to be prepared to use software without any difficulties and to solve using it typical problems. First results of this experiment were to some extend promising. Students gained better scores in this case, but they were not very keen to spend additional time at home watching podcasts. They do prefer to be taught during classes. This problem was easily solved by adding simple point to subject regulations – students should be prepared to computer laboratories and this fact is checked by means of test before the class. In academic year 2013-2014 idea of flipped classroom was used for all groups – nine studying in Polish and two studying in English but only for part of material covering Computer Algebra System MathCAD Prime 3.0. Result of survey concerning students' satisfaction will be presented in the paper.

II. Flipped Classroom

It is much more effective to watch, passive by nature, screencasts at home and solve problems with tutor in class than the other way round. This observation leads to idea to revert the situation. Why not to ask students to perform

This work was supported by Warsaw University of Technology and grant no. 504/00713/1088/40.000103

easier tasks at home and learn from podcasts independently and why not to solve during classes more difficult problems. Such situation is in agreement with Bloom's Revised Taxonomy [10], [11].

Blooms Taxonomy proposed in 1956 by panel of educators chaired by Benjamin Bloom is a categorization of learning objectives as well as activities split up into three areas: cognitive (mental skills, knowledge), affective (feelings, emotional areas and attitude) and psychomotor (manual and physical skills). The cognitive domain, the most significant in higher education, requires mental abilities and also knowledge. Within this domain one can find six major categories outlined from the most straightforward one: knowledge, comprehension, application, analysis, synthesis and finally evaluation. In the middle of 1990's cognitive domain has been modified. Titles associated with different types have been transformed from nouns to verbs. Moreover their order has been somewhat changed. Bloom's Revised Taxonomy demonstrates to a greater extent active way of thinking and also consists of six different categories: remembering, understanding, applying, analyzing, evaluating and finally creating. This taxonomy much better accounts for completely new behaviors and multimedia technology innovations (Fig. 1). Lower order thinking skills like remembering, understanding and applying are gained at home from podcasts which can be treated as recorded classes. Higher order thinking skills like analyzing, evaluating and creating are gained at the university. Such situation requires change of the role of academic staff – from teachers to tutors. This idea was fully described in three books recently published by Bergmann and Sams [12], Bretzmann [13] or Walsh [14].

One of the best definitions of flipped class is given by Bergmann, Overmyer and Willie in The Daily Riff entitled The Flipped Class: Myths vs. Reality. "The traditional definition of a flipped class is: where videos take the place of direct instruction; this then allows students to get individual time in class to work with their teacher on key learning activities; it is called the flipped class because what used to be classwork (the "lecture" is done at home via teacher-created videos and what used to be homework (assigned problems) is now done in class."

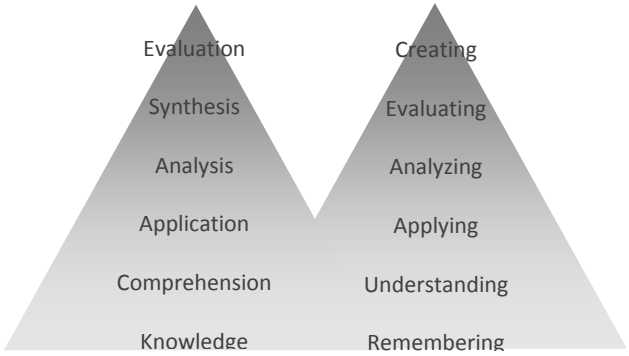


Fig. 1 Blooms Taxonomy and Revised Blooms Taxonomy

They also answered to the question what the flipped Classroom is. "A means to increase interaction and personalized contact time between students and teachers. An environment where students take responsibility for their own learning. A classroom where the teacher is not the "sage on the stage", but the "guide on the side". A blending of direct instruction with constructivist learning. A classroom where students who are absent due to illness or extra-curricular activities such as athletics or field-trips, don't get left behind. A class where content is permanently archived for review or remediation. A class where all students are engaged in their learning. A place where all students can get a personalized education."

III. Computer Science in Civil Engineering

The place and role of Information Systems (IS), which can be treated in narrow sense as a term referring mainly to ICT, and Computer Sciences (CS) in curricula of Civil Engineering (CE) studies was described in previous paper [9]. Hardware and software information revolution has changed radically modern engineering workplace. The exact description of how transformation and circulation of information in the construction industry can and should look like can be found in [15]. In this work there are distinguished three groups of information: about the function, about the structure and about structure's behavior (Fig. 2). In all stages of transformation, it is important to use computer tools. The process of analyzing the structure is invariably dominated by computer programs using the finite element method. In the process of synthesis, where there is room for optimization, there are also available computational tools, such as for example a spreadsheet Solver.

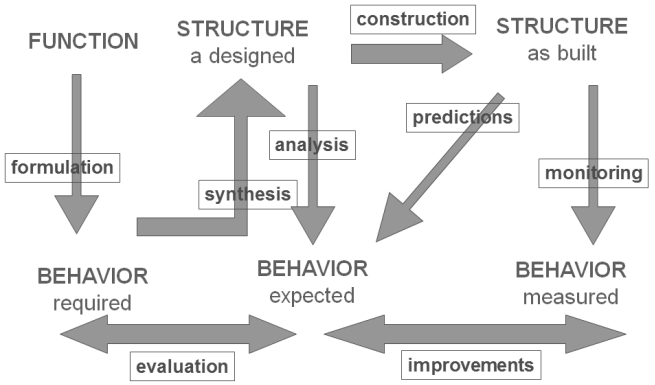


Fig. 2 Information flow in Civil Engineering

The last and the biggest block of classes is devoted to Computer Algebra System (CAS) namely to MathCAD Prime 3.0 which is described in books like [15] and [16]. Previous version of this program is better described in literature [17], [18] and [19]. Its presence in curricula of studies is a source of never ending discussions. In the opinion of many teachers students overuse MathCAD while preparing their design homework. It is enough that one

person creates a file and all remaining can simply enter only the input data.

First part of MathCAD classes is devoted to solving classical mathematical problems like symbolic calculations, defining variables and functions, calculus (integrals, derivatives, limits), matrix and vector operators and functions and solving problems like linear and nonlinear equations, minimization and maximization. Second part is devoted to programming. In the beginning basic instructions (if, for, while) and control statements (return, continue, break) are introduced. The idea of this subject was inspired by the book [20]. Detailed list of screencasts for that part of the subject is in Table I.

All course materials are stored on learning platform Moodle. Taking into account different learning styles [21], [22] learning materials are available in different forms ranging from PDF files to screencasts (Fig. 3). The vast majority of materials are in the form of screencasts which are recordings of traditional classes (Fig. 4).

IV. Survey

The research concerning students' satisfaction with flipped classroom was conducted in academic year 2013/2014 on a group of 222 students studying in Polish (PL) and a group of 51 students studying in English (EN). Out of 222 PL students the questionnaire was filled by 211 students. Similar data are for students studying in English. Questionnaire was filled by 49 out of 51 students. One third of students studying in English were foreigners.

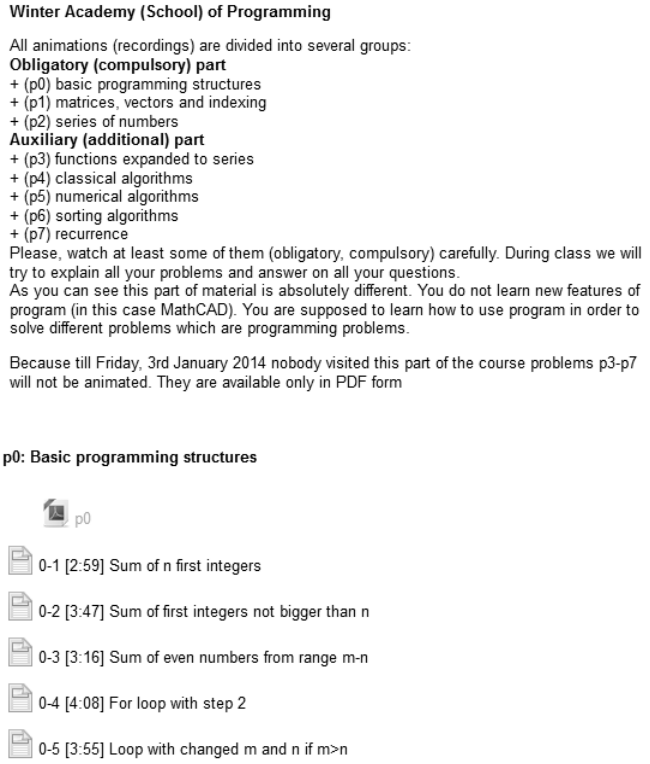


Fig. 3. Sample Learning Materials on Moodle platform

Four weeks of classes during which material covering text editor Word and spreadsheet Excel was discussed were

TABLE I.
DETAILED TIMETABLE OF MATHCAD PRIME 3.0 CLASSES

Week 1 – two hours	Week 2 – two hours	Week 3 – two hours	Week 4 – two hours
MathCAD Window MathCAD ribbon Customizing worksheet Text and graphic regions Math region Grouping and formatting Symbolic calculations Simplifying expressions Expanding expressions Factoring expressions Collect keyword Coeffs keyword Substitute keyword Partial fractions Series	XY plots Formatting XY plots Range variables and XY plots Parametric plots Contour plots Formatting contour plots Arrays and tables Creating arrays Contour plots for scattered data 3D plots	Linear equations - matrix and lsolve Linear equations - solve block Nonlinear system of equations - solve block Nonlinear system of equations - polyroots Finding roots Parameterizing solve block Optimizing functions Optimizing with constraints Distance between two curves Solving ODEs with solve block	<u>Basic programming structures</u> Sum of n first integers Sum of first integers not bigger than n Sum of even numbers from range m-n For loop with step Loop with changed m and n if m>n Loop with step k+2 or -2 Testing different solutions (loops) Not nested if and wrong nesting Properly nested if
Identifiers Defining variables Defining functions Units and label Range variables Derivatives Integrals Limits Sums and products Complex variables and functions	Generation of arrays with if Matrix operators Matrix functions (1) Matrix functions (2) Linear equations File access - output File access - input Curve fitting (1) Curve fitting (2) Keyword explict Function root	Creating a program Defining functions Using operators Writing if statements Writing if - else if statements Function for different ranges Loop with for Loop with while Structure try on error Recursion	<u>Matrices, vectors and indexing</u> Minimum and maximum element in vector Minimum and maximum in one function Minimum element and its index Sum of even numbers in vector <u>Series of numbers</u> "Theory" of series of numbers Series with for and while loops Slow convergence series Alternative stop condition Alternatives and art of programming

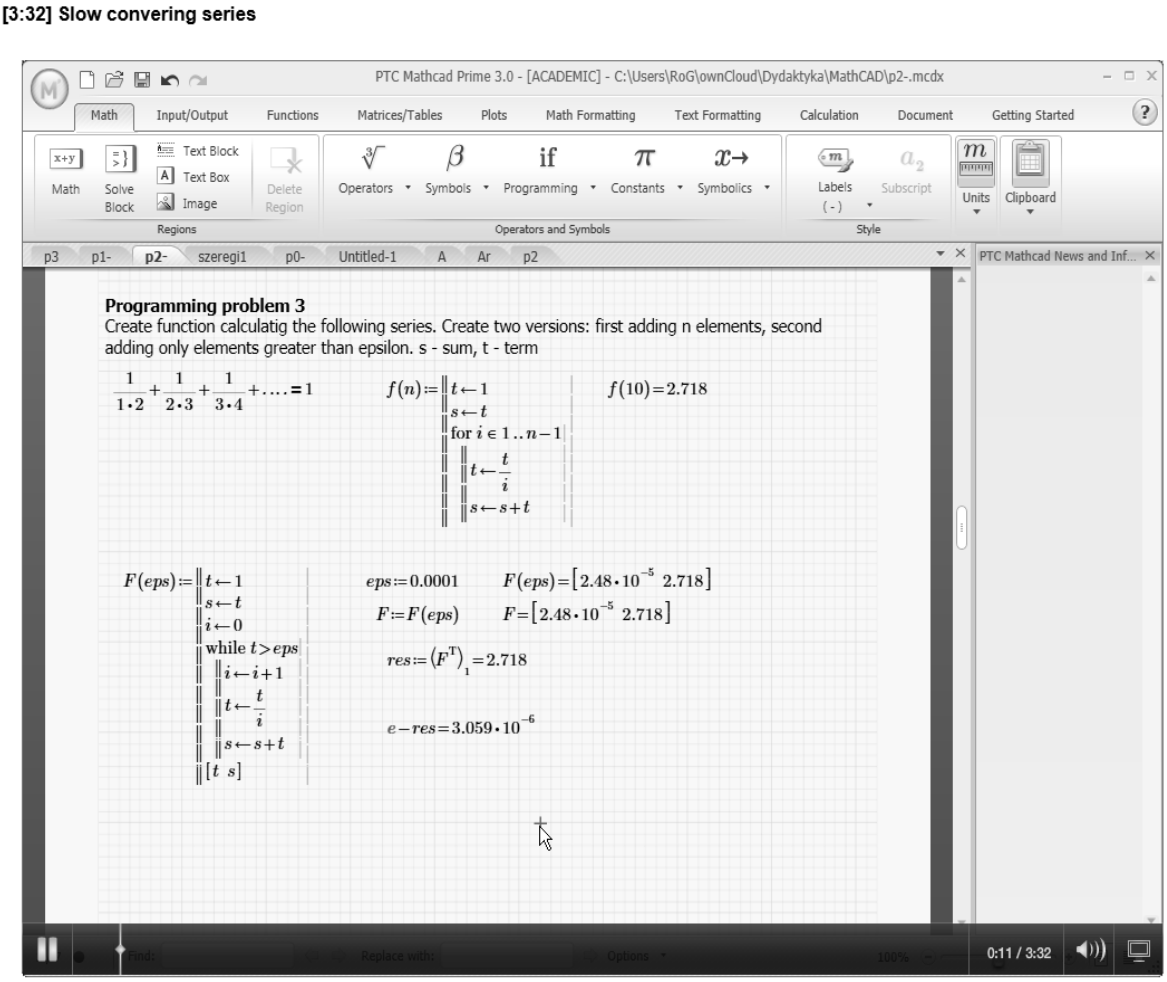


Fig. 4. Sample Screencast – Slow Converging Series

carried out in the traditional manner. In the computer laboratory equipped with 30 workstations there were two teachers. One of them demonstrated with multimedia projector solutions of the sample problems. The second teacher assisted students.

In order to perform survey Google Docs were used (Fig. 5, Fig. 6).

Fig. 5. Survey for Polish language students

Questionnaire used in this survey consists of fifteen closed form questions and 6 opened form questions. Due to the nature of answers all questions were divided into three groups. In order to compare results of survey with other outcomes some of the questions were based on similar surveys: first one conducted in Canada [27] and second one described in blog Flipping with Kirch conducted by Mary Kirch from United States.

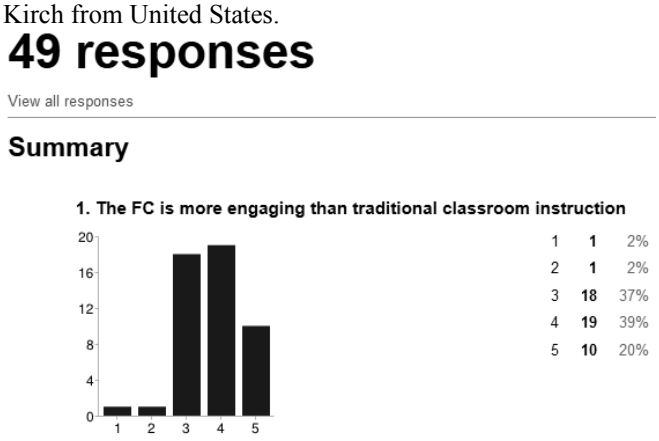


Fig. 6. Survey for English language students

Observations from traditional classes were rather pessimistic. More than half of the students did not follow the presentation and did not perform similar tasks to those presented by the teacher. As presented in previous article [9] this material was rather unknown for students but they were simply not interested in learning new things. It is rather difficult to motivate digital natives to learn [23]. This is still a very important and significant problem even there were many books written on that subject [24]. The answer to this question is difficult when we consider digital natives who don't care and who also think that they know everything [25] in the field of subjects like Applied Computer Sciences and Computing in Civil Engineering. Mendler presents one of the existing solutions – five key processes that motivate: emphasizing effort, creating hope, respecting power, building relationships and expressing enthusiasm. But digital natives being real partners for learning [26] are difficult and demanding partners.

V. Results of Survey

Scale of answers for all first five questions is from "strongly agree" to "strongly disagree". Results for Polish language and English language students were compared with surveys from Canada. First of the asked questions was about level of engagement in traditional classroom instruction and flipped classroom (Fig. 7).

40% of students studying in Polish language strongly disagree or disagree with the statement what is in accordance with the observation, that nearly half of the students was not interested in traditional classes. Answers of students studying in English language are closer to the answers from survey conducted in Canada.

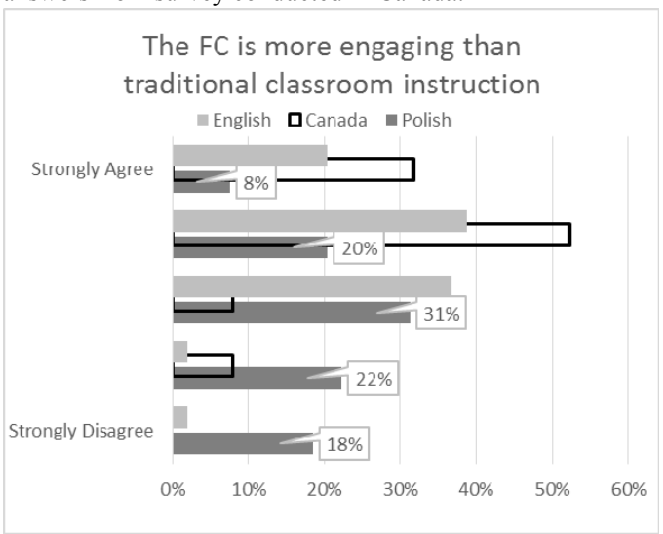


Fig. 7. Answers on question 1.1

Second question from that group was about potential recommendation of flipped classroom to a friend (Fig. 8).

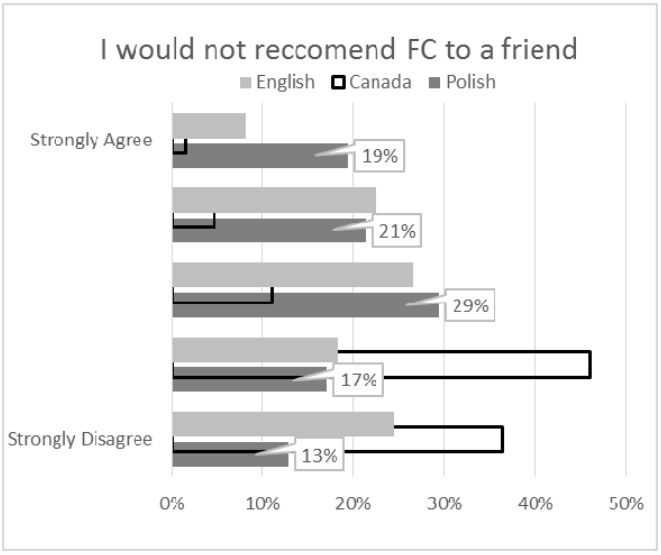


Fig. 8. Answers on question 1.2

For this question answers of students studying in Polish and English languages are similar but they definitely differ from the results of survey conducted in Canada. Nearly six times more students studying in Polish language in comparison to Canadian agree or strongly agree with the statement that they would not recommend flipped classroom to a friend.

Next question (statement) was very simple – I like watching lessons on video (Fig. 9). In this case answers for all three groups were very similar.

Fourth question in this group of questions was about bigger motivation to learn in the flipped classroom mode (Fig. 10).

In the case of this question answers of students studying in Polish language differ from the answers of two other groups. Nearly 40% of them strongly disagree or disagree with that statement that they are more motivated to learn in flipped classroom mode.

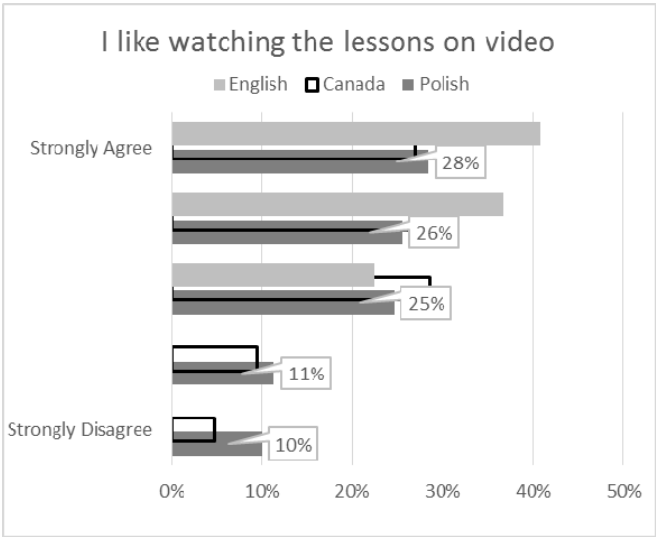


Fig. 9. Answers on question 1.3

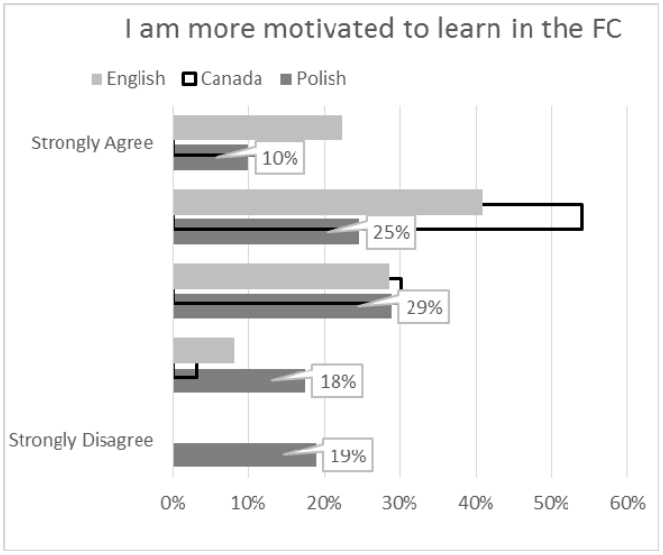


Fig. 10. Answers on question 1.4

The last question in this group is about improvement of learning in flipped classroom mode (Fig. 11).

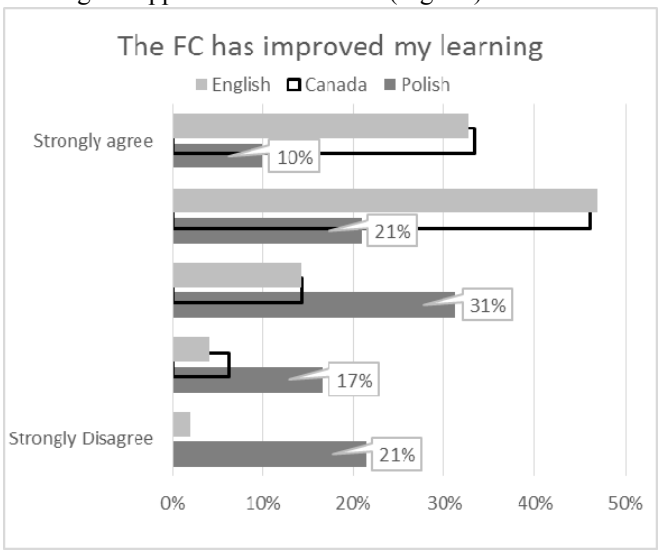


Fig. 11. Answers on question 1.5

Also in the case of this question answers obtained for students studying in Polish language are definitely different from results for two other groups. Nearly 40% of them strongly disagree or disagree with that statement. Answers for students studying in English and answers from survey in Canada are nearly the same. Nearly 80% agree or strongly agree with statement that flipped classroom has improved learning.

Second group of five questions is based on research conducted by Mary Kirch. Also in this case answers are on scale but they differ from question to question.

First question in this group is about rating flipped classroom. Scale of answers is from “very bad” to “very good” (Fig. 12).

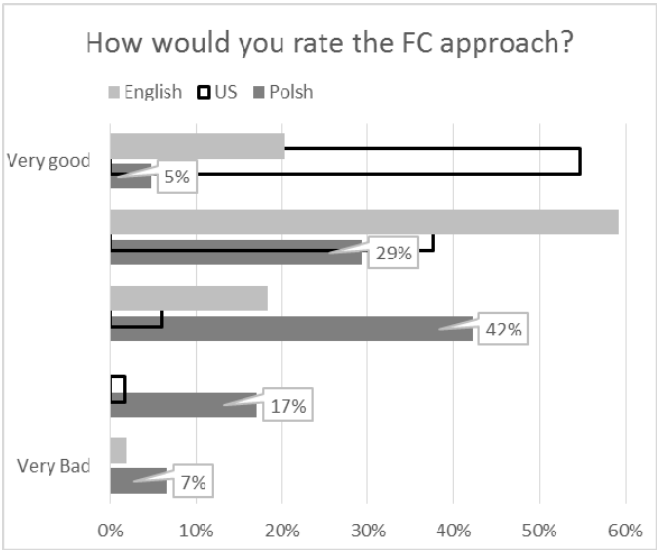


Fig. 12. Answers on question 2.1

Also in the case of this question answers for all groups are different. American students are most enthusiastic – more than 90% rated flipped classroom approach as very good or good. Students studying in Polish are definitely less enthusiastic and more skeptical – nearly 25% of them rated flipped classroom approach as very bad or bad.

Next question in this group compared feelings how flipped classroom helps to learn the material in comparison to traditional approach (Fig. 13).

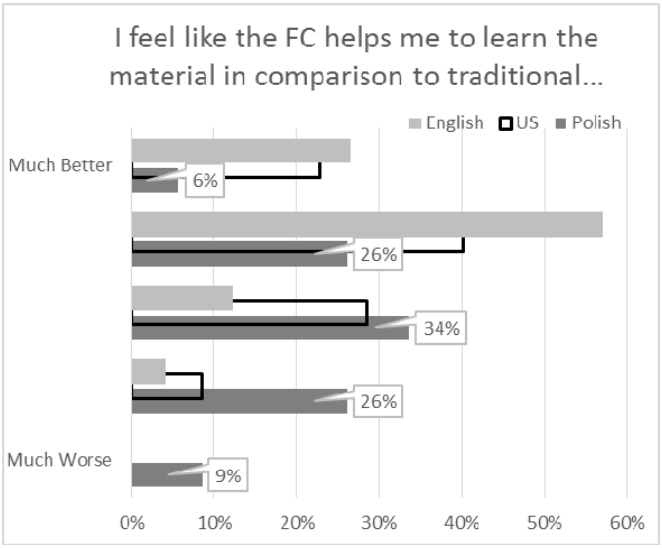


Fig. 13. Answers on question 2.2

Answers to this question are to some extend surprising. Number of answers much better or better is among students studying in English the highest – nearly 85% of such answers. Traditionally students studying in Polish gave the highest number of negative answers – 35% of them answered that flipped classroom helped them to learn in comparison to traditional approach much worse or worse.

Fourth question was about feelings towards flipped classroom approach. Scale of the answers was from “I hate it” to “I love it” (Figure 14). Results in this point differ also

because American scale was four point scale without neutral answer. Students were supposed to express only negative or positive feelings.

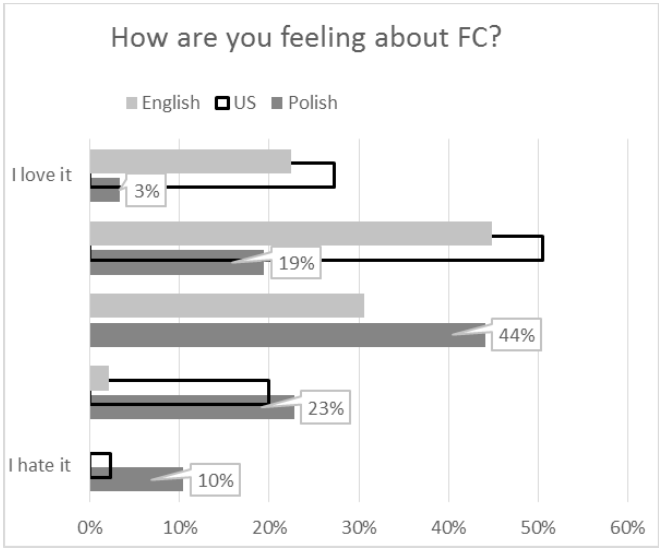


Fig. 14. Answers on question 2.3

As in the case of previous questions for students studying in Polish language the number of definitely positive answers I love it is the lowest – only 3%.

In next point question how much did you learn in flipped classroom in comparison to traditional was asked (Fig. 15).

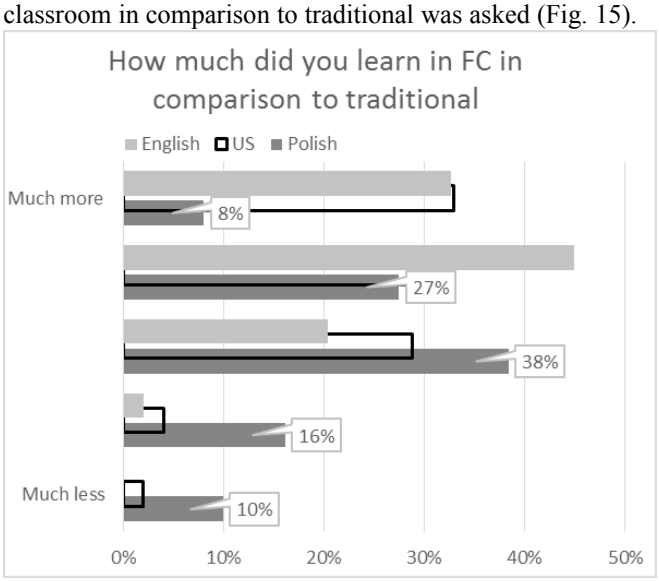


Fig. 15. Answers on question 2.4

More than 25% of Students studying in Polish language answered that they learned much less or less while more than 30% of students studying in English language or American students answered that they learned much more.

The last question was: how much were you challenged as a student in flipped classroom comparing to traditional one (Fig. 16).

Results for all three groups are in the case of this question very similar.

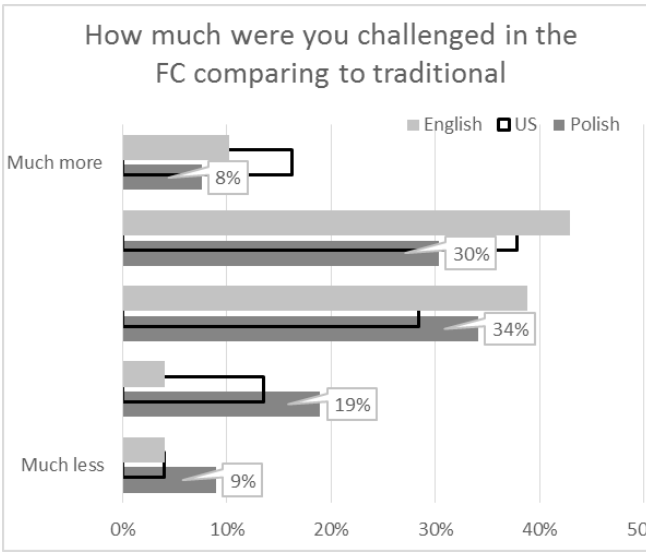


Fig. 16. Answers on question 2.5

I. Final Remarks

One of the pressures on universities is the rapid development of new information and communication technologies for the provision of education and training. Wide opportunities in open and distance learning create new markets. Moreover the principle of life-long learning extends the age groups to which university can offer education. Additionally the principles of new techniques can be applied to traditional markets – regular intramural students. All Polish universities willing to use modern information and communication technologies face common opportunities, threats and constraints. A constant struggle between pressure to change and fear and resistance to change is visible in Poland. Teachers' attitudes are a major obstacle to the introduction of change. There is an internal brake on the efforts to make changes through using new technologies: resistance from people. Reference can be made to a "frozen middle" resisting attempts to change from both the top of the institutions (authorities) and from the bottom (students). Students' demands are a powerful factor forcing universities to exploit the potential of new technologies to improve learning experience. But the question "how to change the unchanging" is still open [28]. Moreover one should add to this question a new one – "how to motivate digital natives to learn?" [23].

Flipping the classroom is not the only way in which "how to change" and "how to motivate" problems can be solved. One of them is classification of learners using linear regression [29], because proper classification of learners is one of the key aspects in e-Learning environments .Second possibility is to enhance online educational environment what can be obtained by providing platform side intelligent functionalities in the shape of a recommender system or learning path builder [30]. Another possibility is to use system that allow the organization of a set of teaching and learning activities and meets, at its highest level of detail all the elements that compose the curriculum: setting goals and content, design and development of activities and

evaluation, organization of space and time, and providing the necessary resources [31]. Next chance is to implement a Learner-centered learning which is constructivism based and Competence directed in which we define general competencies, domain competencies and specific course competencies [32]. Last but not least there are also another innovative teaching methods for blended learning like for example Drawer [33] or recognizing different learning styles while designing e-courses [34].

ACKNOWLEDGMENT

The authors wish to thank L. Wlasak, T. Dubilis and T. Warda for their support during flipped classroom experiment and survey. Thanks are also due to Dean of Faculty of Civil Engineering Professor H. Zobel who allowed us to conduct this educational experiment. Last but not least the authors wish to thank all students who filled very long questionnaire at the end of semester.

REFERENCES

[1] M. J. Lage, G. J. Platt, and M. Treglia, "Inverting the Classroom: A Gateway to Creating an Inclusive Learning Environment," *Journal of Economic Education*, vol. 31, no. 1, pp. 30–43, 2000. http://dx.doi.org/10.1080/00220480009596759

[2] J. Foertsch, G. Moses, J. Strikwerda, and M. Litzkow, "Reversing the Lecture/Homework Paradigm Using eTEACH® Web-based Streaming Video Software," *Journal of Engineering Education*, vol. 91, no. 3, pp. 267–274, Jul. 2002. http://dx.doi.org/10.1002/j.2168-9830.2002.tb00703.x

[3] R. Scheines, J. Smith, G. Leinhardt, and K. Cho, "Replacing lecture with Web-based course materials," *Journal of Educational Computing Research*, vol. 32, pp. 1–26, 2003.

[4] J. A. Day and J. D. Foley, "Evaluating a Web Lecture Intervention in a Human-Computer Interaction Course," *IEEE Trans. on Educ.*, vol. 49, no. 4, pp. 420–431, Nov. 2006. http://dx.doi.org/10.1109/TE.2006.879792

[5] G. C. Gannod, "Work in progress; Using podcasting in an inverted classroom," in *Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE '07. 37th Annual*, 2007, pp. S3J–1–S3J–2. http://dx.doi.org/10.1109/FIE.2007.4418119

[6] M. T. Helmick, "Integrated Online Courseware for Computer Science Courses," in *Proceedings of the 12th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education*, New York, NY, USA, 2007, pp. 146–150. http://dx.doi.org/10.1145/1268784.1268828

[7] G. C. Gannod, J. E. Burge, and M. T. Helmick, "Using the Inverted Classroom to Teach Software Engineering," in *Proceedings of the 30th International Conference on Software Engineering*, New York, NY, USA, 2008, pp. 777–786. http://dx.doi.org/10.1145/1368088.1368198

[8] R. Gajewski, "Towards a New Look at Steaming Media," in *WCCE 2013 10th IFIP World Conference on Computers in Education*, vol. 2: Practice Papers, Torun: IFIP, 2013, pp. 98–103.

[9] R. Gajewski, L. Wlasak, and M. Jaczewski, "IS (ICT) and CS in Civil Engineering Curricula: Case Study," in *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems*, Krakow: IEEE, 2013, pp. 717–720.

[10] B. S. Bloom, *Taxonomy of Educational Objectives Book 1: Cognitive Domain*, 2nd edition. Addison Wesley Publishing Company, 1984.

[11] L. W. Anderson, D. R. Krathwohl, P. W. Airasian, K. A. Cruikshank, R. E. Mayer, P. R. Pintrich, J. Raths, and M. C. Wittrock, *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, Abridged Edition*, 2nd ed. Pearson, 2000.

[12] J. Bergmann and A. Sams, *Flip Your Classroom: Reach Every Student in Every Class Every Day*. International Society for Technology in Education, 2012.

[13] J. Bretzmann, *Flipping 2.0*. Bretzmann Group LLC, 2013.

[14] K. Walsh and P. J. Walsh, *Flipped Classroom Workshop in a Book*, 1 edition. Kelly Walsh, 2013.

[15] H. Wesselingh and H. de Waard, *Calculate & Communicate with Mathcad Prime*. Delft: VSSD, 2012.

[16] B. Maxfield, *Essential PTC® Mathcad Prime® 3.0: A Guide for New and Current Users*, 1 edition. Academic Press, 2013.

[17] R. W. Larsen, *Introduction to Mathcad 15*, 3rd ed. Prentice Hall, 2010.

[18] B. Maxfield, *Essential Mathcad for Engineering, Science, and Math, Second Edition*, 2nd ed. Academic Press, 2009.

[19] P. Pritchard, *Mathcad: A Tool for Engineering Problem Solving + CD ROM to accompany Mathcad*, 2nd ed. McGraw-Hill Science/Engineering/Math, 2011.

[20] D. Harel and Y. Feldman, *Algorithmics: The Spirit of Computing*, 3rd ed. Addison-Wesley, 2004.

[21] F. Romanelli, E. Byrd, and M. Ryan, "Learning Styles: A Review of Theory, Application, and Best Practices," *American Journal of Pharmaceutical Education*, vol. 73, no. 1, pp. 1–5, 2009.

[22] C. Scott, "The Enduring Appeal of 'Learning Styles,'" *Australian Journal of Education*, vol. 54, no. 1, pp. 5–17, 2010. http://dx.doi.org/10.1177/000494411005400102

[23] G. B. Johnson, "Student Perceptions on the Flipped Classroom." The University of British Columbia, 2013.

[24] L. Wlasak, M. Jaczewski, T. Dubilis, and T. Warda, "How to Motivate Digital Natives to Learn?," in *WCCE 2013 10th IFIP World Conference on Computers in Education*, vol. 3: Book of Abstracts, Torun: IFIP, 2013, pp. 78–79.

[25] J. E. Brophy, *Motivating Students to Learn*. Routledge, 2010.

[26] A. Mendler, *Motivating Students Who Don't Care: Successful Techniques for Educators*. National Educational Service, 2000.

[27] M. R. Prensky, *Teaching Digital Natives: Partnering for Real Learning*. Corwin, 2010.

[28] R. R. Gajewski, "How to change the unchanging? Restructuring Polish universities for the XXI century," in *TelE-learning, The Challenge for the Third Millennium*, 2002, pp. 297–300.

[29] C. Mihaescu, "Classification of Learners Using Linear Regression," in *Proceedings of the 2011 Federated Conference on Computer Science and Information Systems*, 2011, pp. 717–721.

[30] M. C. Mihăescu, "The Design of eLeTK – Software System for Enhancing On-Line Educational Environments," in *Proceedings of the 2012 Federated Conference on Computer Science and Information Systems*, 2012, pp. 739–744.

[31] A. P. Cipres, H. M. Fardoun, and A. Mashat, "Cataloging Teaching Units: Resources, Evaluation and Collaboration," in *Proceedings of the 2012 Federated Conference on Computer Science and Information Systems*, 2012, pp. 825–830.

[32] J. Schreurs and A. Al-Huneidi, "Design of Learner-Centered constructivism based Learning Process," in *Proceedings of the 2012 Federated Conference on Computer Science and Information Systems*, 2012, pp. 1159–1164.

[33] F. A. Marco, V. M. R. Penichet, and J. A. G. Lázaro, "Drawer: an Innovative Teaching Method for Blended Learning," in *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems*, 2013, pp. 727–734.

[34] O. Mironova, T. Rüütmann, I. Amitan, J. Vilipõld, and M. Saar, "Computer Science E-Courses for Students with Different Learning Styles," in *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems*, 2013, pp. 735–738.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 803–808
DOI: 10.15439/2014F398
ACSIS, Vol. 2

New Teaching Methods: Merging "John Dewey" and "William Heard Kilpatrick" Teaching Techniques

Habib M. Fardoun, Abdullah Almalaise Alghamidi
Information Systems Department
King Abdulaziz University (KAU)
Jeddah, Saudi Arabia
Email: {hfardoun, aalmalaise}@kau.edu.sa

Antonio Paules Ciprés
European University of Madrid,
Madrid, Spain
Email: apcipres@gmail.com

***Abstract*—This article shows our research oriented towards improving doctoral theses by publications. This process is justified by John Dewey's knowledge theory and various phases of Kilpatrick's work. After a theoretical foundation, the research work set the steps in a thesis by publications, which is both validated by members of the scientific community and created from the knowledge of the experience of the candidate in the field of study. Validation of the PhD thesis in this case is found by members of the scientific community in a research validation along publications.**

I. Introduction

THIS article proposes a didactic technique using John Dewey's knowledge theory as the starting point. With experience as key factor, theory and practice have a strong relation and focus education on an experimental and pragmatic approach.

Pragmatism is a school of philosophy originated in the United States in the late 19th century. The term comes from the Greek work praxis, which means action. Pragmatism is not really a philosophical theory but a "mindset" that includes different theories that can be applied to different disciplines. Pragmatism is also regarded as a theory of human beings from the cognitive perspective.

John Dewey's theory of knowledge [1] in conjunction with William Heard Kilpatrick's technique of project-based work produces an improvement of the teaching-learning process.

The combination of both is a different approach to the project-based method from many perspectives: epistemological, psychological, educational purpose, and curriculum development. This article shows how to apply the project-based work to doctoral theses by publications, namely theses composed of a compendium of publications, where there is a project-based work as well as a team that supports the PhD student. By applying a didactical methodology, the study tries to get PhD work and students closer.

A doctoral thesis is usually an original idea from the author. This idea is a consequence of a series of individual and smaller ideas coming from other authors involved in previous scientific researches, with a set of problems to solve.

The development of a thesis then involves the creation of a team aimed at solving a problem using a project-based approach. This team is often comprised of tutor, co-director, supervisor, collaborators and the student. This makes the final results are validated both by the team and the publications, which are a form of validation.

This approach is innovative and unusual, but we think it is necessary to begin the development of scientific research methodologies closer to reality, as teamwork in an interdisciplinary thesis is common. Also, there is a constant exchange of ideas among scientists through published articles, workshops and forums that are a way to provides both corrections as validating the research since its inception.

II. Objective

The aim of this article is to begin research and start a discussion on how to make a thesis by publications around ICTs in education. The starting point is knowledge methods of recognized educators, which are the source of the didactic-scientific methodology nowadays.

The team based the research on its previous three-year period of work on educational curricular organization and classroom methodologies. This experience was vital to develop this research and create a methodology for theses by publications, targeted at the application of theory to the productive, research or corporate sectors, thus creating theses applied to real environments such as industry, education or health.

III. State of the art

John Dewey sees education as the sum of processes by which a community or social group transmits its acquired powers and objectives, to ensure its own existence and its continued growth. He built an entire pedagogical philosophy with experience as a base, defining education as "a reconstruction or reorganization of experience, which gives meaning to experience and increases ability to guide subsequent experiences" [3].

The main element related to Dewey's theory of knowledge is the concept of experience. This concept is also the most important of his theory. Dewey proposes defines experience as a dynamic concept, an exchange between humans and

978-83-60810-58-3/$25.00

their physical and social environments, not simply a matter of knowledge. Moreover, experience entails a series of actions and affections, and cannot therefore be as a simply subjective term. Dewey states the experience is based on connections and continuities, implying processes of reflection and inference and linking experience and thinking [4].

Summarizing Dewey's proposal, the key element of knowledge and also the most important is a progressive education. The general method is an aim-driven action that takes into account activities such as artistic activities, past, tradition, and techniques that development that activity. The management of the scientific method must be guided by scientific knowledge. Because of this, Dewey believes educative methods must derive from the scientific method, adding any necessary adaptations. This is the so-called "Dewey method" or "problem method", a sequential process used to outline learning as a research activity. With this method, learning becomes a chapter in the general research method [5], comprised of five stages [6]:

- The student has a situation of authentic experience. In other words, there is a continuous activity in which the student is interested.
- A real problem arises in this situation as a stimulus for thought.
- The student posses the information and performs necessary observations to tackle the problem.
- Suggested solutions make the student realize she is responsible for developing them in an organized way.
- The student has the opportunity and the chance to test his ideas for their application, to clarify its meaning and to discover their validity.

Alternatively, Kilpatrick created the following method that "arises from a proposal made for the purpose of teaching the Dewey's concept of thought. This method became a more concrete way of acting in the field of practice than in theory [7]:

- First stage: students choose a topic from a set proposed by the teacher.
- Second stage: once students reveal their interest, we must assess their prior knowledge about the topic. This phase is called "What Do We Know?"
- Third stage: the following question is asked: "What do we want to know?" The answer is objectives and contents the students should acquire through analysis and research on the chosen topic.
- Fourth stage: "How can we know it?" phase or "investigate and learn." Here we must take into account the elements of the methodology (organization of space and time, materials and resources to use, people who will help us, sources of information and activities).

• Fifth stage: "What Have We Learned?". This stage corresponds to the evaluation. Students tell about what they liked most, the least, specially difficult tasks... and will reflect our work in a final document to provide greater value to our work (reports, letters, drama, book project) that will be stored in the classroom library.

The project-based method leads to a significant and globalizer learning process where students assume more responsibility for the own learning and use acquired skills and knowledge in real problems. We therefore need to take into account the theory of significant knowledge by David Ausubel [8], which states that significant knowledge occurs when new information "connects" with relevant concept a ("subsunsor") that already existed in the cognitive structure. Thus, new ideas, concepts or relevant propositions can be meaningfully learnt, as others are clear and available in the cognitive structure of the individual, working as an anchor.

Once we have justified the cognitive perspective of project-based work, with Dewey's theories of knowledge and Kilpatrick's project-based work, we need to evaluate the research methodology in order to adapt it to research on the use of ICTs in classroom education.

IV. Thesis Methodology: summary of publications

At the point of state of the art we have developed the necessary starting points for this item, we will develop as they could be the steps for developing a thesis by publications, adapting the method steps for projects Kilpatrick, is a methodology for creating and decorating that could be adapted to their needs, we must also note that we have to take into account the meaning of a thesis by publications where doctoral dissertation develop through the articles achievement of previously set objectives.

The figure presented below find the different phases of the period of creation of the thesis, there are five stages in white, these phases are designed to obtain the different phases that are the definition, validation of ideas, the specific development, application of the results and final phase of integration and evaluation of results. The blue line found the feedback from the doctoral thesis of companies, researchers conducting a review of exports and the items made. These sections in conjunction form a series of phases, as described below.

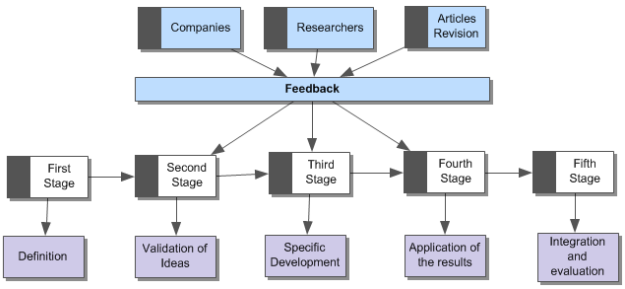


Fig. 1 Stages

A. First stage

In this first phase the doctoral student must select the subject of the thesis, this process must be from the starting point of the knowledge of the candidate, director of the thesis proposes possible research. In this first phase of the doctoral student must submit inquiries to define their own line of research. This period is a period of reflection, which must be accompanied by a practical part, this practical part must include a period of experimentation must also include scientific methodologies a practical part that is as close as possi-

ble to the productive sector which will intended research. This PhD received feedback both the method and the specialized staff working in these sectors may thus incorporate new elements of judgment and experience by providing both solutions and establish the objectives, hypothesis and problems.

We are therefore faced with the definition of the research of the thesis as we see include the problems that you are encountering in the productive sectors, this is a way of bringing doctoral reality and allow the implementation of doctoral thesis the productive sectors of the industry, thus bringing research computer schools, both companies like the different sectors, since the computer is a highly interdisciplinary and cross science, in which the approach to the problems from companies improve the concept of applied computer science.

The starting point for the hypothesis may be that hypotheses are possible solutions to the problem are expressed as generalizations or propositions. These are statements that consist of elements expressed as an ordered system of relationships, which are intended to describe or explain conditions or events not yet confirmed by the facts. [9].

The PhD student may pose for a series of scientific articles that should culminate in the precise definition of objectives that will develop and what the scope of the investigation.

In this first part of the doctoral student should perform a series of articles in order to ensure the research and thus allow other members of the scientific community to review the established line, this series should end with the completion of A final item that we could call "Phd Thesis Objectives Article" which clearly embodying the following points:

- Problem
- Hypothesis.
- Objectives.
- Scope of the problem
- Development methodology of the thesis

It is therefore a series of articles that trigger at the origin of the research and the idea of the author is protected from the start by posting the same and its transmission to the members of the scientific community in the concrete explanation that general solutions can be a bear to achieve the objectives.

B. Second stage

In the previous phase have described that we do, through a process of validating ideas and methodology publications, in this second phase begins developing the doctoral thesis, verifying and enhancing their knowledge, this knowledge is not science specific on which the PhD is done, but the novelty here ye must learn and see what knowledge and must be purchased on the implementation of the thesis, dissertation that apply to a specific production sector.

In this case depending on the possibilities the director of the thesis, shall perform a work accompanying the doctoral candidate in the search for work experience in a company of the productive sector, as in the formation of this in the field, thus we get that you obtain a better view doctoral researcher field applied to the reality of the productive sector, in turn the doctoral student must complete the items necessary for the realization that solutions can be, how it will be if your metodolgoia7y seen an increase and modification of the objectives, hypotheses and the emergence of new problems.

This phase must be completed with the creation of an article in which the objectives will be translated and how they will achieve these goals, technologies and designs all the justification necessary in order to introduce the members of the scientific community and its progress if need help from a researcher in validating their results or have an investigator which states something. This is the turning point where the work really begins, once the revision of the article from which the solution begins with the objectives.

C. Third stage

Once validated and completed the objectives, problem and methodology doctoral students proceed from scientific publications a specific development for each of the objectives, i.e. scientific publications will be aimed at the solution of each of its objectives separately, where maintaining a classical structure of the article, abstract, introduction, aim of the thesis being treated, points of development and verification of the aim of the thesis.

During the review process these items reviewers may seek clarification on these items or possible improvements for sharing the results thereof, such information shall be forwarded to the director of the thesis in order to recapitulate the development data items if necessary a greater realization of the objectives or a breakdown thereof, the doctoral student must retake these changes and make publications in order to clarify and modify the changes that reviewers indicate.

D. Fourth stage

At this stage we continue the exploration of the research, it is time to validate the objectives and whether the solution is true, for it would have to perform the doctoral an article of inclusiveness, we might call integrator article dissertation, this article must integrate items developed in previous phases, this is where the doctoral student researchers and will see the entire investigation.

This point is important because the overall view can offer both weaknesses and strengths in the investigation thereof, is another as another feedback point of the thesis. This can be done by hand, a redefinition of the objectives or the appearance of new targets to be treated as a whole, at this stage the direct application of the results obtained in a case of direct application is also included.

There are multiple options for an application of the results obtained, we believe that such validation items that have been made, it would be interesting for the creation of articles, presentation application sectors of the thesis, we thus validate Objectives from the IT point of view but from the point of view of the application of research. This is an analysis of the results expose these results to anyone who will use and proceed to the validation of these from the point of view of the application.

E. Fifth stage

The evaluation process, feared by all doctoral candidates, one way to do this is to make a final scientific article in which the doctoral drafted the level of achievement of the objectives and the solution of the problem has been achieved, this can be done many forms depending on the subject of the thesis, a way is made an article with realizer integrative conclusions based on the main items of a solution to the whole problem, in this paper besides an integration under one thread should capture the level of achievement of the objectives through the articles, a section that can clarify this situation is the main product, a case study in which the solution to be integrated, this has been developed in different scientific papers but is a way to facilitate an understanding of their research as a whole.

The combination of the phases can be performed the same research methodology, we highlight one hand the organization of the work we have previously pointed out the phases and other research methodology, this methodology must also develop in the realization of the articles, the reason for this is that the development of the items directly affects a good realization of the final conclusions and also have to have external factors as shown previously described. In the figure below we present such as the research methodology within each of the phases.

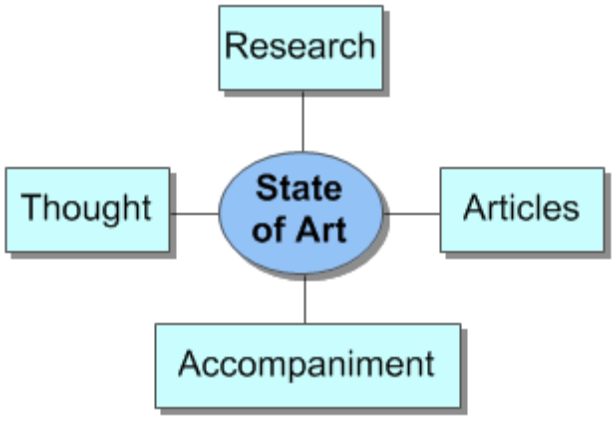


Fig. 2 Research methodology

From the state of the art, a research by the scientific articles that tries to find solutions to the objectives and problems encountered in the translation of the thesis is done, the presentation of scientific papers in international journals and conference allows the option to review other researchers, we put in the focus state of the art, this state of the art is not that we are used to now, is to make a state of the art in two parts, first we make a study of what the needs or as the productive sector which is destined. It is also important accompaniment of the thesis supervisor and the doctoral student in situations where you will apply the final result of the thesis in order to produce feedback to the thesis.

Once the sections of the thesis done we can think of the following index for the thesis, as these consisted of a doctoral thesis by publications is necessary to capture very specific things, and a good way to see it is through an index, which it can be shown below:

1) Problem Hypothesis and Objectives.
2) Scope of the problem.
3) Theoretical Framework: State of the art where
4) Need.
5) Articles published, plus should consist of the full text article for a description of the stage at which it is and has been achieved with this article.
6) Conclusions inclusive, this point should in addition to the final conclusions of the thesis describe, ideally write a case study integrating papers made in order to see an overall solution to the problem described in another section and the level of achieving the objectives through the items.
7) Future work.

V. Centralized research resumé

Once we have described the methodology and the different phases, in previous research we have described as the centralization of student resume is performed [10] [11]. In this section, we plated the goals of our future research in order to extract information from the presentation of this publication will be a starting point for a series of articles in which we set the following objectives.

Centralize the doctoral curriculum, depending on the work done during the development of the doctoral thesis, in which it is to storing all the work done, this main objective we can be subdivided into:

- Classification of the knowledge of the candidate technologies starting with handles and transverse fields has been working.
- Targets developed on each of the items, giving a score by the reviewers as to the level of achievement of the objectives in the articles.
- Search for researchers under a similar profile, in order to enable collaborative research work with similar objectives, although these are in a different line of research
- Communicate with companies achieved the objectives in research; this would open the way to the outside research in the investigation periods.
- Curriculum vitae researcher from the milestones it has been marking the researcher, making unique and validated cataloging items.
- Integration of the resumes of researchers in a body that also lend credibility to this research allow feedback, this organism can be for example the top universities, can use this feedback to generate other lines of research.
- Recognize the idea of the researcher in scientific forums, in order to ensure that the PhD was who came up with the idea and avoid potential problems in the development of research that can last five years.

These objectives are the beginning of a new line of research and here with comments from reviewers and members of this conference can be a start to know the real possibility of whether a line of research is possible or necessary.

VI. CASE OF STUDY

In this section we will make the assessment objectives which would be a summary of the findings and integrating validation of submitted articles for my doctoral thesis, the starting point was 4 years ago when the computer was working doctoral technician for School 2.0 program [12] in Spain, the field of application of the thesis is the use of ICT in classroom education, so we set from the beginning to facilitate the tasks of the teacher through a system that allows the creation, maintenance curricular programming for the educational stages of Primary, Secondary and Vocational Education, from:

1. Creating curricular schedules that allow teachers to structure their programming following the official curriculum, divided into teaching units have a distribution in sessions over time, justification, objectives, core competencies, provide content, materials needed, activities, qualification students, attention to diversity for students with special needs, methodology.
2. Structuring and preparing the working sessions within their curricular programming, taking into account the learning unit wherein said working session is located.
3. Track the teaching-learning process of their students in collaboration with parents, attending the assessment, rating and monitoring of the student in this process, which will also improve mentoring.
4. Ensure a flexible system that allows the inclusion of activities and creating them from existing educational platforms, including such activities in the lesson plans the teacher develops within its programming.
5. Centralize the educational curriculum vitae student, taking into account the results and the level of achievement of the objectives in addition to basic skills attained.
6. Improve the interaction between students in the classroom by Distributed Graph User Interfaces (DUI), thus improving the teaching methodology in the classroom.

These objectives we have worked throughout the research time and have been monitoring the level of achievement of the objectives through the items, as you can see in the summary of the conclusions, which we integrate below:

First we defined the cloud services necessary for the development of curricular schedules of teachers, including the creation of curriculum schedules following the official curriculum and following a structured work sessions where the teacher indicates the activities to resolve considering objectives, core competencies and the specific needs of each student [25].

Once you have developed these services is necessary to establish the curricular structure of the official curriculum and curricular programming that allows storage managers of database systems with the information associated with each of these [14] [15].

One of the most important aspects is to track student evaluation. In this respect we track the teaching-learning process based on objective and core competencies of the contents worked by students in activities and assessment tests have been conducted, establish a system to score each of the activities taking into account basic objectives and the teacher determined the creation of each one of the activities that make up the didactic units that form the annual teacher curricular programming skills. The inclusion of activities in curricular programming allows teachers to better development of their daily work and provides flexibility that is possible thanks to the integration capacity of the systems in the cloud and the ability to leverage interoperability of Web services [16] [20] [21] [23] [24].

On the other hand improving student resume and centralization is possible because the system on which the assessment is made is a score of basic skills that students obtained in academic courses [10] [11].

The teaching methodology is achieved by improving the interaction, by using the graphic user interfaces distributed (DUI, Distributed User Interfaces) and insertion patterns of education as a service in the cloud, allowing the inclusion of teaching methodologies in programming and that the teacher can adapt these methodologies to the student group [19] [22] [24].

In the research process is important also apply learning, which is why publications have made a different line of research to validate what we are doing and allow us to continue our research in the future to other lines of research [17] [18].

VII. CONCLUSIONS AND FUTURE WORK

The development of doctoral theses by workers compendium also includes collaborative work we must also give a definition and approach to achieve specific objectives, these objectives must abide by and comply to reach a good end development, these goals are dismembered objectives pertaining for each item in order to perform a better specification of problems to solve. Surely many of these objectives will be converged with other researchers so flexibility is important and here comes the opening teamwork among researchers, so it will be important to re plan, refocus and rediscuss the scope and needs from know the real development of the doctoral thesis.

The Director of the thesis is a fundamental task as it must ensure the best options for the doctorate and the completion of the thesis in a while possible, this teamwork between different doctoral can make different lines appear very close research so the director of the thesis may take this opportunity to look for new and continuing doctoral research.

The type of evaluation process is enclosed in a unstructured learning environments, with specific objectives but flexible and open, conducted by a research activity that encourages told and team learning environments between different doctoral research and also took the research experience to the production system in real contexts, experiential learning theory of John Dewey suggests in his theory of knowledge.

REFERENCES

[1] Dewey, J., Llavador, F. B., & Llavador, F. B. (1997). Mi credo pedagógico. Universidad de León.
[2] Kilpatrick, W. H. Los proyectos y su secuencia de trabajo. Revista digital enfoques educativos, 42.

[3] Dewey, J. (2004). Experiencia y educación.
[4] Carreño, M., Colmenar Carmen; Egido Inmaculada y Sanz Florentino. (2000). Teorías e instituciones contemporáneas de educación. Madrid: Síntesis educación
[5] Joyce, B. R., Weil, M., & Calhoun, E. (1985). Modelos de enseñanza. Anaya/2.
[6] Dewey, J. (1995). Experiencia y pensamiento. Madrid: Morata, 124-34.
[7] Claves de la gestión de proyectos. Gestión eficiente de proyectos y de trabajo en equipo grolimund carlos, ed. Fund. Confemetal, 2011
[8] Ausubel, D. P., Novak, J. D., & Hanesian, H. (1976). Psicología educativa: un punto de vista cognoscitivo (Vol. 3). México: Trillas.
[9] Meyer, W & Van Dalen, D. (1978) Manual de técnica de la investigación educacional. Editorial Paidos.
[10] H. M. Fardoun, A. Paules y D. M. Alghazzawi, "Centralizing Students Curriculums to the Professional Work," Elsevier Procedia - Social and Behavioral Sciences, vol 122, 2012, pp. 373-380.
[11] H, M. Fardoun, A. Paules y A. S. Mashat, "Improvement of Students Professional Formation Currículums to meet the market Work," Elsevier Procedia - Social and Behavioral Sciences, vol. 122, 2014, pp. 416-420.
[12] Ministerio de Ciencia e Innovación. "Proyecto de investigación. Las políticas de un «ordenador por niño» en España. Visiones y prácticas del profesorado ante el programa escuela 2.0. Un análisis comparado entre comunidades autónomas".
[13] H. M. Fardoun, B. Zafar, A. H. Altalhi y A. Paules, "Interactive Design System for Schools using Cloud," Journal of Universal Computer Science, vol. 19, no. 17, 2014, pp. 950-964.
[14] H. M.Fardoun, A. H. Altalhi y A. Paules, "Improvement of students curricula in educational environments by means of online communities and social networks," Springer Online Communities and Social Computing, vol 8029, 2013, pp. 147-155.
[15] H. M. Fardoun, A. Paules y K. M. Jambi, "Educational Currículum Management on rural environment," Elsevier Procedia - Social and Behavioral Sciences, vol 122, 2014, pp. 421-427.
[16] H. M. Fardoun, A. AL-Malaise, A. Paules y B. Zafar, "Improvement of Students Access to Work by means of TICs," Elsevier Procedia - Social and Behavioral Sciences, vol 122, 2014, pp. 367-372.
[17] H. M. Fardoun, A. Paules, D. Alghazzawi y M. Oadah, "KAU e-Health Mobile System," Proc of XIII Congreso Internacional de Interacción Persona-Ordenador, 2012, pp. 171-175.
[18] H. M. Fardoun, A. Mashat y A. Paules, "Mecca Access and Security Control System," Proc of XIII Congreso Internacional de Interacción Persona-Ordenador, 2012, pp. 199-205.
[19] H. M. Fardoun, D. Alghazzawi y A. Paules, "TabletNet: Utility, Usability and User Interface Quality," Proc of XIII Congreso Internacional de Interacción Persona-Ordenador, 2012, pp. 365-366.
[20] A. Paules, H. M. Fardoun y A. Mashat, "Cataloging teaching units: Resources, evaluation and collaboration," Proc Federated Conference on Computer Science and Information Systems, 2012, pp. 825-830.
[21] H. M. Fardoun, B. Zafar y A. Paules, "Using Facebook for collaborative academic activities in education," Springer Online Communities and Social Computing, vol 8029, 2013, pp. 137-146.
[22] H. M. Fardoun, A. Paules Ciprés y D. M. Alghazzawi, "Distributed User Interfaces to Enrich Collaborative Teaching Methods," Proc. of the 3nd Workshop on Distributed User Interfaces: Collaboration and Usability, 2013, pp. 37-41. DBLP. SPRINGER
[23] H. M. Fardoun, A. Paules Ciprés, A. AL-Ghamdi, "Tutor Platform for Vocational Students," Proc Federated Conference on Computer Science and Information Systems, 2013, pp. 703-707.
[24] H. M. Fardoun, A. Paules Cipres, D. M. Alghazzawi, "Distributed User Interfaces in a Cloud Educational System," Distributed User Interfaces: Usability and Collaboration, M.D. Lozano, J.A. Gallud, R. Tesoriero y V. Penichet eds.,Srpinger, 2013, pp. 151-163.
[25] H.M. Fardoun, A. Paules Ciprés, D. M. Alghazzawi, "CSchool - DUI for Educational System using Clouds," Proc of the 2nd Workshop on Distributed User Interfaces: Collaboration and Usability, 2012, pp. 35-39.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 809–814
DOI: 10.15439/2014F397
ACSIS, Vol. 2

New Teaching Techniques of Mathematics Subjects by means of Artificial Genesis

Habib M. Fardoun,
Daniyal M. Alghazzawi
Information Systems Department
King Abdulaziz University (KAU)
Jeddah, Saudi Arabia
Email: {hfardoun, dghazzawi}@kau.edu.sa

Antonio Paules Ciprés
European University of Madrid,
Madrid, Spain
Email: apcipres@gmail.com

***Abstract*—This paper presents an adaptation of the Brousseau method. The concepts of didactic variations and interactions are introduced by using didactic and a-didactic situations. With the main aim of guaranteeing alumni improvement and evaluation processes, a case study application was development founded on a cloud architecture, which is also based in previous research, curricular organization and interaction processes within the class.**

I. Introduction

GUY Brousseau introduced theory of situations, didactic contracts and didactic and a-didactic situations in 1998. This is a teaching theory seeking the conditions for artificial genesis of mathematical knowledge. Brousseau says that "knowing mathematics" is not only about knowing definitions and theorems to recognize the opportunity to use and apply them, but also about "dealing with problems" in a broad sense, which means finding good questions and solutions.

Computer science professors and researchers play an important role at first stages of the computer science teaching process and should be able to start asking questions about the process of teaching and learning that is used with their students in their daily work.

Curriculum development in university subjects lacks structure focused on the student evaluation process. Evaluation and assessment is an inexistent process, since in most cases it consists of an evaluation of the practical part of the subject, accompanied by a short presentation by the student, and a final exam to asses theoretical contents.

This teaching and learning process does not apply any clear pedagogical basis, with the exception of objectives achievement and is often subject to the teacher's decision. For example, in a Master thesis, the final evaluation is done in a presentation where the student talks about her work during a fixed time.

Sometimes, didactic methods are not used due to either ignorance or lack of transversality among subjects. However, there exists a wide use of teamwork [1] or project work [2], mainly because these methodologies adapt perfectly to work that students will perform in their professional future.

Didactic adaptation of subjects at the university level is likely to grow over time. One of the goals of educational outreach is the university of education, which will lead to an increment of the population percentage with higher education. Universities will need to include teaching methods that allow curricular adaptations, significant [3] and not significant [4].

Furthermore, the enhancement of students creative thinking from the Brousseau perspective can lead to the creation of new products and creative solutions to solve complex problems. In this way, students of Computer Science can improve their abilities to solve problems with training and adaptation methods. On the other hand, university teachers could use these methods autonomously and adapt their needs over time.

II. State of the art

During our research work, we have carried out the adaptation of the Montessori method [5] and the leverage of social networks in educative environments [6]. This paper presents an adaptation of the Brousseau method.

Brousseau developed the "Situation Theory", in which he searches conditions for an artificial genesis of mathematic knowledge. Under the assumption that they are not built spontaneously, it is based on the constructive conception of thought.

Pedagogic constructivism is to equip students with the necessary tools, allowing them to create their own procedures to solve problems.

There are four essential features of a constructivist action [7]:

1) It is based on the conceptual structure of each student: part of the ideas and preconceptions that students bring to the classic topic.
2) Anticipates conceptual change is expected from the active construction of the new concept and its impact on mental structure.
3) Confront the ideas and preconceptions related topic of education, with the new scientific concept being taught.
4) Applies the new concept to concrete situations and relates to other concepts of cognitive structure in order to expand its transfer.

Necessary conditions to boost constructivist teaching are the following:

978-83-60810-58-3/$25.00

1) Generate dissatisfaction with prejudices and preconceptions, allowing students to identify their mistakes.
2) The new concept begins to be clear and distinct to the previous.
3) The new concept shows its applicability to real situations.
4) The new concept generates new questions and expectations.
5) The students observe and understand the causes of their prejudices and misconceptions.
6) Create a climate for free expression of student, without coercion or fear of making mistakes.
7) Foster the conditions for the student as a participant in the teaching-learning process: planning, selection of activities, sources of information queries, etc.

In the Brousseau theory of constructivism, students learn to adapt to an environment that is a factor of contradictions, difficulties and new knowledge is acquired through responses that are evidence of learning. This is achieved with didactic and a-didactic situations.

A didactic situation is a set of relations explicitly and / or implicitly established between a student or group of students in an environment, including, eventually, instruments and objects, and an educational system (teacher). Student work should, at least in part, reproduce the characteristics of scientific work itself, to guarantee an effective construction of relevant knowledge. Each didactic situation is governed by a certain type of training contract, i.e. a set of implicit and explicit obligations, namely an action between teacher and students.

An a-didactic situation is essentially characterized by the fact of representing certain learning moments in which the student works independently, not subject to any direct control of the teacher. In this situation, the student becomes able to reuse the knowledge that he is acquiring, in a situation not covered in any teaching context and in the absence of any teacher.

Summing up, a didactic situation is one that inherently contains the intention of someone to learn something. This intention does not disappear in a-didactic situation, where the lack of intention means that the student must address the problem by replying to it based on his knowledge, motivated by the problem, not by the desire to please the teacher [10].

Brousseau establishes a didactic contract. The didactic contract was defined by Rousseau [11] and there are several functions performed:

1) Promotes student autonomy and responsibility.
2) Attends to the particularities of the student.
3) Facilitates the interest and motivation of students in their own learning process.
4) Academically guides student work.
5) Democratizes the education to take into account instances of negotiation in setting learning objectives, course content and assessment process.
6) Stimulates the ability of self-reflection on the student's own learning enhancing critical thinking.

Brousseau also adds that the training contract refers to the actions established between teacher and student, thus comprising the set of behaviors that the teacher expects from the student and the set of that the student expects from the teacher.

The didactic contract, along with educational and a-didactic situations, can be modified by the didactic situations are theoretical objects whose purpose is to study the set of conditions and own a well-defined relations knowledge. The contract can use values that enable students to understand and resolve the situation with their prior knowledge, and then make them face the acquisition of new knowledge by setting a new value of a variable.

The contract has a structure similar to a teaching unit that we developed in prior research. In this previous research we also included a study on how to evaluate and assess students in conventional curricular situations [12] and in social networks [13]. Also, we studied the necessity of curricular organization in rural environments [14]. The outcome from this research showed that evaluation and assessment should be done by taking into account the whole teaching and learning process. Student evolution should be registered and logged during the process.

III. Definition and objectives

This article shows an adaptation of the Brousseau method, developed in order to improve the learning of Computer Science and change the approach of subject from a raw contents perspective. The starting point is to be able to address problems from the computer science point of view. These capabilities involve the use of techniques and technologies designed for problem resolution. Furthermore, this approach can improve students' creativity in the development of new products and particular solutions to problems.

IV. Didactic methodology

Teaching methods drawn from this technique are a part of a master class [15], where the teacher presents concepts and takes the approach to solving problems in addition to the qualification criteria (training contract). At this point, students should start having doubts regarding these content. These doubts or questions are contract breakers. To avoid, the contract breaking, the teacher interferes with the development of problem proposing new situations for the student to solve. At this point project work, collaborative work, or individual work can be used.

The introduction of educational variables in this methodology complements a problem during the teaching-learning process, creating milestones in the teaching methodology and turning points that allow students to obtain new solutions from these variables.

These variables can be adapted by the teacher for students with learning and teaching needs with significant or not significant curricular adaptations, thus adapting the activity of students to the specific needs of these or group level students each time.

These educational variables are milestones, which must be solved by the students. These new standards raise new

questions for students to be overcome and entail the creation of new solutions to the problem proposed completing its complexity or making new methods to its solution.

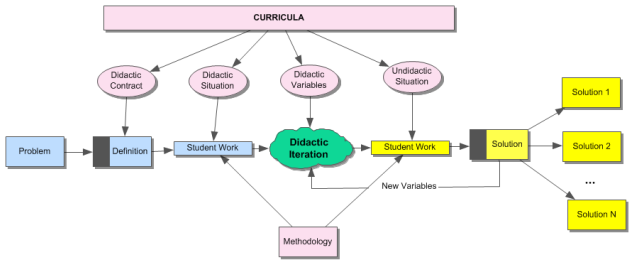


Fig. 1 Didactic Methodology

The above diagram depicts the operation of this methodology. The methodology starts with a problems and ends with a solution. This solution is not unique because during that modify variables included.

• Blue boxes represent the beginning of the process. The teacher defines the problem and the students begin working on its definition. During this process, students adapt the problem to their academic interests, mainly based on the knowledge acquired in previous experience of other activities.

• Didactic Iteration is in green. The teacher introduces variables that cause the a-didactic situation in students and create confusion enabling them to continue the development of the solution. This a-didactic situation can cause various solutions the problem.

• Yellow boxes highlight the iterating process where the teacher introduces new and triggers a process where different teaching methodologies are applied.

• Purple boxes represent the curricular part of the technique described above. This requires access to information in the official curriculums, which are created on the teacher's teaching schedule.

Summing up, the teacher develops a problem statement, which is completed by the students in the teaching-learning process. Then, by creating the conflict, students learn and develop complex problems, from their own knowledge, while teacher supervises the development of the activity and the achievement of objectives.

V. Models and States

The design of models for a solution to this teaching methodology is defined by the following:

• Problem: The problem statement must be constructed from the knowledge gathered from the student group considering transversality and curriculum contents. The process must take into account the objectives and core competencies the activity will be based on.

• Definition: In this model, the teacher helps students continue with the development and realization of the problem statement, using knowledge acquired in other subjects (transversality of education).

• Student Work: The "individual student" work, influenced by the different teaching methods that teachers adopt. These methodologies entail personal, individual work and collaboration, and should be applied after periods of discussion that generate ideas, thus achieving greater articulation of such proposals or solutions as appropriate.

• Didactic Variable: Within the process of learning the teacher can add new challenges to the student. These challenges are raised by questions, which are asked in order to break the didactical contract and sow doubt in the chosen solution, allowing the student progress in this process.

• Teaching Iteration: Starting from the doubt created by the a-didactic, the teacher introduces didactic variables so that students can continue their work. These didactic iterations can last until the teacher considers the objectives are met.

These models must be accompanied by states, which are changes produced by the introduction of a model. In this way, the teaching-learning process is configured and can modify the behavior of the model itself. It is a way to introduce a change in the model by autonomic self-employment in a situation that produces a change in both solution and in the definition.

• Didactic Situation: Not a model itself, but a state in the process where the teacher creates the curricula through the objectives of the practice or problem statement. In this state, the student collects information about what to do and how to do it. The end of the situation is the acquisition of a training contract containing the work to be performed, which may be modified posterity.

• A-didactic situation: This situation is a state of reflection on the proposed problem in which students develop their cognitive elements of an evolution problem. More restrictions to the problem are raised, thanks to the introduction of didactic variables added before.

For the development of this methodology, we consider the situation of the student group and the level of achievement of the objectives. Therefore there is a direct relationship with the official curriculum [12] and in teaching methodology used the classroom [15] [16] hence the need for data collection of the teaching-learning process.

VI. Method Application

The structure we have for the development of the subject of IT projects is a matter that also needs the support of other subjects in the field of software engineering, database management systems and data programming.

From the learning outcomes and assessment criteria the teacher proposes the problem:

Problem: short form corresponds to the problem statement. "...Create a management application that allows professionals to manage their company store..."

Definition: Students begin to define the problem and begin to questions arise that will lead to the creation of the didactic contract. This methodology can be accomplished by collaborative work and well predefined roles.

Didactic contract made by the students and the teacher, which is the result of acquired learning and assessment crite-

ria (goals to achieve). The contract provides students with the ability to get that learning outcome, according to the curriculum of the course:

1. **Identify needs of the productive sector, relating them to projects that can satisfy these needs.** The following results are obtained:

a) Related companies were classified according to their organizational characteristics and the type of product or service they offer.

b) Firms were characterized indicating the organizational structure and functions of each department.

c) Most demanding business needs were identified.

d) Predictable business opportunities in the sector were assessed.

e) Type of project was identified in order to meet the anticipated demands.

f) Specific characteristics of the project were specified as required.

g) Tax, labor and risk prevention and implementation conditions of obligations were determined.

h) Potential aid or subsidies for the incorporation of new production technologies were identified.

i) A roadmap was developed for the project.

2. **Design projects related to the competences described in the statement, explicitly developing the phases that compose it.** The following results are obtained:

a) Information was compiled on the aspects that will be addressed in the project.

b) Technical feasibility study of the project was done.

c) Project stages were defined, specifying its content and deadlines.

d) Targets by scope were set.

e) Development supporting activities were determined.

f) Necessary resources and personnel to complete the project were predicted.

g) Funding needs were identified.

h) Project documentation was defined and created.

i) Aspects that must be controlled to ensure project quality were identified.

3. **Schedule the execution of the project, determining the intervention plan and associated documentation**, The following results are obtained:

a) Tasks were scheduled according to implementation needs.

b) Logistics and resources required for each task were determined.

c) Roles and permissions to carry out each tasks were identified.

d) Procedures have been identified for implementation of tasks.

e) Risks were defined, defining the risk prevention plan and means.

f) Allocation of material and human resources were planned.

g) Economic assessment that responds to the conditions of the project implementation was carried out.

h) Documentation necessary for project implementation was defined and developed.

Once the teacher has presented learning results, objectives and assessment criteria, students define the problem they want to solve, through collaborative work or moderated discussion. The definition of the problem is done following these points

• Conduct a marketing job.
• Conduct a study of economic feasibility.
• Create a project plan.
• Create an implementation plan for the company.
• Specify a requirements document.
• Study legal legislation in the creation of a software project
• Balance project risks
• Perform an economic assessment of the project.
• Characterize the productive sector. Students decided that it would be garage to and a car parts store.
• Establish the garage modules: warehouse management, repair management, billing and online sales.
• Establish phases of application design and make the application requirements in order to ensure the success of it.
• Create a testing phase.
• Tighten the planning of project duration and resources required for its development.
• Perform a marketing study of the developed product.
• Teamwork: Create teams in the initial phase to enable the design, analysis and implementation of the project, also defining the role of each team:

- Design team: system design, databases design, UML and database management systems.

- Analysis Team: conduct the analysis of the final creation of a requirements document, also will perform market research and project feasibility.

- Implementation Team: initially undertake technological evidence necessary for the implementation of the subject and phase necessary to select databases and programming languages needed to implement tests.

Once the didactic contract with students is done, we can see the challenge that students have been established, as there is a considerable lack of order and some aspects remain undefined. At this point, the teacher plays an important role as a project leader, providing the students a pseudo-real situation as close as possible to a corporate project, without forgetting the academic taking into account that the a-didactic situation should let students progress on their own.

A. Introduction to syntactic variables and interaction production

Students showed their findings in a presentation to the group. During the questions and answers time, the teacher asked questions to the group of students from each team. These questions should make students question their own activities and complete their work.

For example, the following are questions made to the different teams. This was asked to the analysis team was: “Is all your work aimed to find the customer's needs, divided into different types of requirements, as you specified in the anal-

ysis phase?" The following comment was made to the design team: "You have decided which technology to use, I see you chose a language taught in other subjects (Java), but I do not see how the analysis team could use that work, you must design thinking to support the analysis team. Should the customer to see something during the requirements phase?" The implementation team was asked the following: "You carried out stress testing, you designed a database cluster and client-server communication, but what if I have to offer an external service such as sending data to a central provider or upgrade fees from the central shopping?"

At the end of the class, a student from the analysis team points out the following: "How should we organize the staff? This is a mess; it took us a week and we have nothing, we can not offer a joint solution for business… in software engineering classes the teacher told us about CASE tools, should we use them?"

A member of the design team questioned this decision and indicated they already had the Rational Rose tool at their disposal, which could be integrated with Requisite Pro tool, a specific tool for requirements definition, also saying that they could create database table in the design stage.

A member of the implementation team showed the possibility of defining a logical structured to develop a modular application. Another member said that they could create the accounting module with an ERP and save time.

The teacher finally decided to step in this chaos, with the sole purpose of moderating, and also indicating that he wanted to see them at his office the next morning. Next morning, only one member of each team showed up. Groups had just defined the team leader role without the teacher intervention.

During this mentoring sessions, issues discussed the day before are solved. Students established a modular structure and decided to continue using Rational Rose and Requisite Pro, as their combination is a way of linking design and analysis phases, finishing up with the definition of use cases. Regarding requirements definition, students decide to start developing IGU prototypes in order to achieve a better definition of requirements and ease the creation of a database. The also decided to do a reverse engineering process on industry applications, to includes industry data in the analysis, which is a good way to make an analysis and design when software teams do not have own data.

In this interaction, with the introduction of variables, the progress was considerable in terms of project organization. Students made decisions based on prior knowledge and sought appropriate solutions to problems, creating a project development methodology.

VII. Architecture

We need to create architecture for incorporating the method of Brousseau. This architecture is focused on the cloud to ensure the possibility of growth and integration of services built on previous research. It is essential to consider the organization the teacher does in the academic environment, the documentation of teaching the set of objectives that the student has reached. Also, the teacher has to conduct the evaluation and monitoring of the goals of the students. This requires the creation of a workspace.

Figure 2 depicts the cloud architecture that enables this work, as we will now need a part that facilitates interaction and creation of workspaces based on the teaching methods and other assessment and preparation activities for the teacher.

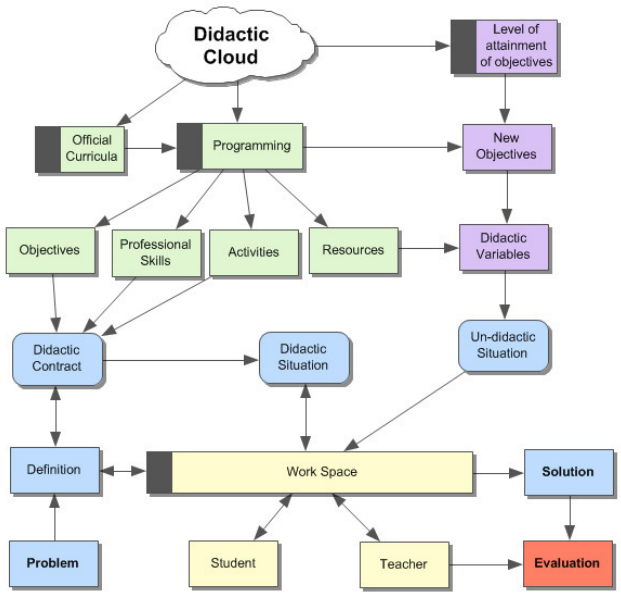


Fig. 2 Architecture in the cloud

We set a conceptual level architecture in the cloud, as we can see there are several parts.

Green modules represent the organization that the teacher created for the subject. Data is gathered from the subject curricula, divided into objectives, skills, activities and resources. The curricular organization that the teacher has developed for a subject must be adapted to the Brousseau methodology, thus conducting a curricular inclusion in the method we are developing.

Parts of the method are depicted in blue. They are some of the patterns and conditions described in the previous section. Data is collected both from the didactic plan and transversality necessary for the advancement of the data. The method begins by defining the problem and creating a learning contract that allows the teacher to start the process. Once the student work and under a teaching situation, the a-didactic situation comes into play, which should finish with the solution of the problem.

In purple, we highlight didactic variables selection that create the a-didactic situation and break the didactic contract. These educational variables are defined by the teacher based on the objectives of the students obtained in other subjects and also new goals. In both cases, the student level of objectives achievement is taken into account, thus guaranteeing transversality and the fact that students will be able to come up with an overall solution.

For the evaluation, the teacher needs objectives developed by each student and the student group as a whole in order to ensure a quantitative assessment of the objectives developed in the problem statement. At this point the teacher needs to

change the perspective to see entire process. This means that activities performed by the students must be complete and must be exported to the teacher assessment notebook.

The workspace or interaction space for students and teachers is where they perform the work. The workspace must conform to the methodology that the teacher has selected and also have access to resources, this can be achieved with the CSchool Interactive Design tool [17] and use a cloud application environment [18].

VIII. Conclusion

The method of Brousseau, raises the goal that students should learn to solve problems. We use this method for computer courses. A case study was presented in which students develop the problem by introducing variables. Teacher's work in such situations is to guide students through the educational variables and produce iterations in each stage of the process. On the other hand, it is important to set limits to ensure the achievement of the objectives of the course and also control the development of the sessions. Course planning comes from the subject curriculum and, in this case in particular, needs other subjects to achieve the same goals. Currently the problem is the inhibition for students to enroll in certain courses if they have not passed previous more basic courses. This is not a problem with our approach, as there is a possibility that students review the lessons learnt in other subjects. This suggests that future work on this methodology should be directed towards improving students' curriculum and promotes the objectives of subjects from others that require such knowledge. For example, if a student does not have the minimum knowledge to pass the programming databases course, he could be placed in a team with another student who masters this subject, supporting the other student. Activities like this one would need to know what to do with interdisciplinary work between subjects, how to evaluate it and how allow academics the student profile improvement.

References

[1] Hernández, F., & Oller, M. V. (1992). La organización del currículum por proyectos de trabajo. Graó.
[2] Ander-Egg, E. (2001). El trabajo en equipo. Editorial Progreso.
[3] Ballester Vallori, A. (2005). El aprendizaje significativo en la práctica. In V Congreso Internacional Virtual de Educación.
[4] Stainback, S., & Stainback, W. (Eds.). (1999). Aulas inclusivas: un nuevo modo de enfocar y vivir el currículo (Vol. 79). Narcea Ediciones.
[5] Habib M. Fardoun, Abdullah Saad AL-Malaise Al-Ghamdi, Antonio Paules Cipres, Creating new Teaching Techniques with ITCs Following the Montessori Method for Uneducable Young Students, 15th International Conference on Enterprise Information Systems (ICEIS-2013), ESEO, Angers Loire Valley, France-2013.
[6] H. M. Fardoun, B. Zafar y A. Paules, "Using Facebook for collaborative academic activities in education," Springer Online Communities and Social Computing, vol 8029, 2013, pp. 137-146.
[7] i Gallart, I. S., & Font, C. M. (1996). Asesoramiento psicopedagógico: una perspectiva profesional y constructivista. Alianza Editorial.
[8] De Zubiría Samper, J. (2006). Los modelos pedagógicos. Hacía una pedagogía dialogante. COOP. EDITORIAL MAGISTERIO.
[9] D'Amore, B., & Brousseau, G. (2005). Bases filosóficas, pedagógicas, epistemológicas y conceptuales de la Didáctica de la Matemática. Reverté.
[10] Johsua S. Et Dupin J. J. (1993) Introduction à la didactique des sciences et des mathématiques, Paris, PUF
[11] D'Amore, B. (2002). Influencias del contrato didáctico y de sus cláusolas en las actividades matemáticas en la escuela primaria.
[12] A. Paules, H. M. Fardoun y A. Mashat, "Cataloging teaching units: Resources, evaluation and collaboration," Proc Federated Conference on Computer Science and Information Systems, 2012, pp. 825-830
[13] H. M. Fardoun, A. H. Altalhi y A. Paules, "Improvement of students curricula in educational environments by means of online communities and social networks," Springer Online Communities and Social Computing, vol 8029, 2013, pp. 147-155.
[14] H. M. Fardoun, A. Paules y K. M. Jambi, "Educational Currículum Management on rural environment," Elsevier Procedia - Social and Behavioral Sciences, vol 122, 2014, pp. 421-427.
[15] H. M. Fardoun, A. Paules Ciprés y D. M. Alghazzawi, "Distributed User Interfaces to Enrich Collaborative Teaching Methods," Proc. of the 3nd Workshop on Distributed User Interfaces: Collaboration and Usability, 2013, pp. 37-41. DBLP. SPRINGER
[16] H. M. Fardoun, A. Paules Cipres, D. M. Alghazzawi, "Distributed User Interfaces in a Cloud Educational System," Distributed User Interfaces: Usability and Collaboration, M.D. Lozano, J.A. Gallud, R. Tesoriero y V. Penichet eds.,Srpinger, 2013, pp. 151-163.
[17] Habib M. Fardoun, Bassam Zafar, Abdulrahman H. Altalhi, Antonio Paules Ciprés: Interactive Design System for Schools using Cloud Computing. J. UCS 19(7): 950-964 (2013)
[18] H. M. Fardoun, B. Zafar, A. H. Altalhi y A. Paules, "Interactive Design System for Schools using Cloud," Journal of Universal Computer Science, vol. 19, no. 17, 2014, pp. 950-964. JCR, IF=0.762. SCOPUS. Thomson Reuters. DBLP.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 815–817
DOI: 10.15439/2014F383
ACSIS, Vol. 2

Global Unification Model of Studies based on similar subjects

Habib M. Fardoun
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Kingdom of Saudi Arabia
Email: hfardoun@kau.edu.sa

Daniyal M. Alghazzawi
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Kingdom of Saudi Arabia
Email: dghazzawi@kau.edu.sa

Lorenzo Carretero González
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Kingdom of Saudi Arabia
Email: lgonzalez@kau.edu.sa

***Abstract*—We propose a Global Unification Model of Studies where the subjects, which follow a same educational plan inside of a specific state or country, can be selected by students from any place inside of that state/country. This proposal increases the university's flexibility providing to the students the option of selecting subjects given for teachers of other universities. In addition we promote the unification of studies' plans to facilitate the subjects' selection and to step forward for a future global unification.**

I. Introduction

WE have to take into consideration the current trend to a high education studies globalization to allow students of different places to work in a foreign country without problems derived from their curricula or previous studies. Therefore, a particular student that had done his/her studies in Spain could go to any country of the European Union without to normalize the degree or the titles obtained previously. It facilitates a fluent movement of prepared people inside of the continent. At the same way that frontiers and different kinds of currency made difficult the transactions and emigration, different kinds of studies make that students of separated areas don't have the flexibility and freedom needed to continue promoting a globalization.

To contribute to continue with a studies' globalization, we propose an approach that consists of unifying most of the subjects of a specific degree, which will allow students to choose among different universities the subjects that they think is better for them. In other words, a student will be able to have many subjects of different universities.

At the following sections we are going to explain the content of this approach, where will take into account the current state of the high education studies, the benefits of our proposal and other interesting issues related with it.

II. Current state of higher education

The most common way of doing a high education degree is to study all the subjects proposed by the university's plan for that degree, and doing it assisting to the classes of that specific university. However, the current trend of e-learning is updating the existing model, mixing the face to face classes with online lessons [1][3]. Indeed, some degrees allow students to perform their practices and to read material from their homes through an e-learning platform. In addition, there are diverse kinds of agreements to send students to other countries for studying there, and after that, their passed subjects will be ratified with the local ones in their base university. This program is called Erasmus (European Community Action Scheme for the Mobility of University Students), and as its name says, it consists in generate a mobility environment for students who want to study some of their subjects out of their country to improve their foreign language among other characteristics. It is a beginning of globalization outside of the frontiers, but it still continues being so far to reach what we are talking about.

Therefore, with the introduction of new technologies and the current bandwidth of internet connection we are able to modify our style of doing learning [2]. Thus, everyone can take advantage of the benefits provided by the new models and the new trends.

III. Global Unification Model

The first step to achieve a globalization of higher education contents is to unify the contents into the own country. Once unified inside of a particular country, it is easier to talk about unification outside of it because there are not too many different studies' plans to discuss. Therefore, we are going to start with a local unification to reach a future global unification. However, our main goal is more focused on the students, which will be able to select subjects from many universities to promote the choice freedom and a more flexible learning.

If we take a specific degree from two different places into a specific country like Spain, we can find some changes among them. For this example we are going to take the studies' plan from the University of Castilla-La Mancha and from the University of Barcelona related with the Computer Science degree.

This work was not supported by any organization

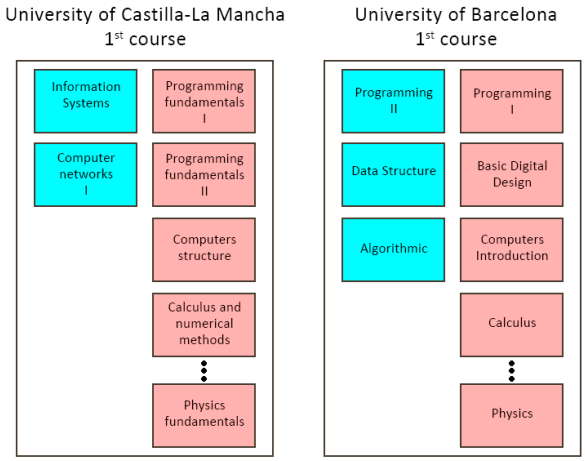


Figure 1. Studies' plan for the first course at the UCLM and UB.

As we can see at the previous figure (Figure 1) there are similar subjects but with different nomenclature. It could confuse students when they want to compare among universities' plans to choose a specific subject. The first approximation to reach the globalization is to unify the subjects, in other words, we have to allow that subjects of different universities but which correspond with the same degree, can be ratified on any university with the same degree. It is the beginning to let the students to choose a specific subject from other university. Therefore, if we get the goal, students will have the freedom to choose in case they want a subject of another university because of the teacher's quality or the methodology followed.

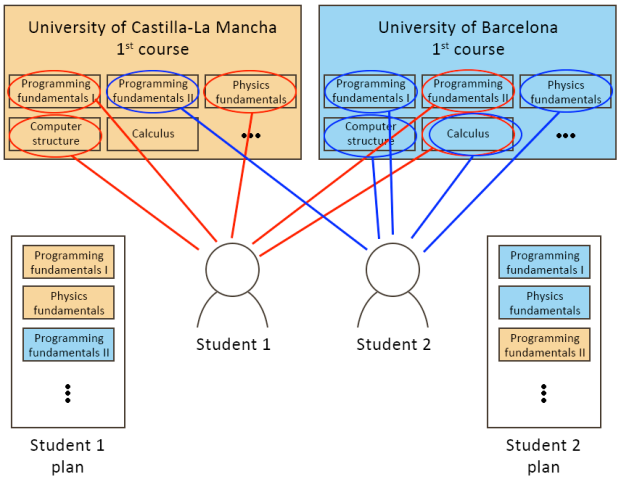


Figure 2. Unification model.

At the figure 2 we can see how two different students (one enrolled in the University of Castilla-La Mancha, and another one enrolled in the University of Barcelona) choose the subjects that they think better for them due to particular circumstances. One aspect to take into consideration is that any student enrolled into a specific university has to select always at least a minimum percentage of subjects of that university, for example the fifty percent. Thus, the student will belong to that university and s/he will be able to take advantage of the benefits promoted by it.

Obviously, each university will have their own specialties, but we are talking about a unification of the bases and main subjects of a specific degree. Thus, the common aspects, which form the degree, will be kept; and each university will propose their own specific areas that make it different [6].

Once we have unified the subjects of a determined country, we can face the challenge of unifying the subjects and plans of the European Union in an easier way. However the scope of this paper is focused firstly on local universities, but in the future we will step forward to reach the European scope [7].

IV. Benefits of the Global Unification Model

The unification of subjects' nomenclature and studies' plans makes easier for students to select them from each university. But the question could be, Why would a student want to choose the same subject from a different university?. Each student has a diverse set of motives that could bring him/her to a determined decision. For example:

- Could happen that a subject is taught by a teacher of another university that explains very well the content.
- Maybe the practices performed in another university are clearer than the practices provided by the current university.
- Due to that more content is explained at the other subject, or in the opposite side, due to that less content is explained at the other subject.
- Maybe the relation with a specific teacher is not as good as a student would want to have and s/he prefers to take the lessons from another person.
- If the student prefer an online learning, it is possible that a subject from another university provides a better experience with that kind of contents.

Previous sentences are related with possible reasons that a determined student could have in mind. However, the student could study the whole set of subjects at the university where s/he is enrolled. This model provides to the students with the possibility of choice, which makes universities' plans more flexible and adaptable to each student [5]. Moreover, depending of the amount of students in a determined subject, it can show that something is happening in relation with that subject. It gives the possibility to the educational institution to pay more attention on it and to find out more about the motives of the lack or excess of students. Therefore, universities can get a lot of conclusions in relation with the enrolments and all the aspects related with them.

This solution is in line with the current trend of emerging technologies and new models of learning related with the online learning [4]. Therefore, it is a modern model that facilitates the access to the university to people that cannot deal with the expenses related with housing or another issues

related with higher studies. In other words, students have the possibility of studying from their homes, saving a big amount of money. Something to take into account is that obviously they have to go sometimes to the university to perform exams or other necessary aspects that cannot be made from home.

The result of promoting the e-learning offers some benefits for both students and university. Some of them:

- A bigger amount of enrollments, due to opening the possibility of studying to more people. It gives more money to the university.
- Less people at class due to the division among offline and online students. It facilitates the personal learning and the university doesn't need big classes to perform the lessons. Fewer students at class allows them to have more attention from the teacher and the labor of teaching is easier for him/her.
- In case of necessity of computers or another type of hardware, with less people the quality of the practices is better because those students have more possibility to use those elements.

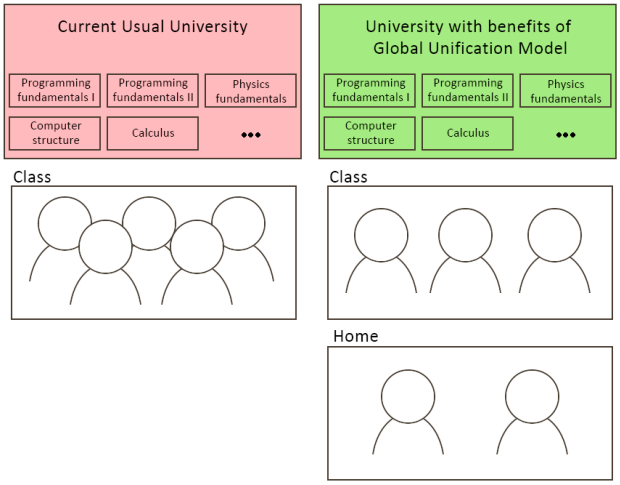


Figure 3. A benefit of the Global Unification Model based on the e-learning.

At the figure 3 we can observe how the density of people at class is lower, which benefits to students and teachers. We also have to take into consideration that each of the students of the university with the Global Unification Model can use his/her right to use the e-learning with his/her own university and obviously it becomes necessary when we talk about the classes of the other university.

However although this model provides some benefits, we have also to take into consideration that teachers have to pay more attention to the online platforms because a part of students use them to make the learning labor. Therefore, they have to get used to the use of these platforms and to the interaction with people outside of the class.

V. Conclusion

Technologies, methodologies and models are evolving to improve the learning; therefore educational institutions have to do the same and don't leave aside this trend. Moreover, the current trend is focused on the globalization and unification of studies plans into the higher education, for that reason we have proposed this model, which pays special attention to improve the flexibility and quality of educational institutions. Students will have more options to choose what they want to study and where they wish to do it. In addition, students that have fewer resources or that don't have the possibility to study in a different place where they live, will be able to make higher studies through internet connection and the e-learning platforms provided by the university. Therefore we are not only globalizing and unifying the subjects and studies' plans, but we are extending the education to the student's home.

In future studies we are going to face the challenge of adapt the model to the European scope. Once unified the model at local scope, the adaptation and unification in a wider environment is easier. However, it will create new issues that have to be faced through negotiations among affected countries inside of the continent. Into these negotiations, the best statistics and results derived from the application of the model proposed will have more weight. Thus, we will have real feedbacks to take into consideration and to improve both the model and the learning methodology.

References

[1] Habib Fardoun, Francisco Montero, Víctor López Jaquero. eLearniXML: Towards a model-based approach for the development of e-Learning systems considering quality. Advances in Engineering Software 2009 Elsevier, pp 1297-1305.

[2] Habib M Fardoun, Antonio Paules Ciprés, Daniyal M Alghazzawi. CSchool-DUI for Educational System using Clouds. Proceedings of the 2nd Workshop on Distributed User Interfaces: Collaboration and Usability. In conjunction with CHI 2012 Conference Austin, Texas, USA, pp 84-695.

[3] H Fardoun, Francisco Montero, V Jaquero. Designing e-Learning Systems to Support new Teaching Techniques. Journal of Computer Science and Engineering, 2010.

[4] Habib M Fardoun, Sebastián Romero López, Pedro G Villanueva. Improving E-Learning Using Distributed User Interfaces. Distributed User Interfaces, Human-Computer Interaction Series, 2011, pp 75-85.

[5] Robert F. Arnove, Carlos Alberto Torres, Stephen Franz. Comparative Education: The Dialectic of the Global and the Local. Rowman & Littlefield Publishers, 2012, pp 472.

[6] David A. Gruenewald, Gregory A. Smith. Place-Based Education in the Global Age. Routledge, 2014, pp 408.

[7] Frank Newman, Lara Couturier, Jamie Scurry. The Future of Higher Education: Rhetoric, Reality, and the Risks of the Market. John Wiley & Sons, 2010, pp 304.

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 819–826

DOI: 10.15439/2014F237
ACSIS, Vol. 2

Benu: Operating System Increments for Embedded Systems Engineer's Education

Leonardo Jelenković, Domagoj Jakobović, Stjepan Groš
University of Zagreb
Faculty of Electrical Engineering and Computing,
Unska 3, 10000 Zagreb, Croatia
Email: {leonardo.jelenkovic, domagoj.jakobovic, stjepan.gros}@fer.hr

***Abstract*—Most of today's computer systems, including rapidly emerging embedded ones, rely on an operating system. Consequently, the development of embedded systems and related software often requires a deeper understanding of operating systems. This paper presents a new incrementally built operating system and a learning course formed around it. Each increment builds on the previous one and introduces new system elements, new concepts and solutions, and a new set of assignments for improving or extending operations or simply demonstrating its use. Increments and assignments are designed to extend theoretical and practical knowledge in the operating system domain, give experience with non-trivial software systems and their development tools, familiarize the learner with basic computer hardware components and demonstrate device driver construction. The audience targeted by this operating system and course materials includes advanced students with (basic) knowledge of computer architecture, programming and operating systems. In addition, materials may be used individually as part of a lifelong learning process.**

I. Introduction

EMBEDDED computer systems are ubiquitous and have become increasingly integrated into our environment, thus increasing the need for engineers with appropriate skills. Developing and maintaining such systems involves hardware and software considerations. The software complexity inherent in embedded systems may vary from a simple controlling program running on a small micro-controller to a complex distributed system. Furthermore, complexity and logical correctness are not enough. For embedded systems, software (as well as hardware) must conform to additional requirements, such as deterministic behavior, dependability, security, low power consumption, long term operation and connectivity. To fulfill such requirements, an engineer must have appropriate skills and expertise. Computer science is a fast-developing domain, and it is common for computer science engineers to expand their skills long after getting their degrees, qualifying them for new challenges. Some of the challenges for the embedded system software developer are discussed in the following.

Software development for a new project may start from scratch or may reuse code from existing projects. Some simple systems may be built from scratch in a short period, but to achieve the required functionality in shorter periods, it is usually better to start with existing elements, e.g., a similar system or a collection of components. If there are no similar systems or useful components to reuse, one may start by selecting the operating system and appropriate development tools that will be used as a base. The operating system and development tools might be selected from the available commercial off the shelf tools (COTS), such as μC/OS-II [1], QNX [2], VxWorks [3], or from freely available tools, such as Linux [4] and and FreeRTOS [5]. Possible disadvantages of COTS systems may include price and possible problems with customization because all system details and internal component sources might not be available. The main problem with freely available systems lies in the absence of technical support (except in the form of free community forums). Whether we use COTS, freeware or a custom-built system, at some point of development, customizing the core system is likely to be required. The reason for customization might be a change in system requirements or a hardware change (e.g., a new or changed device, part or system, prior to the existence of official support). When system customization is required, even if we have the complete source code, an operating system (as for the core of any system) is a complex system, and changes are difficult to implement. Deep knowledge of operating systems is required if the desired change is to be made without compromising the existing functionality.

To be a successful embedded system engineer, one must understand computer architecture, have programming skills and have a deeper understanding of operating systems. The basics of those skills are usually acquired through education. Improvement of these skills is possible with practice and experience. While knowledge of computer architecture and programming language skill is usually improved through courses, operating system expertise is harder to attain. Depending on the instructor and the available options, operating system exercises usually concentrate only on using operating system operations through its interface (and improving understanding in this way). Knowledge obtained in this fashion may be adequate for a regular software engineer but is not sufficient for an embedded systems engineer who might be asked to customize some (operating system) core component.

A deeper knowledge of operating systems might be acquired individually, using the literature and resources from the Internet or through special seminars or courses. In this paper, we present the educational system, Benu, whose source code is freely available [6]. Benu was built primarily for

978-83-60810-58-3/$25.00

education on operating systems in embedded and real-time systems (RT), but it is generic enough that it can also be used for education on operating systems in other areas. The source code is accompanied by a textbook (currently only in Croatian), as Benu is used in the course "Operating systems for embedded computers" (OSFEC) [7].

Remainder of this paper is structured as follows. A comparison with similar systems is made in Section II. Section III presents the basic concepts and ideas behind Benu. The main part of the paper details the contents of the Benu increments, which are presented in Section IV. Benu usability is discussed in Section V. Conclusions are presented in Section VI.

II. Related work

Creating an operating system for education is an old idea and has been performed many times in the past because examples are the best educational tools, especially those examples that the teacher is comfortable using. A review of many such systems, often called "instructional operating systems," has already been presented in several papers, e.g., [8] and [9]. A large number of such educational systems indicate the complexity in teaching this subject and the possibility of many different approaches, emphasis on different aspects of operating systems, different abstraction levels, influences of teacher preferences and the number of students in groups and their competences, i.e., their previous education. In this section, we compare only a few instructional operating systems, primarily highlighting features that are important for those systems when comparing them to Benu.

MINIX [10] is a well-known instructional operating system, modeled on UNIX, which was first introduced in the early 1980s and has since evolved to version 3. In addition to educational use, mostly for understanding UNIX system architecture and micro-kernel concepts, MINIX was later developed for systems with minimal resources, embedded systems, and systems requiring high reliability. A classical operating system textbook [11] details MINIX internals, providing its usage for education.

Linux [4] is not built to be an educational tool. However, because it was free from initial release, it has become fairly popular in academia, and it is often used as a base for student assignments in operating system courses. Because of the magnitude of the Linux source code, the assignments are mostly concentrated on utilizing operations that Linux provides, not on modifying its internal coding. Linux is a complete operating system, built to be effective and used on a variety of computers, primarily targeting personal computers, workstations and servers. Therefore, Linux source code, while freely available, in authors view is too vast and complicated to be used as an educational system for begginers. Only highly persistent students can master Linux complexity and use its internals as part of assignments.

The operating system μC/OS-II [1] is a small system, designated for embedded and real-time systems, with very limited hardware resources. It comes with a companion book, and it is free for educational use. μC/OS-II was primarily created for use in real-time systems, but because it is simple, it is also adequate for education, perhaps more for individual use by enthusiasts than in coursework.

Nachos [12] is a system skeleton prepared for student assignments (in C++) that complete some of the Nachos functionalities. Topics covered by Nachos include thread and process management, paging, file systems and network subsystems. Because solutions from previous assignments are used in the next ones, students are highly motivated to do their best in every assignment. Nachos comes with a MIPS processor emulator for the UNIX environment, which somewhat limits its use. Similar ideas used in Nachos are behind PortOS [13]. PortOS runs in an emulated environment, as a process in a Windows operating system. Assignments prepared for PortOS include multithreading, network and file subsystems.

Nachos and PortOS start with threads and are therefore on a higher abstraction level than Benu. In addition, Benu highlights the building process, building tools and advanced features using C. MINIX, and especially Linux, are complete operating systems, made to be used in more than one role, while Benu is built just for education and research. μC/OS-II is specifically designed for embedded and real-time systems and has many mechanisms for low-level system control exposed directly to user programs. For example, a user program may temporarily disable some kernel components, such as interrupts and scheduling. Although a few of such mechanisms can be found in Benu, we do not encourage their usage; we prefer accomplishing all such operations through the kernel.

III. Benu basic concepts

Benu is a collection of increments that uses a step-by-step presentation of the core operating system operations, data structures and algorithms, where each new increment introduces only a few new subjects. Other educational operating systems, while presenting a single topic, still use the complete system, highlighting related elements from it. In Benu, using increments, the student can focus better on only the subjects introduced in that increment, thus simplifying the learning process of an otherwise very complex system. The evolution of operating system components is roughly presented through increments, starting with the basic functionality in one increment and adding extended functionalities as they are needed. In addition to operating system topics, using Benu in education might improve the other skills required for embedded system development. These skills include the advanced use of the C programming language, experience with development and debugging tools and methods, and familiarization with POSIX for real-time and embedded systems.

For education, Benu can be used without supervision, simply progressing through prepared assignments. Better results can be achieved faster if assignments are preceded with some theoretical introduction, e.g., that presented in [7], with topics covered, such as those presented in [14], or any operating system textbook, e.g., [15], [11] or [16]. Source code dissection, assignments and other experiments should follow theoretical

introductions to broaden students' understanding. Every increment in Benu is associated with example assignments. New components, methods or principles are carefully chosen and added such that each new major increment brings the proper number of new elements for ease of understanding. Some concepts that are more radical and complex, such as threads and processes, are introduced in several smaller increments. Changes from one increment to the next can be easily tracked using text or graphical tools, such as Meld [17].

The current version of Benu is prepared for both Intel i386 and ARM platforms. Although the i386 platform is not typical for embedded systems, it has the advantage of providing educators with access to development tools, emulators, computers and documentation. Adding support for other processors is supported by the layered architecture of Benu, with a separate hardware abstraction layer. As a development platform, Linux is selected because all of the required tools, i.e., GNU development tools [18], are freely available and easy to install on Linux. Linux itself may be run as a development platform in an emulated environment, requiring only the emulator to be installed on the host computer (if the host computer is not already Linux-based).

Based on our experience in teaching computer architecture, programming languages and operating system basics as well as using Benu in an advanced operating system course, we observed that using Benu accomplished the following:

- A deeper understanding of the operating system, its components, data structure, operations, algorithms and limitations,
- Improvement of advanced programming skills, which in turn produces shorter, more efficient, extendable and more readable (and thus reusable) code,
- An understanding of the capabilities of developing and debugging tools and computer emulators,
- The ability to build embedded system software from scratch, not relying on any operating system interface, and preparing images to be loaded into systems with variety of memory configurations,
- Expertise with POSIX interfaces for timers, threads and signals for real-time and embedded systems,
- The ability to navigate within and use larger source code projects (written by others), the discovery of usual concepts and practice in source code naming, file structure and management tools, and
- An improvement in the student's problem-solving skills.

We do not claim that Benu is the best choice in the field of embedded operating system education, but it may be among the easiest for beginners. Starting increments are simple and do not require preparation. Students can start early, familiarize themselves with the development environment and be prepared for the more demanding increments that follow.

IV. Contents of increments

Benu is created using basic operating system principles as a base [15], [16] and is modified to better suit the embedded system environment, simplified for educational purposes, and built for incremental topic introduction. Benu contains eight major increments in the source code, which, for the rest of this section, are just "increments" or "chapters". Each chapter has several minor increments, depending on the complexity of the subjects presented in the chapter, as shown in Table I. Details about each chapter, its purpose, the components it presents, and possible assignments, are presented in the following sections.

TABLE I
Major and minor increments in Benu

Major increments – chapters	Minor increments
Chapter_01_Startup	01_Startup
	02_Example_clock
Chapter_02_Source_tree	01_Source_tree
	02_Console
	03_DEBUG
	04_Debugging
Chapter_03_Interrupts	01_Exceptions
	02_PIC
	03_Dynamic_memory
	04_Interrupts
Chapter_04_Timer	01_Time
	02_One_timer
	03_Timers
Chapter_05_Devices	01_Devices
	02_Keyboard
	03_Serial_comm
Chapter_06_Shell	01_Shell
	02_Arguments
	03_Programs
	04_Makepp
Chapter_07_Threads	01_User_threads
	02_Threads
	03_Ext_context
	04_Synchronization
	05_Messages
	06_Signals
	05_Sched2
Chapter_08_Process	01_Syscall
	02_User_mode
	03_Programs_as_module
	04_Programs_as_process
	05_Static_processes
	06_Processes

A. Chapter 01 – Development environment

The goal of Chapter 01 (with related materials from the textbook) is to present the environment used for developing system software, i.e., Benu, which will run on bare-bone hardware (real or emulated). Compiling and running system software requires special steps during the compilation and linking phases, supported with appropriate configurations, and is thus significantly different from compiling and running application programs. Because of this straightforward goal, the code is purposely simple; it just displays the "Hello World" message on the console.

There are only three files in Chapter 01: two with source

code (one in assembly and one in C) and one with shell script used for compiling. The assembly code is small, but it is required for processor initialization (stack pointer and status register). The assembler then transfers control to a function written in C. A single function is placed in a C file, i.e., a function that writes a text string into video memory, thus displaying it on the console. While low-level operations implemented in assembly code and device drivers might be interesting to some, they are not essential for using and understanding Benu and the principles it describes. It is possible to learn most operating system concepts without a knowledge of assembly language or device driver details; therefore, assembly and device drivers are placed into a separate directory (from the next chapter) to isolate them from the other increments and lessons. To compile and link the source codes into an executable and run it, a shell script is used only in this chapter. A shell script better reflects the necessary steps involved and therefore better serves the purpose of this chapter, which is to show how the appropriate tools are used. The script illustrates how to start the compiler and linker and how to create the system image and start emulator, using all necessary flags and parameters. Beginning with the next chapter, Benu uses standard build tools, i.e., `make` with appropriate definitions across `Makefiles`.

The second minor increment of the first chapter demonstrates how very simple systems can be implemented without having an OS in a traditional sense. For that purpose, a simple clock is implemented that uses hardware timers and displays a counter on the screen.

Assignments for the first chapter should include only the preparation of the development environment, e.g., installing the required tools, downloading Benu, and compiling and running it in an emulated environment. The first few assignments should also focus on using a revision control tool, such as Git [19] or Subversion [20].

B. Chapter 02 – Layered structure

The instructional goal of the second chapter is to demonstrate the modular and layered approaches to operating system design. To reduce complexity, systems are often developed modularly, using the "divide and conquer" approach. A single module or component is simpler than an entire system, and it is thus easier to develop and test. Furthermore, modules can be concurrently developed by different programmers, reducing development time.

Operating systems are designed and built in a modular manner; each module is a "subsystem". Typical subsystems include input-output, memory management, thread and process management, network, security and file subsystems.

Operating systems are also layered, and the layers communicate with one another using strictly defined interfaces in a top-to-bottom fashion, i.e., a higher layer uses the services of the immediately lower layer only. One of the purposes of introducing layers is to separate smaller, architecture-dependent code that sits immediately above the hardware, often called the "hardware abstraction layer" (HAL), from the larger architecture-independent code in the layers above. This separation makes it easier to port the operating system across different architectures because only the HAL has to be modified or rewritten. The architecture-independent code consists of several layers, one built upon another. The first layer, called the "kernel", is immediately above the HAL and provides the core operations for system resource management. System resources include hardware components, such as devices, and software components, such as synchronization objects and the task scheduler. The kernel implements fine-grained operations that manipulate system resources and provide an interface for accessing them from the layer above, i.e., the application programming interface (API) layer. The API layer (henceforth referred to as "api") uses the kernel layer to implement any designed interface (e.g., POSIX [21]) for user programs, i.e., the next higher layer. User programs use the "api" to implement operations that are required by the user (which is the highest layer of the operating system).

Benu adheres to the described layered approach with four layers, named after directories where their code is placed: arch (HAL), kernel, api and programs. There is also an additional pseudo-layer, the library layer ("lib"), which captures the utility functions used by the other layers, e.g., string manipulation and list operations. A list of all of the layers as well as their containing directories in Benu is shown in Fig. 1.

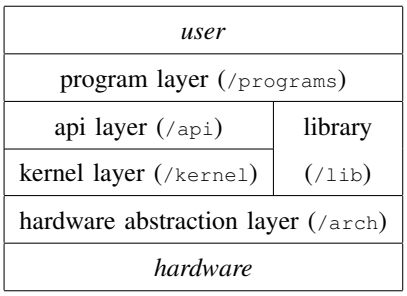


Fig. 1. Layers in Benu

In Chapter 02, the layers contain no functionality: the corresponding directories are mostly empty, and they are just placeholders for the components that will be implemented in subsequent chapters.

Each layer has a clearly defined interface for the higher layers. Fig. 2 shows an interface example introduced in Chapter 04, when the user program uses the `clock_nanosleep` operation. The function `clock_nanosleep` is defined in api as a system call to the kernel function `sys__nanosleep`. The kernel function uses the HAL interface `arch_timer_set` which calls `i8253_set_time_to_counter` from the device driver (via the timer interface function `set_interval`).

```
programs: [api]     clock_nanosleep
kernel:   [kernel]  sys__clock_nanosleep
          [kernel]  ktimer_settime
          [arch]    arch_timer_set
arch:     [arch]    timer->set_interval
          (=>i8253_set_time_to_counter)
```

Fig. 2. Interface chain for operation `clock_nanosleep` (in Chapter 04)

Interfaces that one layer provides for the next one are defined in header files placed in the `include` directory (placed in top level, same as those from Fig. 1).

The naming convention used throughout Benu follows certain prefix rules. Kernel functions that provide the interface to the higher layer (i.e., to api) have the prefix `sys__`; internal kernel functions (i.e., the ones used only within the kernel) are prefixed with `k_` or just `k`; functions that are part of HAL are prefixed with `arch_`; and the device driver's functions are prefixed with a short device name.

Devices and subsystems are used through a predefined interface. Chapter 02 defines the interface for printing characters on the console, i.e., `console_t` (defined in `include/types/io.h`) with the following elements:

```
int (*init)(void *p);
int (*clear)();
int (*gotoxy)(int x, int y);
int (*print)(int attr, char *text);
```

Each structure element is a function that implements a specific operation. The same functionality may be achieved by different devices or components. For example, for simple console display, a graphics card can be used, as in Chapter 02, or a serial port connected to a terminal, as in Chapter 05. Switching from one console to another is accomplished using different objects (variables) that implement the same interface. In Chapter 02, in `kernel/startup.c`, consoles for kernel and user programs are selected using variables `k_stdout` and `u_stdout`, both referencing variable `vga_text`, defined in the device driver file `arch/devices/vga_text.c`.

The same principle for defining an interface is frequently used in Benu, i.e., using the structure with functions and parameters like `console_t`. Interfaces are defined for the interrupt device (Chapter 03), timer devices (Chapter 04), general devices (Chapter 05), and dynamic memory allocators (Chapter 03). Interfaces make separations easier, replacing one device or component with another simpler and source code more readable.

Chapter 02 also presents possibilities for tracing and debugging. Because the system being created uses (emulated) hardware directly, debugging is harder than debugging a traditional program which can be paused and inspected at any moment. One primitive debugging method is to insert print commands in the source code, e.g., with the `printf` operation or with `LOG` and `ASSERT` macros that will be executed only in the `DEBUG` mode. Another method is to use the appropriate tools that enable the developer to stop (and examine) the system while executing, such as the GNU debugger and the QEMU emulator [22], which are used in the demonstration example.

Assignments for Chapter 02 can be focused on layered architecture and on debugging. For example, the assignment can be to divide the console into two parts, one for kernel messages and the other for program output. Implementation through two different `console_t` objects, almost the same as the one provided, will require little coding but will need implementation in nearly all layers. Debugging can be practiced by requesting stops and system inspections at defined points or by discovery of covertly inserted bugs that cause system failures.

C. Chapter 03 – The interrupt subsystem

Interrupts are very important mechanisms, not only for managing devices but also for other purposes, e.g., protection from system calls, thread scheduling (timer interrupts), memory management and program failure detection. The instructional goal of this chapter is the presentation of interrupt handling methods in the operating system. The presented material is, by necessity, simplified and therefore has some limitations, but it also offers space for possible extensions using student assignments.

The primary function of the interrupt subsystem is to provide an interface for connecting the interrupt handler functions (which might be part of other subsystems) with interrupts, i.e., the functions that will be used as interrupt handlers. The interface for registering the interrupt handler function `hnd` with an interrupt identified by number `id` is defined in HAL as follows:

```
arch_register_interrupt_handler(id, hnd);
```

For every source of interrupts identified by an interrupt number, at least one handler function can be defined. When the interrupt occurs, all handlers for that interrupt are called sequentially.

Benu uses an Intel 8259 programmable interrupt controller (PIC) in HAL through the `arch_ic_t` interface, making future replacements with other interrupt controllers (e.g., APIC) easier. Chapter 03 defines the interrupt subsystem but handles only the processor's interrupts because no other device driver is used in Chapter 03. Device drivers are added in ensuing chapters, i.e., the timer in Chapter 04 and the keyboard and UART in Chapter 05.

Registering more interrupt handler functions for a single interrupt number can be accomplished using a static data structure (i.e., an array with predefined size for each interrupt number) or a dynamic data structure, such as a list. A list provides more flexibility and less overall memory consumption, but requires dynamic memory management. Because dynamic memory management will also be required for other subsystems, it is introduced in this increment. Two algorithms for dynamic memory management are presented: the simple "first fit" (FF) method with a "last in, first out" list of free blocks and the more complex "two level segregated first" (TLSF) method [24]. FF is simpler, and, on average, faster than TLSF. However, TLSF provides reduced fragmentation; and more importantly, the execution time complexity of TLSF is $O(1)$, while the worst case for FF is $O(n)$, where n equals the number of free blocks. Therefore, TLSF is a candidate for use in real-time systems.

Student assignments for Chapter 03 include improvements to the interrupt subsystem, such as adding priorities to existing interrupt handlers. Then, when multiple interrupts overlap,

they can be handled according to their priorities. Other assignments can be focused on implementing additional dynamic memory management algorithms, such as "best fit".

D. Chapter 04 – Time management

Most program activities, especially in embedded systems, must be executed in a timely manner. Therefore, the operating system must provide support for time management through a timer subsystem. Most required operations include system time control ("set" and "get" system time), thread execution delays and programmable future actions, i.e., alarms. The operating system also uses time for managing input/output devices, scheduling and maintenance.

The timer subsystem implemented in Benu consists of two components: a lower-level component implemented in HAL and a higher-level component implemented in the kernel. The component implemented in HAL uses an Intel 8253 programmable interrupt timer (PIT) through the `arch_timer_t` interface. The primary operations provided by that interface include keeping system time and a single alarm, which, upon alarm expiration, forwards a call to the kernel. The component implemented in the kernel extends capabilities to multiple alarms available to the kernel and programs and provides operations for program delays.

Assignments for Chapter 04 may include extensions of the timer subsystem. For instance, absolute times can be changed to relative times or the sorted list of alarms may be replaced with some more efficient structure; monotonic clock can be added to the system, a clock that can't be changed with `*set*` interface (as current real-time clock can); a software watchdog timer can be implemented; other more advanced hardware devices than the Intel 8253 can be used to achieve better resolution for time management.

E. Chapter 05 – Device interface

Every device in a computer has its own specifications. To simplify device management, devices are grouped into classes, and an interface is defined for each class. When creating a device driver for a device, the appropriate interface must be implemented. The simplest device driver interface must include functions for sending data to the device and functions for reading data from the device. Such an interface may not be as efficient as a more complicated interface that, for example, uses direct memory access capabilities of the devices, but it is a good starting point for illustrating the integration of device control into an operating system. For that reason, Benu uses a simple interface defined by the structure `device_t` (defined in `include/arch/device.h`) with the following functions:

```
int  (*init)(uint flags, void *parm, device_t *dev);
int  (*destroy)(uint flags, void *parm, (...) );
int  (*send)(void *data, size_t size, (...) );
int  (*recv)(void *data, size_t size, (...) );
void (*irq_handler)(int irq, void *dev);
int  (*callback)(int irq, void *dev);
```

The usage of the `device_t` interface is demonstrated on three devices for which the device driver is prepared within Benu: the display driver (replacing `console_t`), a keyboard driver using the Intel 8042 controller, and a serial port using the 16550 UART device. The interface `device_t` is intended for use only within the kernel. The kernel allows programs to access these interfaces indirectly through `sys__device_*` system calls (`*open`, `*close`, `*read`, and `*write`).

Assignments for this chapter may include introducing the program (thread) blocking state to read/write operations until they are completed on a device. Other assignments include adding device drivers for other devices or improvement to current devices, e.g., adding scroll history capabilities to the console display driver.

F. Chapter 06 – Command shell

The interfaces offered to users on today's computer systems range from graphical interfaces, with buttons or menus, to console-oriented interfaces, such as the command line interface, where the user types commands to be executed. Because Benu only has a text-based console, a command-line interface is implemented and presented. From an educational point of view, the implementation of a command line user interface is useful for two reasons: parsing the command line and sending parameters to programs (as strings, without interpretation).

The second novelty introduced in Chapter 06 is in the compile script (`Makefile`). Every program from the `programs` folder may be independently included in or excluded from compilation. This change further distances the program layer from the kernel, making the kernel (and thus HAL and api) potentially usable for many purposes.

Assignments for this chapter may include improvements to the shell program, e.g., adding history and auto-complete features.

G. Chapter 07 – Thread management

The systems presented in the previous chapters or the systems based on them (e.g., created as assignments) are very simple. Still, they may be sufficient for numerous embedded systems. More complex systems require additional features, such as multithreading and processes. Introducing those features has a strong impact on all of the system components, making the system significantly more complex and larger, and thus it is not recommended if those features are not required by the embedded system.

Multithreading support simplifies complex system implementation. Independent tasks may be run independently as threads, with their own timings and resource requirements that are more easily coded and satisfied at runtime.

Based on our teaching experience, we believe that multithreading programming is one of the most difficult subjects in computer science education. Thinking "in parallel" is required, and any shared resource must be considered and properly protected. Synchronizations via semaphores and monitors have to be carefully designed to achieve desired sequences and avoid deadlocks or simultaneous changes on

any shared resource. Multithreading is increasingly important because modern processors are multicore and manycore and require multithreading for using all of the processing power the processors can provide. Thus, many operating system and programming courses emphasize multithreading with the other subjects.

The multithreading covered in Benu includes both lower-level kernel operations, such as thread creation with resource allocation and context switching, and higher-level operations, such as scheduling, synchronization and communication. A priority scheduler is used with "first in, first out" as the second level scheduling criteria (for threads with equal priority). Semaphore and monitor synchronization mechanisms are included, and communication is provided through messages and signals.

Many assignments can be created for Chapter 07. Threads can be used for kernel operations, e.g., in an interrupt subsystem for processing individual interrupts. Existing thread operations can be improved or extended. Semaphores and monitors may be extended with "try-wait" and "timed-wait" operations or improved with priority inheritance protocols. New synchronization and communication mechanisms could be added, such as barrier, read/write locks and pipes. Program assignments that solve some synchronization problems can also be created.

H. Chapter 08 – Process Management

Programs, especially more complex ones, may have bugs that might compromise the system. In a multitasking environment, there should exist mechanisms that protect the kernel and other tasks from a malfunctioning task. The usual protection mechanisms include processor operation modes and memory protection, both requiring hardware support from the processor. If the system is running in unprivileged processor mode, the thread may not be able to execute instructions that could compromise the entire system. For example, if a thread cannot disable interrupts but must instead use a synchronization function for a critical section, the eventual error that leads to an infinite loop in a critical section will only affect that thread and other threads that use the same critical section object, while the rest of the system will be unaffected. The same reasoning is true with memory separation methods, such as memory protection and virtual memory. If a thread cannot change memory locations outside its defined boundaries, it cannot compromise kernel data or other programs.

Grouping threads that work on the same operation into a single process (i.e., threads that are created within the same instance of a single program) will isolate them from other threads, and vice versa. An error in one thread will usually have only a local effect on threads in the same process. Errors that compromise a shared system resource, like a device, will, however, still be an issue for the entire system.

Chapter 08 brings a further separation of programs and the kernel by introducting the privileged and unprivileged processor modes, forcing software interrupts as mechanisms for calling kernel functions (syscalls). Additionally, memory protection is introduced using segment registers of the Intel 80386 processor family for simple virtual memory implementation. Threads use logical addresses and cannot reach outside the boundaries of their processes. Any attempt to do so will trigger an interrupt, and the thread will be terminated. Compiling programs using logical addresses complicates the building process. Since kernel and programs must be prepared for different locations (physical and logical) they are separated into different objects after compilation. To simplify emulation in those increments GRUB was used as boot loader where program objects can be prepared and loaded as modules.

Assignments for Chapter 08 might be same as those for Chapter 07 because major differences in the chapters are in the implementation of the syscall mechanism via a software interrupt with the address space changing from the user to the kernel, from logical (process) to physical address space. Both changes require special syscall parameter handling, which provides a sufficient challenge to adopt.

V. Using Benu

Even in the last increment (that includes all) there are 56 source code files (.c) and 73 header files (.h) (including the 16 example programs). Furthermore, about half of them are only for layering purposes (parameter checking and forwarding call to lower level function). The combined source code of the kernel, HAL, library and api (all except headers and example programs) have approximately 7,500 lines of code (as counted by the `cloc` program). A system this small cannot have advanced components, such as paging, file systems and networking, which are required in more complex systems. However, some simpler systems, such as the ones used in embedded computers, may use an operating system like Benu because they may not need advanced components at all. Future work on Benu includes developing those advanced components, though in some minimalistic form that has yet to be devised. Otherwise, such complexity will significantly reduce its educational value. The components present in Benu are built on basic principles, avoiding too many complexities that may have better properties. From an educational viewpoint, this approach leaves the basic course straightforward and allows for advanced student assignments and projects.

Although Benu is built on somewhat different ideas than Linux and MINIX, it can be a good prelude to studying them. Because Benu is not a complete operating system, it can provide a simpler example for embedded systems that do not need advanced components. Due to their simplicity, Benu source codes do not have as complex interconnections in their kernel as real systems have (e.g., Linux). Compared to other systems, the components in Benu are easier to change and replace, and new components are easier to integrate and evaluate, thus making Benu also usable in operating system research. For example, current synchronization mechanisms can be changed and extended with other models (e.g., [23]), priority inheritance and priority ceiling protocols can be embedded, deadlock detection can be implemented, thread scheduling can be upgraded, and interrupts can be processed

with threads [24]. As example extensions, "round robin" (RR) and "earliest deadline first" (EDF) schedulers are implemented and presented in Benu.

In addition to Benu source code, a companion textbook is available to students (currently only in Croatian). The textbook is tightly coupled to Benu but includes theoretical explanations of operating system topics, excluding advanced components such as networking and file systems. A quick start with Benu is possible with a basic knowledge of computer architecture and operating systems, and a moderate knowledge and experience in the C programming language. Therefore, a Benu course should best be placed after courses that cover fundamentals. In addition, Benu could be learned as a post-graduate, as part of the lifelong education process. Benu targets students interested in operating system internals and experimentations with it and students interested in software development for embedded computers.

Benu has been used as a teaching tool since the 2009/2010 academic year, when OSFEC was offered as an elective course in master computing science studies. Most students who take this course show much interest, and most of them successfully complete the exams. However, for some students, the assignments and exams are harder to complete. After an analysis, it was found that the students who had problems did not have the recommended prerequisite courses in bachelor studies, especially the courses that exercise C programming skills.

VI. Conclusion

Mastering operating system topics, including theory, implementation details, tools and common practices, can be faster and more interesting for students if a simple system such as Benu is used in teaching and assignments. The uniqueness of Benu among other instructional operating systems is in its incremental build structure, allowing gradual introduction of operating system components. Each increment logically extends the previous one with only a few new elements, which are, consequently, easier to adopt.

An operating system is a complex system, and, even with simplifications, as in Benu, many students may find it difficult to master. However, the majority of students do not need to master all of the components of an operating system. For some, it will be enough to master the kernel layer or even just a specific components, like the interrupt subsystem, timer subsystem and threads. For others, Benu may be just a starting point, one step toward understanding more complete and complex systems.

A quantitative comparison of Benu with other (instructional) operating systems is not performed. To perform such a comparison, we cannot just use the other systems, but we would also need to prepare teaching materials and assignments closely coupled with them, as the current OSFEC materials are coupled to Benu. Nevertheless, based on our experience with Benu, we can conclude that Benu offers a great deal for independent study and exercise and provides a base for faster learning, not only in the operating system domain but also in embedded system software development.

Acknowledgment

This work was supported by FP7 project Embedded Computer Engineering Learning Platform (E2LP).

References

[1] J. J. Labrosse, *MicroC OS II: The Real Time Kernel*, 2nd ed. CMP-Books, 2002. ISBN 1578201039
[2] QNX operating systems. [Online]. Available: http://www.qnx.com/products/neutrino-rtos/
[3] Wind River VxWorks. [Online]. Available: http://www.windriver.com/products/vxworks/
[4] The Linux kernel archives. [Online]. Available: http://www.kernel.org
[5] FreeRTOS. [Online]. Available: http://http://www.freertos.org/
[6] L. Jelenković. (2012) Benu source code. [Online]. Available: https://github.com/l30nard0/Benu
[7] ——. (2010) Operating system for embedded computers, course homepage (in croatian). [Online]. Available: http://www.fer.unizg.hr/en/course/osfec
[8] C. L. Anderson and M. Nguyen, "A survey of contemporary instructional operating systems for use in undergraduate courses," *J. Comput. Sci. Coll.*, vol. 21, no. 1, pp. 183–190, Oct. 2005. [Online]. Available: http://dl.acm.org/citation.cfm?id=1088791.1088822
[9] Y.-P. Cheng and J.-C. Lin, "Awk-Linux: A lightweight operating aystems courseware," *IEEE Trans. Educ.*, vol. 51, no. 4, pp. 461–467, nov. 2008. doi: 10.1109/TE.2007.912571. [Online]. Available: http://dx.doi.org/10.1109/TE.2007.912571
[10] MINIX 3. [Online]. Available: http://www.minix3.org
[11] A. S. Tanenbaum and A. S. Woodhull, *Operating systems design and implementation*, 3rd ed. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2006. ISBN 0131429388
[12] W. A. Christopher, S. J. Procter, and T. E. Anderson, "The Nachos instructional operating system," in *Proceedings of the 1993 Winter USENIX Conference*, 1993, pp. 479–488.
[13] B. Atkin and E. G. Sirer, "PortOS: an educational operating system for the post-PC environment," in *Proceedings of the 33rd SIGCSE technical symposium on Computer science education*, 2002. doi: 10.1145/563517.563384 pp. 116–120. [Online]. Available: http://dx.doi.org/10.1145/563517.563384
[14] C. Yang, "Computer operating systems in electrical engineering curriculum," *IEEE Trans. Educ.*, vol. 36, no. 1, pp. 177–180, 1993. doi: 10.1109/13.204841. [Online]. Available: http://dx.doi.org/10.1109/13.204841
[15] A. Silberschatz, G. Gagne, and P. B. Galvin, *Operating system concepts*, 8th ed. Wiley, 2011. ISBN 1118112733
[16] L. Budin, M. Golub, D. Jakobović, and L. Jelenković, *Operating systems (in Croatian)*, 3rd ed. Zagreb: Element, 2013. ISBN 9789851976107
[17] Meld: Diff and merge tool. [Online]. Available: http://meld.sourceforge.net
[18] GNU operating system. [Online]. Available: http://www.gnu.org
[19] Git (distributed version control system). [Online]. Available: http://git-scm.com/
[20] Apache Subversion (version control system). [Online]. Available: http://subversion.apache.org/
[21] IEEE and O. Group. The open group base specifications issue 7. [Online]. Available: http://pubs.opengroup.org/onlinepubs/9699919799/
[22] QEMU: open source processor emulator. [Online]. Available: http://wiki.qemu.org
[23] P. A. Buhr, M. Fortier, and M. H. Coffin, "Monitor classification," *ACM Comput. Surv.*, vol. 27, pp. 63–107, March 1995. doi: 10.1145/214037.214100. [Online]. Available: http://dx.doi.org/10.1145/214037.214100
[24] S. Kekckler, A. Chang, W. Chatterjee, and W. Dally, "Concurrent event handling through multithreading," *IEEE Trans. on Computers*, vol. 48, no. 9, pp. 903–916, 1999. doi: 10.1109/12.795220. [Online]. Available: http://dx.doi.org/10.1109/12.795220

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 827–833
DOI: 10.15439/2014F257
ACSIS, Vol. 2

Experience with Real-Life Students' Projects

Jaroslav Král
Masaryk University in Brno,
Botanická 68a, 602 00 Brno, Czech Republic
Email: kral@fi.muni.cz
and
Charles University in Prague
Malostranské nám. 25, 118 00 Praha 1, Czech Republic
Email: kral@sisal.mff.cuni.cz

Michal Žemlička
Charles University in Prague
Malostranské nám. 25, 118 00 Praha 1, Czech Republic
Email: zemlicka@sisal.mff.cuni.cz
and
University of Finance and Administration
Estonska 500, 101 00 Praha 10, Czech Republic
Email: michal.zemlicka@post.cz

***Abstract*—Student software projects are often focused on training coding skills and on model-driven software system design. The projects rarely develop skills needed for the proper formulation of system visions and requirements specifications. To solve this issue the projects must deal with real-life software projects issues. The projects should solve main commercial aspects of real-life – they must include looking for project topics in practice and there should be possible to communicate and collaborate with future project users. Successful projects should be rewarded (optimally paid) by the users like other commercial products. We discuss here the quite successful experience with a "prototype" implementation of the concept.**

I. Introduction

THE SOFTWARE development is a risky process. The proportions of failed and challenged projects is about 1/4 and 1/2 respectively [1], [2], [3]. It is known that it is due to poor quality of project visions (aims) and project requirements specification. Late stages of software development processes are seldom the source of the issues.

The development issues are besides the management failures caused by underestimation of the complexity of the vision and specification stages.

A detailed analysis of the problem indicates that the development team members are in fact not aware enough of the software engineering aspects of the projects, i.e. that the development of software systems is a technical (engineering) problem. It holds, surprisingly enough, not only for users (project sponsors and project stakeholders inclusive) but also for IT experts taking part in the projects. They all must be aware that a seemingly simple requirement need not be implemented and used due the technical aspect easily.

Our experience from academia as well as form industry shows the reasons for failures and challenges in software projects are often caused by IT experts unable to effectively and properly take part in project vision and requirements specification.

Young people are often excellent in coding. It especially holds for young people active in theoretical disciplines. They know very well that their coding skill is difficult to overcome. This fact is overemphasized. Soft software development skills are difficult to develop. It holds especially for the people trained in hard skills needed for academic or scientific career. They are excellent coders but usually not excellent in design and especially in requirements specification. They are often proud on their intellectual abilities and scorn the users being unable to write and test programs. They underestimate or disregard the importance and the complexity of real-life problems. They are unable to take part in such agile software development processes when they must take users as partners. They even do not admit that they should take part in reqirements specification. They are unable to tune specifications in cooperation with users. We will discuss our experience with student projects aimed to solve the above issues at IT side. The projects were a part of postgradual studies.

The issues can be solved if IT students take part in team projects covering very early stages of software development – marketing, vision formulation, and requirement specification. The students should cooperate with prospective users of the developed systems. It implies that the students project can fail. We know that real-life projects fail frequently.

We will discuss here the structure of student projects we have used to solve the issue and present experience with them.

The paper is structured as follows: First, we discuss the problems discussed above in details. Then we describe two variants of the projects intended to train analytical skills. Finally, we summarize the experiences gained in the projects.

II. The Soft Facts on Project Failures and Challenges

The proportion of failed projects is remarkably stable, the proportion of failed projects remains very high for decades – see the surveys Why Projects Fail [4] or the Chaos Report [3]. The reasons are not solely at project stakeholder and project management side.

Our experience indicates that the issue is a bit more difficult and complex. People involved in software system development, maintenance, and use are not aware enough that software systems are complex technical entities requiring engineering attitudes. All developers, IT experts inclusive, are not ready to apply software engineering knowledge, attitudes and processes. IT experts are then not able to convince users

978-83-60810-58-3/$25.00

and managers that the application of software engineering is necessary and that requirements specification is no straightforward matter.

We must conclude that the project failures are in this sense due to IT experts more frequently than it is presented in project failure analyses. It is quite difficult to change this attitude as developers, especially coders, dislike developing and applying needed skills. Model-driven architecture and related attitudes are helpful partly only.

Our experience with the students' projects indicates that our students like and are able to propose an elegant solving of technical problems. They are often not interested in seemingly boring problems of real life. They, for example, are able to design a very complex SQL statement but they are not aware that a small modification of it could be substantially more usable in real-life projects.

We also experienced that people tending to be managers underestimate the fact that a more detailed technical knowledge could be very useful for them too. It concerns especially the cases when managers make decisions having important technical consequences.

III. Education and Analytical Skills

We can conclude that the education and mastering of soft knowledge and skills necessary for analytical tasks in software development is of growing importance but it does not meet needs. IT students and even some of their teachers do not understand the importance of the issue. They often consider analytical documents to be something like gossips whereas the only valuable thing is a big piece of (good) code. This attitude is supported by the fact that young people are often very good coders.

The poor understanding of the importance of the early stages of software development is the crucial reason of lower quality of analysis oriented student knowledge. The issue is further strengthened by the fact that the needed development skills are not trained enough. It leads to the lack of good project management and to poor results of agile methods of software development.

We discuss here possible ways of solving this issue. The training of necessary social skills is a key issue here. The training can be realized only if the requirements specification and, if possible, the vision and market analysis are trained in students' team projects. The goals of the projects should be the development of small information systems or the development of their autonomous parts.

The projects should be designed so that they meet the needs of real-world people and in collaboration with them, i.e. they should be small commercial projects. The development process should have as much common as possible with standard development of information systems in small software firms – market analysis, business, and technical risks inclusive.

We present below the main principles of the implementation of the above requirements and the experiences with the projects at two Czech universities. We further discuss issues of our attitude to be still solved. The most important problem of the projects is the trend to the fragmentation and extreme specialization of scientific knowledge and of education processes, and a growing gap between hard a soft knowledge of our alumni. They are not trained to be good project managers as recomended, e.g., by Ebert [5].

IV. Information Systems and Professional Knowledge

The basic goal of the postgradual study of IT experts is the education of professsionals being able to propose, develop, maintain, and execute software, usually information systems. The development of such systems are enhanced variants of classical waterfall model [6] consisting of:

1) Analysis done in cooperation of the system developers with future system users: by formulation of visions (answering the question "why?" and outline of the basic requirements) and by requirements specification,
2) design,
3) coding,
4) testing:
   - unit testing,
   - integration testing,
   - function testing,
   - system testing;
5) integration,
6) deployment.

The scheme can be enhanced and modified in various ways, see iterative development (agile development, scrum [7], XP [8]) and incremental development (various SOA methodologies [9], [10], [11] inclusive). The variants of development can be viewed as repeated and integrated waterfalls.

Knowledge and skills needed in the initial stages of software development are often missing. It is known [3], [12] that the user involvement and management issues are crucial for the success of information systems development. The faults in these stages of the development can cause more than 80% costs spent to remove development errors in not failed projects. Almost half of software projects significantly overrun expenses and terms.

The overall losses caused by the errors are significantly higher. For example conceptual errors are the main reason of most project failures. According to [3], [13] it happens to about $1/4$ of projects. These misconceptions usually stay behind poor maintainability of software systems and behind their early or permanent obsolescence [14].

There are yet other snags: The conceptual errors can hiddenly lead to a user discomfort or can even reduce their long-term wellness feeling and even their health. It reduces productivity and worsens social climate in firms. It destroys firm culture and threatens its future.

Coding and testing are usually without substantial issues. The education of the skills needed for coding and testing is therefore successful.

It is the reason for looking for new methodologies and development paradigms like the concept of SOA ecosystem by

OASIS. The stages 1, 2, and partially 5–7 require according [15], [1], [3] and the experience of software developers:

- collaboration of developers with users,
- support by management,
- requirements specifications specific abilities of the involved team,
- etc.

It requires a collection of soft skills and, what is more important, the skills must be trained.

The students should take part in the early development stages of projects designed so that they are "approximations" of real-world projects usually for small to medium enterprises (SME). The knowledge (often even sole information) collected in books need not be sufficient. The needed knowledge should be broad open and flexible to be usable for the specification of user needs.

Some services in SOA can be developed (may be as prototypes) as students projects. It is also possible to use sprint subprojects of SCRUM or XP methodology. It is a crucial proposition as real world projects must be able to bring some real world effects. They should incorporate the vision and even some marketing. It is a very hard task in academia environment. We must cope with the fact that such projects can face real-world risks. The existence of the risks contradicts the concepts and organizational principles of the IT education. It, together with the trends in the academic research, leads to the overspecialization and fragmentation of student knowledge. It further tends to grow contempt of social and generally soft knowledge at the students' side. The study plans must cope with the fact that some of our projects may fail.

V. Basic Requirements on the Real-World-Like Projects

The analysis of the problems of our alumni is strongly restricted by data privacy protection. The data collection requires explicit approval of the students with the data collection and use. Even in the cases when alumni are successfully contacted, that their approval will not cause selective/choice effects. Based on the facts and our experience with students projects there can be made the following conclusions:

The students overestimate the importance coding skills. They are partially right as they are in such skills peerless. One perfect coder (people from the top 5% quantile) can replace some 20–30 average coders ([16], [17]). It leads to some conceit and underestimation of soft and social skills, to unacceptable team attitudes and to some negative aspects of hacking – see the Corncob Antipattern [18].

It is more important that such individuals strengthen the feeling of exclusivity and underestimate the necessity of good relations with users and ability to understand their knowledge and needs. It is not easy to communicate with them what is strengthened by the fact that they underestimate knowledge outside IT. Persuasion that it is neither good nor simple is usually a hopeless task. The only way is to create situation showing explicitly that they are not right.

Let us note that social and organizational skills are getting obsolete significantly slower than the knowledge and skills necessary for coding where within 5 years about one half of them get obsolete. Partial solution of this issue is possible if project aims are properly selected. We have tried two main variants of the projects:

1) Projects having features typical for SME/SMB projects. They train especially the skills of vision setting and requirement specification.
2) Projects solving partial tasks in software development in and for large enterprises. Such projects improve skills necessary for cooperation in such environments and enable understanding the environment and its processes.

VI. Involvement of students in commercial projects of large software vendors

Large vendors, e.g. IBM, have started a close collaboration with some Czech universities. The collaboration is implemented as the engagement of selected students in development teams of the vendors. The students often take part in the development of information system of the commercial partners as assistant analysts.

The students successively take up relevant roles in analysis and negotiation with a vendor client. He/she is induced to use vendor policies, processes, and tools. The students often take also part in design, especially in model building. In the cases when the students prove themselves useful they can be rewarded and have quite often opportunity to become gradually the employees of the firm.

The firms hire in some sense the students. It follows that it is applicable for some students only. The results and experiences with the concept are surprisingly good. The students are often warmly accepted by the enterprise team members. They are pleased to collaborate with the students.

The development of excellence centers supported e.g. by Europe grants enables further enhancement and a broader use of this concept. It simplifies the involvement of start-up firms and open new possibilities for master and dissertation projects.

The main advantage of these forms of studies is the training of analytical and design skills in the real world environment (ecosystem). The main disadvantage is that the students do not come in contact with the formulation of visions (aims) and marketing.

VII. Small to medium business oriented projects

Small-to-medium enterprises (SME) are a frequent domain where our alumni get good jobs. Most of our alumni get job in SME. SME need people having skills enabling to take part in all phases of software life cycle starting with market analysis, looking for a client, vision statement, negotiation, business agreements, and requirements specification.

These skills must be trained. We have trained the necessary skills in projects designed and implemented in the following way:

1) The projects are implemented as team seminary works being an optional part of a postgradual study. There are usually about 20 seminary participants.
2) The participants are required to propose themes of possible projects. The themes should have the potential to be commercially successful. The student must therefore, using their contacts market analysis formulate project proposal containing project name, the specification of possible user(s), and a (abstract of) vision. It must be based on the analysis of a real-word organization (partner).
3) The proposals are presented and defended.
4) The most promising proposals are chosen and the participants form teams having three to seven members.
5) Each team choose tools (e.g. an information system) supporting its work. An open source/free project support systems are preferred. The team members present their professional profiles, CVs and experiences.
6) The project name and abbreviated name is fixed. A project logo is welcome. The visions are refined and included into project home page. Achieved results are presented and defended using data projectors. The presentation should include the testable project effects.
7) The teams develop requirements specifications and system models. The involvement of the partner experts or intended end users is welcome.
   a) The teams can use open source software or free systems.
   b) The integration of the system into the partner ERP is highly appreciated.
   c) The team must do market analysis, especially the detection and study od the properties of similar systems.
8) There are at least two progress reports for each project. They have the form of reviews in the software engineering sense.
9) The final project defense at the end of seminary consists of:
   a) Presentation of the proposed system capabilities in the form understandable for users.
   b) Review of the technical properties a concise description of development processes, prototypes, and models.
   c) Interesting experiences.

The projects cover the software development stages at least up to design. There should be a well-founded hope that the project development could be continued and implemented on commerce basis (payment inclusive).

VIII. Schedule of the Seminary

A. Initial Meeting

- The teacher specifies the aims (vision) of the seminar. He/she shows that the main challenge of software project, especially of the project at SME, is the poor formulation of aims/visions and requirement specification of software project. The challenges are strengthened by the fact that the aims and requirements specifications are seemingly easy to understand and therefore easy. It is not the case but to understand the problem the students must take part in real-life projects needed collaboration with people from practice (future system users).
- A seminary support system specified (offered) rules of seminary students communication are specified in cooperation with the lecturer and the student.
- The overall schedule is specified see the points below.

B. Looking for Project Topics and Possible Business Partners

- Looking for Project Topics and Possible Business Partners
- Parallel activities:
  - Students present themselves, their experience, knowledge, and skills. Aims:
  - training the skills applicable during job seeking and for taking part in seminary project

C. Presentations and Competitive Choice of Project Proposals

- Choice of project leaders. The leaders are usually the students bringing the successful project proposals. They usually serve as project contact people. Involvement of the teacher in the choice is possible but not preferred.
- The leaders choice team members.
- The roles in the teams are tentatively defined.

D. Aims and Main Specifications

The aims and main specifications should be formulated in cooperation with business people – tentative users of the system.

E. Organizational Data

- Project identification;
- Project topic;
- Team structure and team roles – minimal team size is 3, maximal 7 members;
- Project supporting system (open source);
- The chosen development tools;
- Semiformal risk analysis.

F. Initial Outputs of the System Development

- the system full and short names;
- the specification of business partner;
- tentative vision.

G. First Presentation of the Project

First presentation of the projects (after a month) is in the form of a review:

- Presentation of project homepage.
- Main outputs.
- Control of the contents and documents in project support system, especially the team meeting reports, reports from negotiation with partners, and other project reports.

H. Closing Customer-Oriented Presentation

- presentation of project home page;
- presentation of goals/visions and final capabilities;
- possible use of prototypes.

I. Closing Technical Presentation

- Project schedule;
- Derived technical models (diagrams);
- Experience, positive and negative aspects.

J. Final Session

- Evaluation of the projects and students activities;
- Informal evaluation of individual students;
- Common personal $+$ and $-$.

The structure should be accommodated to the expertise of the students. Sometimes happens that there are students being professional software development team or software company leaders. In such cases the other students can take significant advantage of it.

IX. Issues Common for all Team Student Projects

Some students' projects can be done individually but in a team it is possible to collect some additional experience.

A. The Team

The students should build the teams independently. The students should fulfill their other study duties in parallel to the project. The team therefore faces the risk that some team members could be busy with other study duties or that their study can be for other reasons terminated. It can lead to one of the following situations:

- The students can close the ranks and start to help each other also in other (not project related) duties to keep the team at full strength.
- The students start to protect themselves by decomposing the project into relatively autonomous parts (or even by choosing projects that are naturally divided in this way) implementable by individuals. It allows them to show that they fulfilled their partial duties. Sometimes the finished components could provide at least basic functionality of the desired system.

We succeeded in some projects to set up the policy that every team member got assigned some development labor (preferably several tasks) in both primary and secondary role in pair programming. The role of a buddy (secondary developer in the pair) means that the person is informed what the primary developer does and why it does and that it could be asked to consult some issues or even to help. Such teams were quite reliable even when a team member was for some (even longer) time out of service (due other school duties or an illness).

A creation of such pairs requires some minimal tolerance and mental compatibility of the involved people. It is moreover advantageous, is a person has different shadows for his/her different tasks or if it plays a shadow of someone else than he/she is shadowed. If someone get into troubles, the shadow could take his/her work over and can delegate some of the shadow's work to his/her own shadows. The load for individual people than could grow by a part only what gives better chance that the person will be able to finish the work without induced troubles.

It has also approved itself if the project has been designed to change its extent (it is, there was a kernel that must be finished and multiple extensions that were optional – developed when nothing bad happened). Such projects could be finished successfully, reliably, and without fatal rushing.

Creation of a good team and tuning the extent of the project is a nontrivial task. It shows that it could take even several weaks yet before the project officially starts.

Side Effects

Working on a project may bring further benefits for the students:

- a deeper familiarization with the topic handled by the project;
- mastering tools for team work;
- act and negotiate with people (it is necessary to get on with other team members as potential team break could usually harm all its members);
- often also the discovery that a teacher can be true even if the student is thinking that he/she know it better (and then it shows that the hint given by the teacher and rejected by the student(s) could save a lot of work and avoid a lot of troubles).

B. Experience with the Projects

We practiced "real-life" student projects for eleven years at two Czech universities in two different cities (Prague and Brno). There were 4–6 projects a year at each university. There were interesting trends. The students in Prague tended to prefer coding. It caused the falling interest of the students in taking part in the projects. There are no active projects of this type in Prague now. There are less visible reasons for it. For example, a limited number of small enterprises developing information systems in Prague and a strong preference of theoretical computer science. Last but not least, it is possible to finish the study in other study branches with less effort and risk.

The trend in Brno has been towards a broad use of free or open software systems (project management tools, modeling tools, documentation tools) used to support the vision, requirements specification, and the team organization. It substantially enhanced the quality of the project processes. There were no project failures provided that their initial defences were successful. The reason was that the project vision must be successfully defended at the beginning of the project. An unsuccessful defense implied an immediate project cancellation. It was rather an exception.

The best projects in Brno were further positively influenced by the following aspects:

- There are many medium-sized firms in Brno employing the students.

- Some students have their own firms.
- There is a well-working collaboration of big firms (e.g. IBM) with Brno universities.
- SOA and Scrum methodologies can provide enough small-to-medium real-life projects.
- Some projects (usually 1–2 a year) were not only accepted and partly paid by the software firms but they were enhanced, commercially finished, and used for several years.

C. Students Projects and Curricula

The students' projects introduce many issues. They are of organizational, financial, and juridical nature. Let us mention some of them:

- Inclusion of project supervision into teachers load (good project supervision takes significantly more time than what corresponds to assigned/scheduled hours).
- Taking time complexity of the project into curricula design (to avoid time coexistence of critical parts of the project with other crucial study duties).
- Maintenance of successful projects. Good projects should be continued after the project evaluation. The issue here is that the evaluation is long term and laborious and the students that developed the application or system have also other study duties. It can also happen that they leave the university. Such projects could be used to train maintenance. There is a risk that an improper maintenance could break the well-working system.
- Turning the project into its commercial phase (in the case of its success) – requires having a procedure making from study result a commercial project (it includes transfer of the rights between the university and the company that will care about the project further). Some universities have this procedure stable and simple, for other universities it is a very hard (if not impossible) task.
- Not every teacher/lecturer is able to properly supervise students projects. It is necessary to find them and to prepare them. It requires some skills and experience that not every professor must have.

X. Conclusions

Majority of our student projects were quite successful. Their participants were as a rule able to find viable topics, contact people from firms, form student teams, find and apply team work supporting software, develop needed diagrams, and present the results. Some projects led to successful commercial products. It was probably partly due to the fact, that almost all the students were part-time employees of software firms. As a valuable byproduct the students discussed their experiences from the firms where they were employed.

The main contribution of the seminaries is a better understanding of the importance of a proper combination of soft and hard knowledge by all students, not only by the ones taking part at the seminaries. The information is spred spontaneously via social networks It is appreciated ex post by our alumni being in practice for several years.

Our seminary model has been successfully applied at Faculty of Informatics, Masaryk University Brno and partly at Faculty Mathematics and Physics, Charles University Prague.

We can conclude that the seminaries described above were successful. The weak point is that their concept is difficult to be applied massively. An implicit precondition of project success is that some of the seminary participants are quite excellent programmers and that some of the students have moreover a broader knowledge and social skills. It follows that they were good in technical abilities and STEM (science, technology, engineering, and mathematics) knowledge. It is a challenge as the STEM education becomes less popular and often not properly taught. It moreover can negatively influence soft knowledge needed in the project as it must take into account some aspects of STEM knowledge. The deteriation of STEM education is a threat for the quality of coding and for the education of coders. They moreover tend not to work together with users and to use user-oriented knowledge domain.

XI. Future research

The present curricula induce the fragmentation of education processes. It is then very difficult to organize long lasting student projects as well as scientific projects. There is a stronger challenge. Current principles of evaluation of universities and their professors based on impact factors prefer very strong specialization of research, narrow knowledge areas, and short term projects. It handicaps multidomain knowledge and skills needed in system analysis and requirements specification.

It is opened how to combine the training of technical skills needed for SME with the training of the skills for large software vendors. It main issue is the difference in enterprise culture, resources, and needs. It follows tthat many aspects of our projects must still be tuned. The idea is, however, crucial as otherwise there were be lack of good project managers having economic, social, as well as technical knowledge and skills. There is a danger that otherwise our alumni will become laborers (line workers) at Scrum-based duplicate production.

We intend to develop methods for development of quite complex systems using multiple student projects. We believe that it is possible if a specific SOA architecture [11], [19] is used. We intent to use open source and free software and combine it with commercial commodity software offered by large software vendors (e.g. Microsoft Excel, Microsoft PowerPoint, or OpenOffice Calc). We will attempt to apply the experience of small Czech software firms here.

References

[1] Standish Group, "The chaos report," 1994, [Online:] http://www.ics-support.com/download/StandishGroup_CHAOSReport.pdf; accessed 2014-02-28.

[2] ——, "Chaos: A recipe for success," 1999, [Online:] https://www4.informatik.tu-muenchen.de/lehre/vorlesungen/vse/WS2004/1999_Standish_Chaos.pdf; accessed 2014-02-28.

[3] ——, "Chaos manifesto 2013: Thing big, act small," 2013, [Online:] http://versionone.com/assets/img/files/ChaosManifesto2013.pdf; accessed 2014-02-28. [Online]. Available: http://versionone.com/assets/img/files/ChaosManifesto2013.pdf

[4] Calleam Consulting, "Why projects fail," 2014, [Online:] http://calleam.com/WTPF/?page_id=1445; accessed 2014-02-28. [Online]. Available: http://calleam.com/WTPF/?page_id=1445
[5] C. Ebert, "Software product management," *Software, IEEE*, vol. 31, no. 3, pp. 21–24, May 2014, DOI: 10.1109/MS.2014.72.
[6] W. W. Royce, "Managing the development of large software systems," in *IEEE WESCON Proceedings*. Institute of Electrical and Electronics Engineers, Aug. 1970, pp. 328–338.
[7] M. Cohn, *Succeeding With Agile: Software Development Using Scrum*. Addison-Wesley Professional, 2009.
[8] K. Beck, *Extreme Programming Explained: Embrace Change*. Boston: Addison Wesley, 1999.
[9] C. M. MacKenzie, K. Laskey, F. McCabe, P. F. Brown, and R. Metz, "Reference model for service-oriented architecture 1.0, OASIS standard, 12 October 2006," 2006. [Online]. Available: http://docs.oasis-open.org/soa-rm/v1.0/
[10] T. Erl, *Service-Oriented Architecture: Concepts, Technology, and Design*. Prentice Hall PTR, 2005.
[11] J. Král and M. Žemlička, "Implementation of business processes in service-oriented systems," in *2005 IEEE International Conference on Services Computing (SCC 2005)*, vol. 2. IEEE Computer Society, 2005, pp. 115–122, DOI: 10.1109/SCC.2005.58.
[12] J. Martin, *An Information Systems Manifesto*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1984.
[13] M. Levinson, "Recession causes rising IT project failure rates," *CIO Magazine*, Jun. 2009. [Online]. Available: http://www.cio.com/article/495306/Recession_Causes_Rising_IT_Project_Failure_Rates_
[14] P. Armour, "The reorg cycle," *Communications of the ACM*, vol. 46, pp. 19–22, Feb. 2003, DOI: 10.1145/606272.606288.
[15] I. Sommerville, *Software Engineering*, 9th ed. Pearson Education, Apr. 2010.
[16] G. M. Weinberg, *The Psychology of Computer Programming*. New York: Van Nostrand, 1971.
[17] B. W. Boehm, "Software engineering economics," 1981.
[18] W. J. Brown, R. C. Malveau, H. W. S. McCormick, III, and T. J. Mowbray, *AntiPatterns: Refactoring Software, Architectures, and Projects in Crisis*. New York: John Wiley & Sons, 1998.
[19] Open Group, "Open Group standard SOA reference architecture," Nov. 2011. [Online]. Available: https://www2.opengroup.org/ogsys/jsp/publications/PublicationDetails.jsp?publicationid=12490

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 835–840
DOI: 10.15439/2014F259
ACSIS, Vol. 2

Strategies for the Individualization of an Informatics Course

Olga Mironova, Irina Amitan,
Jelena Vendelin, Merike Saar
Faculty of Information Technology, Department of Informatics, Chair of Software Engineering, Tallinn University of Technology, Akadeemia tee St. 15A, Tallinn 12618, Estonia
Email: {olga.mironova, irina.amitan, jelena.vendelin, merike.saar}@ttu.ee}

Tiia Rüütmann
Faculty of Social Sciences, Department of Industrial Psychology, Estonian Centre for Engineering Pedagogy, Tallinn University of Technology, Akadeemia tee St. 3, Tallinn 12618, Estonia
Email: tiia.ruutmann@ttu.ee

Abstract—The present paper describes the strategies used to compile and teach an Informatics course developed during last years at Tallinn University of Technology. The strategy is based on the main principles of blended learning and the analysis of the results of experiments with students from different faculties. Various tests were carried out to identify students' levels of knowledge and preferences in their learning process based on their learning styles. Throughout the experiment, students were divided into groups according to the test outcomes. Separate groups were formed of students with different levels of knowledge and learning styles, determined using the Felder-Silverman model.

Adaptive learning tools were provided for the students considering the three main aspects: students' background, the level of their prior knowledge and their preferred learning style. The success of the strategy presented in this article is demonstrated by comparing the achievements of the test group with the reference group, who were not taught using the new strategies.

I. Introduction

E-learning is a rapidly developing world-wide system. Currently, it is not possible to imagine any educational process without e-components or a holistic e-learning system. The main aim of such systems is to provide knowledge in a convenient form for its consumer – the learner. Abundance of information does not guarantee perfect knowledge. Teaching materials should be carefully structured to cater for the needs and preferences of the students.

All present-day knowledge in engineering education is changing so fast that we cannot predict what the 21th century students will need to know tomorrow. Instead, we should be helping them to develop learning skills and strategies so that they will be able to learn whatever they need to. A combined set of knowledge, skills and attitudes is essential to strengthen productivity, entrepreneurship and excellence in an environment which is based on technologically complex and sustainable products, processes and systems. Similarly, we could improve the quality and nature of engineering education. Thus the objective of engineering education today is to educate students who are ready to engineer, and deeply knowledgeable about technical fundamentals.

Computer science is an integral part of the curriculum, which contents change fast. The general aim of this course is to develop logical, analytical and computational thinking by using the computer on the highest level.

Considering the target audience, several attempts were made to design the course material in the way that it would be easy to understand but would still achieve the goals. Nevertheless, the course seemed to be rather difficult for most of the students. It resulted in low examination grades and lack of motivation.

It became clear that more adaptive learning tools and taking into account individual properties of each student would motivate them and, as a result, would lead to better academic achievements. The question remained how to achieve as much individualization of teaching as possible, using the existing time and personnel resources.

II. A Brief Description Of The Course

The Informatics course belongs to the curriculum of the Institute of Informatics at Tallinn University of Technology. The aim of the course, designed for the first year non-IT students, is creation of applications by using standard PC equipment and developing object-oriented computational thinking. The learning process starts with processing information using Excel spreadsheets: formulas, diagrams, built-in functions and facilities. The set of practical assignments depends on the students' specialization: economics, social, chemistry and civil engineering.

Further, students learn the basics of programming in practice and the main principles of algorithmization. Python for technical disciplines and Visual Basic for Application (VBA) for humanitarians have been picked out as the programming languages for the second part of this course.

It should be noted that the programming part of the course was complicated for most of the students, especially for the humanitarians. This issue was solved by implementing Scratch in the course curriculum. This intuitive graphical

978-83-60810-58-3/$25.00

programming language helps students to take on board the main ideas such as brunching and cycle.

The Informatics course lasts for two semesters. During this period, we try to combine different styles of teaching and learning: classic face-to-face classroom methods, group work and learning in the Moodle e-environment [9]. The last one gives us a huge amount of different opportunities for individualizing the learning process, such as adjustment of the learning pace, for example, as well as increase and variety in the number of learning assignments. Furthermore, students get a diversity of ways to learn and possibilities for self-tests in the e-part of the course. During the study time, they can choose between different kinds of teaching materials and use what they prefer based on their knowledge and learning styles.

The course is taught in three languages: Estonian, English and Russian.

III. Research Methodology

A. Overview

It is generally known that it is not possible to provide all students with one-to-one tutoring at a university. However, this fact should not affect the main educational goal - to ensure high-quality competitive knowledge. In our experiments with course design and curriculum, we considered the differences in students' features, especially their learning styles and prior knowledge.

Since the 2010 fall semester, our group of lecturers has applied a new flexible and adaptive approach to designing computer science e-courses [8]. The basic idea of our methodology is to divide students into equal reference and test groups and to compare the results of these two groups. The reference group is taught using the same course materials and the same e-course but they were not helped with any additional system. The students of the test group, however, were directed in choosing their learning materials based on the data obtained through the tests.

The division into groups was random and was not linked to the students' specialization or knowledge. During the 4 years of the experiment, the group sizes varied depending on the number of students. An average number of participants in the experiment each year was about 100-150 students and they were not aware of the research.

To evaluate the students' progress we tested them at the beginning of the course and at the end of it. The test contents for both groups were similar and were based on the topics described in the European Computer Driving License (ECDL) [1]. The tasks focused on creating documents and presentations, processing spreadsheets, and elementary knowledge in programming.

The last category of the tasks was added recently. In this implementation, we proceeded from an elective course 'Basics of Application Development and Programming', which has been included in the curriculum of Estonian secondary schools.

The current situation of teaching computer science at schools is varied. Some schools do not have computer science lessons at all due to the lack of teachers. Unfortunately, most of those pupils who have obtained sufficient IT-knowledge at their schools are not non-IT students at university level, who our course is aimed at and designed.

B. Research stages

Theoretically, it is possible to name three main strategies in the process of improving the Informatics course and individualization of the learning process:

- e-course
- knowledge
- learning style.

The first stage, which was named 'e-course', includes the adaptation of the course materials for the e-environment. Theoretical materials and practical assignments were innovated and supported with videos [6] and self-tests. They made the Informatics course more attractive for students. Both groups, the test and the reference group got access to this renewed course. At the same time, face-to-face lessons were held, too. Here we preferred group work that gave the students an opportunity to try the obtained knowledge in practice and develop teamwork skills. In this case the role of the lecturer was slightly different – it became more of an advisor, motivator and supporter in the student's work with learning materials. During the contact lessons learners ask questions related to their homework and share their skills and experience with other. In addition, students have an opportunity to get the support not only from their teacher but from other students, too. This form of support is equally useful both for the students who get it and, especially, for the ones who give it. To find and correct a mistake is an important skill in computer science subjects.

It makes no sense now to enumerate all the advantages of e-learning – they have been known to all. At this stage, we got the first results of our work: positive feedback from students and increase in academic achievements. These results were extracted from the Studies Information System (ÕIS) – an e-environment, where students and teachers get information about courses and curricula, students declare courses, keep results and give anonymous feedback on their educational process [11].

The second stage, titled 'knowledge', is dedicated to the division of the test group students into three streams based on their subject knowledge at the beginning of the course. Those three groups receive different amounts of different level practical tasks in the Moodle e-environment (Fig. 1). So, we increased the number of practical tasks without increasing the subject hours.

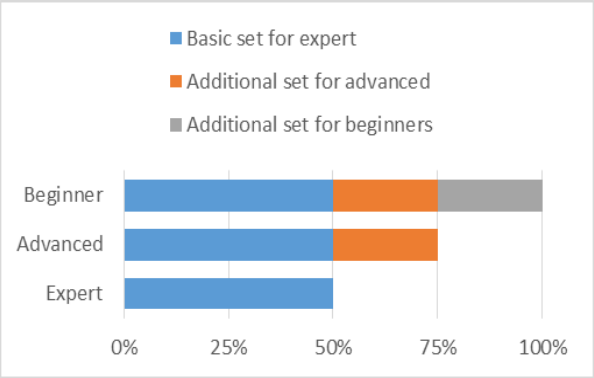


Fig. 1 The distribution of practical tasks

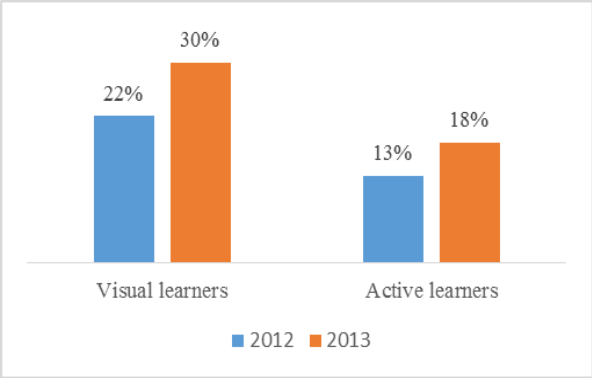


Fig. 2 Increase in the number of visual and active learners

The face-to-face lessons were held as before. The mentioned division provided us with important information about what the students knew before studying our course. Moreover, according to the data we were able to provide them with the necessary learning material.

The success of this stage of experiment was confirmed by positive results of the students' survey in the ÕIS and by the increase in academic results.

The 'learning style' phase of our experiment was the most laborious part. Learning styles are characteristic cognitive, affective, and psychological behaviours that serve as relatively stable indicators of how learners perceive, interact with, and respond to the learning environment. Students learn best when instruction and learning context match their learning style.

There are many studies about the individualization of learning depending on students' ability [7], [3]. Using one of these, the learning styles model of Felder-Silverman, we divided our test group students by their preferences and provided them with corresponding learning materials [3], [4].

Felder distinguishes the following groups of learners depending on their learning styles [2]:

- active and reflective
- sensing and intuitive
- visual and verbal
- sequential and global.

Through this division, we found that the majority of our students were active and visual learners and they had very strong preferences for their learning process. These preferences were detected according to the Felder test which was held at the beginning of the course [10]. Throughout the educational process, students were provided with the necessary learning materials and activities in accordance with Felder´s instructions [5]. For example, active learners received more group work and opportunities to help others; visual learners were provided with visual representation of the educational material.

It should be noted that each year the number of active and especially visual students increased (Fig.2).

The results of this stage of research showed us that the test group students managed with their practical tasks better than the students of the reference group. This has led to the better academic progress of the test group students.

Thus, we were able to create a model of our students´ learning preferences (Fig. 3), which considers their level of knowledge and preferences in the learning process. Using this model, we try to find an individual approach to each student in our course.

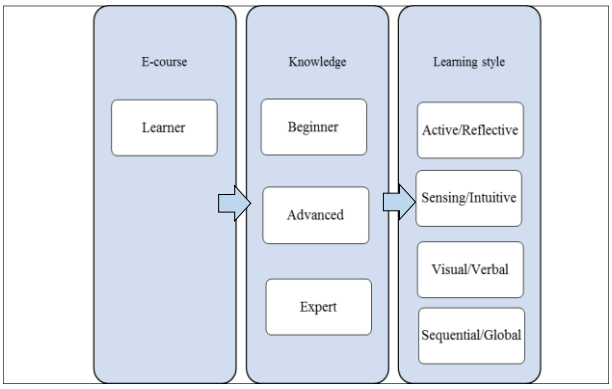


Fig. 3 The model of students' learning preferences

C. The Method of Research

Throughout the experiment, positive students feedback and good exam results showed the positive effect of the course and curriculum modifications. Finally, it was decided to examine the data with statistical methods. The aim of this examination is to check and prove the correctness of the chosen approach to an educational process.

As the method of the hypothesis testing, we chose the Student's t-test for the comparison of the two means. This test assumes a normal distribution of samples and not significant differences between the standard deviations of either samples. Our aim is to show that there were no significant differences between the test and reference group students in September, while in January the results are significantly different.

For calculations we use the equations for the averages $\bar{x}$ (1) and corresponding standard deviation S for both groups (2):

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (1)$$

$$S = \sqrt{\frac{\sum_{i=1}^{n} \left(\bar{x} - x_i\right)^2}{n-1}} \quad (2)$$

After that, we calculated the standard error σ using the equation 3.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} \left(\bar{x}_{Test} - x_{iTest}\right)^2 + \sum_{i=1}^{n} \left(\bar{x}_{Ref} - x_{iRef}\right)^2}{n \cdot (n-1)}} \quad (3)$$

Finally, using equation 4 we calculated the experimental value t_{exp}:

$$t_{exp} = \frac{\left|\bar{x}_{Test} - \bar{x}_{Ref}\right|}{\sigma} \quad (4)$$

In addition, to compare this value with theory we need to calculate the degree of freedom df using equation 5:

$$df = 2n - 2 \quad (5)$$

As initial data for calculations, we chose September 2013 students' test results of the test and reference groups, and the same groups' results in January 2014 (cf. Table I in the Appendix).

Both samples are in the equal size: n=89 and distributed normally (Fig. 4 and Fig. 5).

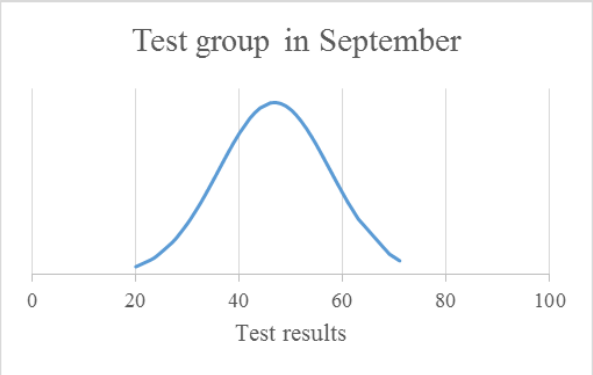


Fig. 4 The distribution of the test group results

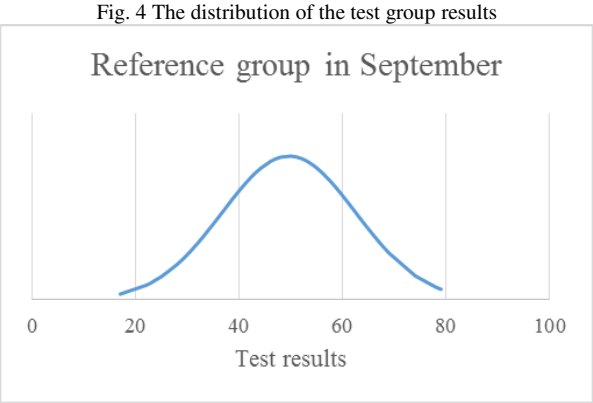


Fig. 5 The distribution of the reference group results

The two means and the corresponding standard deviations are calculated by using the equations 1 and 2:

$$\bar{x}_{Test} = \frac{4164}{89} \approx 46{,}787$$

$$\bar{x}_{Reference} = \frac{4413}{89} \approx 49{,}584$$

$$S_{Test} = \sqrt{\frac{10222{,}94}{89-1}} \approx 10{,}778$$

$$S_{Reference} = \sqrt{\frac{14721{,}62}{89-1}} \approx 12{,}934$$

Now we see that there is no significant difference between the standard deviations in either groups. It means that we can continue with Student tests.

The standard error of the difference between the two means is calculated by using equation 3:

$$\sigma = \sqrt{\frac{10222{,}94 + 14721{,}62}{89 \cdot (89-1)}} \approx 1{,}785$$

Experimental t value is calculated using equation 4:

$$t_{expSept} = \frac{\left|46{,}787 - 49{,}584\right|}{1{,}785} \approx 1{,}567$$

To compare this value with the theoretical t_{th} we need to calculate the degree of freedom using equation 5:

$$df = 2 \cdot 89 - 2 = 176$$

Using equations 1 to 4 we then calculated both groups´ results in January 2014:

$$\bar{x}_{Test} = \frac{8029}{89} \approx 90{,}213$$

$$\bar{x}_{Reference} = \frac{6896}{89} \approx 77{,}443$$

$$S_{Test} = \sqrt{\frac{5092{,}94}{89-1}} \approx 7{,}608$$

$$S_{Reference} = \sqrt{\frac{12518{,}22}{89-1}} \approx 11{,}927$$

$$\sigma = \sqrt{\frac{5092{,}94 + 12518{,}22}{89 \cdot (89-1)}} \approx 1{,}499$$

$$t_{expJan} = \frac{\left|90{,}213 - 77{,}443\right|}{1{,}499} \approx 8{,}489$$

D. Results of the experiment

Using the table of theoretical t_{th} values with the corresponding degree of freedom (cf. Table II in the Appendix) we found that the means of September results are not different at any critical level:

$$t_{expSept} < t_{th}$$

$$1{,}567 < 1{,}65$$

This means that at the beginning of the course both groups of students, the test and reference, had the same level of knowledge.

January results are the opposite – the means are different at critical levels.

$$t_{expJan} > t_{th}$$

$$8{,}489 > 3{,}29$$

It shows that the students who were taught using our system of the learning process individualization obtained knowledge much better then the others.

The progress of both groups is shown in the Figure 6 and Figure 7:

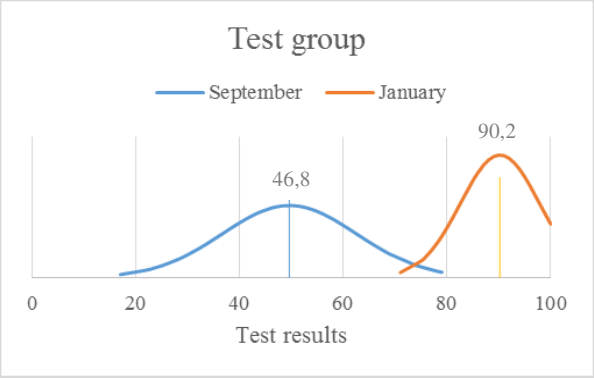


Fig. 6 The test group progress

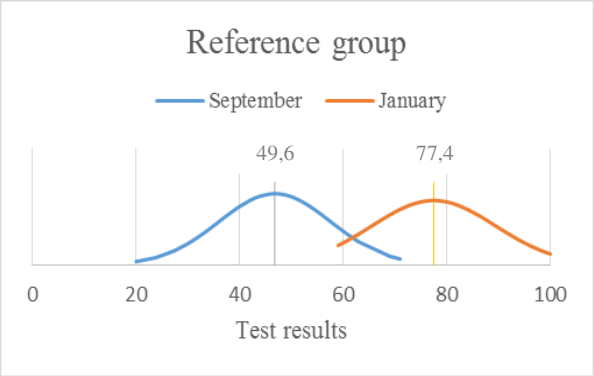


Fig. 7 The reference group progress

These results confirm the validity of our study and the chosen method of course individualization.

IV. Conclusions

The analysis presented in the paper shows positive outcomes of the strategy used. The calculations with two relevant groups, the test and reference, demonstrate significant differences in achievements at the end of the first part of the course.

The feedback, received from the groups, is also different. The test group shows higher motivation for further learning compared to the reference group. The main reason for it, picked out by students, was that there were no unreachable targets in the educational process.

Classroom activities of teachers and students took place in mutual communication. Therefore, the guidance and the formative role of the teacher was realized in the creation and review of the theoretical material and the material in practical classes.

The authors intend to continue developing the created model and Informatics course in the same style, trying to adapt it to individual students as much as possible.

Appendix

The results of each group are arranged in two columns under the name of the group in the Table I.

Table II shows only a part of the table of theoretical t values for Student's test. The whole table could be found in any book on statistical analysis.

TABLE I.
THE TEST RESULTS OF BOTH GROUPS

September 2013				**January 2014**			
Groups' results				**Groups' results**			
Test		**Reference**		**Test**		**Reference**	
63	52	61	39	89	91	79	72
52	40	54	59	81	98	84	100
71	38	78	48	75	91	78	71
63	40	59	46	91	79	80	62
53	56	55	40	100	96	77	74
69	58	58	44	79	98	81	68
61	48	69	65	71	78	64	60
69	54	74	57	83	98	60	72
61	55	60	47	100	100	67	87
63	40	56	50	89	81	71	86
61	51	56	64	88	92	84	100
55	42	61	48	79	98	60	84
42	39	51	42	89	80	100	78
59	44	59	46	87	90	79	82
63	48	53	48	100	90	73	84
54	55	60	54	89	98	100	92
48	47	44	46	77	90	66	84
49	40	55	54	88	93	66	92
55	32	63	44	85	95	78	82
44	54	54	17	90	98	96	76
52	47	61	55	88	86	78	94
57	44	79	51	100	87	66	60
55	46	74	50	81	82	72	59
46	44	53	25	77	91	72	65
59	34	65	23	90	98	84	63
55	35	59	30	86	82	78	70
48	28	36	33	79	98	67	74
46	40	43	59	87	89	75	100
53	37	59	55	98	98	91	98
44	43	55	25	76	96	99	69
47	43	46	52	90	90	69	96
52	30	69	34	100	98	62	78
56	36	68	36	80	94	78	82
48	44	55	50	92	92	79	96
44	34	46	41	80	92	70	68
44	27	47	38	81	100	92	66
55	42	40	50	98	94	65	100
51	33	22	46	98	98	76	98
50	28	67	44	100	92	76	96
49	40	51	42	86	100	60	66
35	28	44	23	93	100	73	68
46	23	57	28	80	94	64	74
48	20	33	32	86	92	64	72
48	24	44	42	98	98	76	79
38		38		100		70	

TABLE II.
THEORETICAL *t* VALUES

Degrees of freedom	Probability			
	0,1	0,05	0,01	0,001
∞	1,65	1,96	2,58	3,29

ACKNOWLEDGMENT

The authors thank anonymous FedCSIS 2014 reviewers for their comments and suggestions. In addition, the authors thank the professor emeritus Leo Võhandu, the Senior Research Scientist of Tallinn University of Technology, for his support and contribution to the present research.

REFERENCES

[1] ECDL Foundation. Retrieved from: http://www.ecdl.com
[2] Felder, R.M. and Brent, R. Understanding Students Differences, Journal of Engineering Education, 2005, 94(1), pp. 57-72. http://dx.doi.org/10.1002/j.2168-9830.2005.tb00829.x
[3] Felder, R.M. and Silverman L.K. Learning and Teaching Styles in Engineering Education, Engr. Education, 1988, 78(7).
[4] Felder, R.M and Soloman, B.A (n. d.) Learning styles and strategies. Retrieved from: http://www4.ncsu.edu/unity/lockers/users/f/felder/public/ILSdir/styles.htm
[5] Felder, R.M. and Spurlin, J. Applications, Reliability, and Validity of the Index of Learning Styles. Intl. Journal of Engineering Education, 2005, 21(1), pp. 103-112. Retrieved from: http://www4.ncsu.edu/unity/lockers/users/f/felder/public/ILSdir/ILS_Validation%28IJEE%29.pdf
[6] Khan Academy. Retrieved from: http://www.khanacademy.org
[7] Kolb David. "Experiental Learning", Engle Cliffs, Prentice Hall, 1984, 256 p. Retrieved from: http://academic.regis.edu/ed205/kolb.pdf
[8] Mironova, O., Amitan, I., and Vilipõld, J. Computational Thinking and Flexible Learning: Experience of Tallinn University of Technology. Lecture Notes in Information Technology; 2012, 23-24, pp. 183 – 188. Retrieved from: http://www.ier-institute.org/2070-1918/lnit23/v23/183.pdf
[9] Moodle. Retrieved from: https://moodle.org
[10] Soloman, B. A. and Felder, R.M. (n. d.). Index of Learning Styles Questionnaire. Retrieved from: http://www.engr.ncsu.edu/learningstyles/ilsweb.html
[11] Studies Information System, Tallinn University of Technology. Retrieved from: http://ois.ttu.ee

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 841–850
DOI: 10.15439/2014F196
ACSIS, Vol. 2

Situational Software Engineering
Complex Adaptive Responses of Software Development Teams

AJB (Barry) Myburgh
Jo'burg Centre for Software Engineering (JCSE)
School of Electrical and Information Engineering
University of the Witwatersrand
Johannesburg, South Africa
barrym@jcse.org.za

***Abstract*— The Complex Adaptive Situational Model (CASM) promotes understanding of establishing conditions which enable software engineering success. Influenced by complexity science, CASM explains aspects of the state of dynamic equilibrium that is achieved under constraining influence of management and production governance. Four states of dynamic equilibrium are defined: Crafted Quality (Agile), Controlled Quality (waterfall), Managed Costs (WetAgile) and Self-Directed Quality. A band of software engineering feasibility is also described and it is suggested that successful software engineering initiatives require teams to operate in that band. The journey across the band of feasibility is explained by introducing SEMAT, with Crafted Quality amounting to applying SEMAT Essence, and Controlled Quality being achieved by introducing additional practices which satisfy the more stringent governance requirements. An enterprise is then described as a collection of CAS's, thereby setting the scene for further research into the complexities of human-driven complex adaptive systems.**

I. Introduction

FOR many practitioners, Agile software development seems the best way to develop software. But old-style management often presents the biggest obstacle to successful adoption of agile approaches. The model described in this paper promotes understanding of what it takes to establish conditions which enable software engineering success, not only with agile approaches, but also traditional, plan-driven software engineering.

Humans easily relate to causal determinism – a thesis based on experience that future events (combined with the laws of nature) are sometimes the result of past and present events. For example, causal determination enables us to catch a ball by predicting in which direction it is going. Causality also enables software developers to design, plan, and predict what software will do.

Immanuel Kant (1724-1804) promoted universal causal determinism. But causality is not enough. We can't accurately predict the weather. We can't predict the full combination of features, qualities, time and resources of a software project. As explained by Jurgen Appelo: *Predictability has a devious sister called complexity* [1].

II. Complexity

Our attempts at understanding complexity involve: Dynamical systems theory; Chaos theory; Network theory; Game theory and other branches of science that are collectively known as the Complexity Sciences. Causality ruled the sciences from the 17th century. Complexity is a product of the 20th century. Complexity theory offers a new way of understanding the problem of producing software and managing organizations - even though our minds prefer causality over complexity.

The human brain is wired to find purpose and causality in everything and we favor "linear thinking" to "nonlinear thinking". So we easily reason that the global financial crisis was caused by bankers. Bad atmosphere at work is caused by the manager. The team didn't make a deadline because of someone's mistake.

The mental addiction to causal determinism has led people to use control to ensure desired outcomes. Engineers and other people with technical minds are particularly susceptible to the concept of control. Engineers developed scientific management - the command-and-control style of management. Engineers devised the kind of control systems we still find today which work adequately with repetitive tasks that don't require serious thought and analysis. But these control systems don't work with creative product development.

Managers also look for causes that would produce the outcomes exactly as they need them: through careful up-front design, with meticulous top-down planning. Appelo explains that agile management derives when hierarchical management embraces complexity and non-linear thinking and is a logical companion to Agile software development [2].

III. Challenges of Software Engineering

The software development industry started in an ad hoc way with the term "software engineering" first appearing in the 1968 NATO Software Engineering Conference where

This work was not supported by any organization

attention was given to the perceived "software crisis" of the time which resulted from the impact of rapid increases in computer power and the complexity of the problems that could be tackled. In essence, it referred (and still refers) to the difficulty of writing correct, understandable, and verifiable computer programs. The roots of the software crisis have been recognized as being complexity, expectations, and change. All too often formal approaches introduced bureaucracy and delivered software much more slowly than the rate at which requirements were changing. At the same time, some teams of passionate and disciplined programmers, with ad hoc processes and flexible requirements, delivered products of higher quality at a fraction of the cost and in a fraction of the time.

The dilemma created by the constantly high rate of software project failure in the midst of a multitude of alternative ways of working, triggered the search for general theories of software engineering that could achieve recognition equivalent to that of, for example, Maxwell's equations in the electrical engineering community [3]. But where Maxwell's equations deal with translating natural phenomena into usable practice, software engineering is all about people applying process and technology to translate their ideas into operational solutions. This translation is enabled by design which, according to John Gero and quoted by Kruchten [4], is a goal-oriented, constrained, decision-making, exploration and learning activity which operates within a context that depends on the designer's perception of the context. In the same article Kruchten explains that he had to extend the boundary of "software design" to include much more than software practitioners' traditional activities as defined in the Software Engineering Body of Knowledge (SWEBOK). In SWEBOK, software design covers only a narrow set of processes and artifacts [5]. But if we accept that design is making choices that will shape the final product, we must include some requirements activities and all coding and testing activities. The significance of this statement is that, contrary to most other engineering disciplines, the design process remains active throughout – virtually up until the very moment that source software is translated into executable machine language. And people drive the design process. As is described later in this article, people are the active agents in a complex adaptive system (CAS) and CAS agents respond to governance forces while applying rules.

An early case study that deals with the tension between approaches is described in Dee Hock's fascinating book *"Birth of the Chaordic Age"* (1999). He describes how, in the 1960's, a management team responded to their concerns when the traditional approach to system delivery was failing. *The team took ownership of the challenge and we shut ourselves in a room and didn't come out until we had an approach to which we were totally committed*. He also confirms that: *out of initial failure grew a magnificent success* [6].

In 2001 a gathering held in Utah resulted in formulation of the "Agile Manifesto" [7] which was, on the one hand, a reaction against the bureaucracy of the formal approaches, while on the other hand, also taking a stand against the "chaotic" processes and low quality products of undisciplined programmers. It gave substance to the search for a middle road between structure and non-structure, between order and chaos.

Evidence demonstrates that Agile software development, when done well, shows a tremendous return on investment. But if Agile methods have such positive effects, why doesn't everyone use them? And why are so many software projects across the world still failing? Appelo refers to a "State of Agile Development Survey 2009" [8] which identified the following factors as contributing to failed Agile approaches:

- Management opposed to change
- Loss of management control
- Lack of engineering discipline
- Team opposed to change
- Quality of engineering talent
- Organizational need for planning, predictability and documentation

This seems to suggest that management preferences are the biggest obstacles to Agile software development. The CASM model described in this article sheds further light on this contention.

In 2009 a group of leading international software engineering personalities started collaborating on an initiative to "re-found" software engineering. Ivar Jacobson (Use Cases, UML, RUP), Bertrand Meyer (Design-by-Contract and the OO Language Eiffel) and Richard Soley (CEO of the Object Management Group (OMG)) established the SEMAT Initiative - Software Engineering Method and Theory. Supporters of the initiative signed a declaration somewhat reminiscent of the Agile Manifesto and since then a great deal of work has been carried out aimed at defining the "kernel of widely-agreed elements" and often referred to as "Essence", meaning the essence of software engineering.

Figure 1 highlights two important aspects of the SEMAT kernel:

- The Areas of Concern (Customer, Solution and Endeavor) and
- The Alphas (Opportunity, Stakeholder, Requirements, Software System, Team, Work and Way of Working)

Areas of concern are addressed in terms of Activity Spaces which involve the actions taken to achieve objectives.

Alphas represent essential aspects of software engineering and each progresses through a number of states (Alpha States) as the team conducts work.

As described in the submission to the OMG [9] and in the published book *"The Essence of Software Engineering: Applying the SEMAT Kernel"* (2013) [10], the SEMAT initiative promises to fundamentally affect the discipline of software engineering.

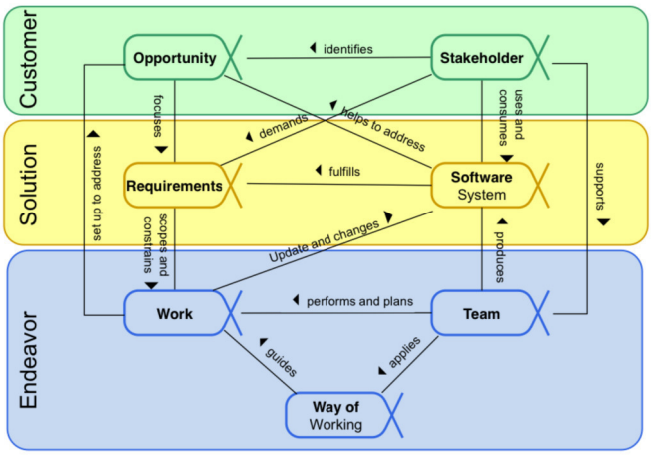


Figure 1: SEMAT Alphas in the Areas of Concern – Components of the Essence of Software Engineering

IV. Complex Adaptive Systems

A view shared by many software development experts and Agile/Lean evangelists is that software projects are complex adaptive systems (CAS's). CAS's are composed of agents - as described by M. Mitchell Waldrop in *"Complexity: the Emerging Science at the Edge of Order and Chaos"* (1992) [11]. CAS agents can be molecules, neurons, web servers, fish, starlings, and people - always forming new emergent structures with new emergent behaviors

Software projects involve people who are constantly organizing and reorganizing into larger structures: Project teams; Social groups; Task forces; Committees; etc.

CAS's are able to adapt to their environments: an infant learning to walk; car drivers evading a traffic jam; a software team adapting to what their customer really wants. Moving to the sweet spot between chaos and order, they learn and adapt and navigate their way with "chaordic" [12] processes that are neither fully ordered nor truly chaotic.

In Agile software development, we often hear reference to scientific terms such as self-organization and emergence. The concepts of emergence and the factors leading to emergent results lie at the heart of CAS theory's relevance to software development.

Examples of self-organizing systems include an ant colony, the brain, a Scrum team, a CMMI-Dev team, a team applying SEMAT Essence. Scrum, CMMI-Dev and Essence are not methodologies with defined processes or sets of procedures - they are development frameworks. And the frameworks provide rules and constraints on behavior that can cause a CAS to self-organize into an intelligent state of dynamic equilibrium.

When applying complex systems theory to software development and management, we are treating the organization as a system.

System dynamics - not to be confused with dynamical systems theory - is a technique from the 1950's to help managers understand and improve their industrial processes. System dynamics recognized that structure is often a more important contributor to an organization's behavior than individual parts themselves.

Systems Thinking was developed in the 1980's and popularized by Peter Senge's book *"The Fifth Discipline"* [13]. It's about understanding how things influence each other in the whole, a problem-solving mindset that views problems as parts of an overall system. In some ways similar to System Dynamics, but more subjective.

Social complexity is the study of complexity in social systems and to manage social complexity, we need to understand how things grow - not how they are built. This is an extension of ideas promoted by Fred Brooks in 1987 when he explained that the very essence of software engineering lies in complexity, conformity, changeability and invisibility [14].

Appelo's *"Management 3.0"* applies complexity thinking and assumes that managers cannot construct and steer a self-organizing team. The team must be grown and nurtured. Productive organizations are not managed with models and plans; they must emerge through the power of self-organization and evolution. Appelo suggests that complexity thinking is like the light that feeds all that grows and goes on to explain that at the project level, new emergent structures form and new emergent behaviors are displayed [15]. Like any other CAS with interconnected agents (people) interacting with each other to form an integrated whole. Even though software projects have many elements, only people are the real agents - the active elements. Teams themselves are agents on the next higher level.

Items that are not agents include: Requirements; Features; Artifacts; Deliverables; Tools; Technologies; Processes; Practices. They cannot actively organize and reorganize themselves. They cannot initiate interaction with any of the other elements in the project.

Appelo emphasizes that the primary focus of any manager should be to energize people - to make sure that they actually want to do what's required of them. Like a gardener looking after plants in a garden, a manager looks after the employees on the team/s [16].

For centuries mathematicians have preferred to work with linear (ordered) systems and considered nonlinear (complex) systems to be a special group. But nonlinear systems are the norm and abundant throughout the universe, whereas linear systems are a rare and special breed. From the beginning of the universe, everything in it was shaped by self-organization. Self-organization is the process where a structure or pattern appears in a system without any central authority or external element imposing it through planning. Self-organization is the norm. It is the default behavior of dynamic systems, whether these systems consist of atoms, molecules, species, businesses or software developers. Appelo emphasizes that self-organization is not a "best practice" - it is "default practice" [17]. No matter how a team

is managed, there will be self-organization. People will discuss and agree on lunch meetings, folder structures, workplace territories, birthday parties. Everything that management does not constrain - and much that it attempts to - will self-organize. Humans have behaved that way for 200 000 years.

But is what happens also happening in the "right direction"? Though every self-organizing system can have its own direction, the possible directions are limited by its environment. No self-organizing system exists without context. And the context constrains, governs and directs the organization of the system.

Environmental constraints affect the direction taken by a self-organizing system. This is illustrated by considering the Game of Life - a simple zero-player game invented in 1970 by the British mathematician John Conway. It is "played" on a grid of cells, where each cell has eight neighbors, one in each direction, including the diagonals. The cells can be born and stay alive or die as determined by the application of rules. The Game of Life is an example of a cellular automaton - a mathematical system in which cells are influenced by other cells according to some set of predefined rules. It is particularly interesting because it is a fine example of a system with a small set of simple rules, having complex behavior and ordering itself. The game also shows us that, whatever the initial situation is, the system will eventually always stabilize.

There is, however, one catch: the set of rules has to be chosen carefully. We therefore observe that rules must be tuned for a system to be both stabilizing and lively. A different set of rules leads to a different system with different behavior

As described by Waldrop [18] Stephen Wolfram proposed a classification scheme for cellular automata - named universality classes.

- Class I: These are the systems with “doomsday rules”. No matter what pattern of living and dead cells at the start, everything dies within a few generations.
- Class II: These systems are a bit livelier, but not much. Each initial pattern quickly collapses to a set of very boring, static configurations.
- Class III: These systems are at the opposite extreme: they are too lively. Each initial pattern in the system results in total chaos with no configuration stabilizing and nothing being predictable.
- Class IV: These are the systems with a set of rules not leading to dead, static, or chaotic configurations. Emerging patterns in this category are lively, creative, often surprising, but also stabilizing.

In dynamical systems, Classes I and II correspond to order. Class III corresponds to chaos. Class IV (of which the Game of Life is a famous example) corresponds to complexity. Given that complexity is usually explained as the region between order and chaos, this means that class IV finds itself between II and III.

Complex adaptive systems are systems that can find their own way toward that sweet spot of complexity, between order and chaos, where life blooms and creativity thrives. Scientists call it the edge of chaos, but they also could have called it the edge of order. This sweet spot represents a state of dynamic equilibrium between governance forces, parameters and rules that influence emergent behavior of the CAS.

Self-organization takes care of the edge of chaos when certain parameters fall within a critical range. The manager is not a game designer and is not concerned with the low-level rules of the game. Rather, the manager configures the high-level parameters, like diversity of team members, information flow between people, and connectivity between teams. When setting up governance in an organization, one responsibility of a manager is the development of a self-organizing system, defining the boundaries of the board but not the rules of the game. When a manager takes rule-making into own hands, self-organization will be significantly influenced and frustrated. And then creativity, innovation, and adaptability in the system will suffer.

Self-organization is fundamental for every complex system. But in a human social system, self-organization alone is not enough. Appelo explains how Glen Alleman described the need for management by pointing out that there is a difference between self-organizing and self-directing and this is the role of management [19]. This is not "directing" in the command and control sense. It is directing in the "required business value" sense. If self-organizing teams serve their customers, who "manages" the customer, when the customer is not prepared to behave in a "well-mannered" way? If there is more than one self-organizing team working on the same project, who coordinates the activities between these teams? When there are conflicts in resources, funding and requirements, who coordinates resolution of these conflicts? At least a little management is needed to steer self-organization in a direction that is of value to everyone in the system. Appelo points out that Sanjiv Augustine calls it "light-touch leadership". Appelo calls it alignment of constraints [20]. This author calls it balancing the governance forces.

Directed self-organization in software engineering is a matter of manipulating governance so that a group of people produces results valuable to the goals of the project.

V. The Complex Adaptive Situational Model

Humans, with the introduction of consciousness, invented morality, laws and authority. We defined preferred directions for self-organizing systems because some results are seen as valuable and others as harmful. We value human lives therefore consider malaria parasites and HIV viruses an undesirable result of self-organization. Appelo points out

that we value many irrational and unnatural things too, like non-discrimination, peace, monogamy [21]. Self-organization makes no distinction between good and bad, between virtues or vices, between valuable and harmful. Systems simply do whatever the environment allows them to do. Whatever they can get away with. And so, humans embraced the concept of command-and-control which enables attempts to steer self-organizing systems (businesses, teams, countries) in the direction that stakeholders considered to be valuable. That's how managers got their positions and how governments try to run countries. They care about results. They want to make sure that self-organizing systems either produce valuable things (products and services), or refrain from harming valuable things (human lives, economic growth, natural resources). Managers want software teams to create valuable software with which to make money or deliver good service.

Key constraints affecting the emergent behavior of a team of software engineers as a CAS are broadly identified by this author as:

- Management Governance and
- Production Governance.

Management Governance is a method or system of management practices that range from formal, high ceremony practices on the one hand, to informal, low ceremony practices on the other. The formal approach to management provides work products that could lead to high levels of visibility – producing project plan/s and progress reports, risk management plan/s and reports, quality management plan/s and reports, configuration management plan/s and status accounting reports, meeting agendas and minutes, etc. On the other hand, the informal, low ceremony approach depends less on detailed, written communication, hence leaving less visible evidence trails.

Production Governance is a method or system of production practices that range from engineering, "Waterfall" practices on the one hand, to organic, iterative or "Agile" practices on the other. A key engineering practice is to work according to the sequential stages of the life cycle, e.g.: Requirements Analysis; Design; Implementation and Unit Test; Integration and System Test; Qualification Test. Visible artifacts are then produced, including software requirement specifications, software architecture documents, software design documents, programming standard/s and test records. At the organic extreme we experience a situation where there is little emphasis on the life cycle stages and associated documentation, and high focus on the technical practices of software development.

As illustrated in Fig. 2, this author's hypothesis is that different combinations of governance constraints influence emergent behavior, resulting in four possible states of dynamic equilibrium:

- Crafted Quality (Agile)
- Controlled Quality (Waterfall or Plan-Driven)
- Managed Costs (WetAgile)
- Self-Directed Quality

While CASM identifies these four domains, it is important to realize that for a particular team and at a particular time, only one of the domains will be dynamically active to represent the emergent behavior of that team under particular circumstances.

Today's "Complex Adaptive Situational Model" (CASM) illustrated in Fig. 2 was first described as the model of "Situational Software Engineering" by Myburgh in 1992 [22]. Continuous application and research gave rise in 2005 to the second generation of the model, viz. the "Situational Process Model" (SPM), illustrating interaction between production processes and management & control processes [23]. CASM represents the third, published generation of the model and it identifies four behavioral domains that represent states of dynamic equilibrium as responses to the environmental governance constraints.

The terms "Plan-Driven" and "Agile" have been used by Boehm and Turner [24] to describe what are essentially the Controlled Quality and Crafted Quality approaches.

CASM in no way implies that the essence of software engineering is any different in Controlled Quality and Crafted Quality domains, but life cycle models will be different as later explained in this article.

A. Crafted Quality (The Curved Arrow)

This domain suits the information age organization where management formality is relaxed (low ceremony management governance) and production processes are accelerated by doing things in parallel (organic production governance). Crafted Quality is tantamount to taking an Agile approach. A key benefit of the Crafted Quality approach is faster delivery – exactly what is required in the

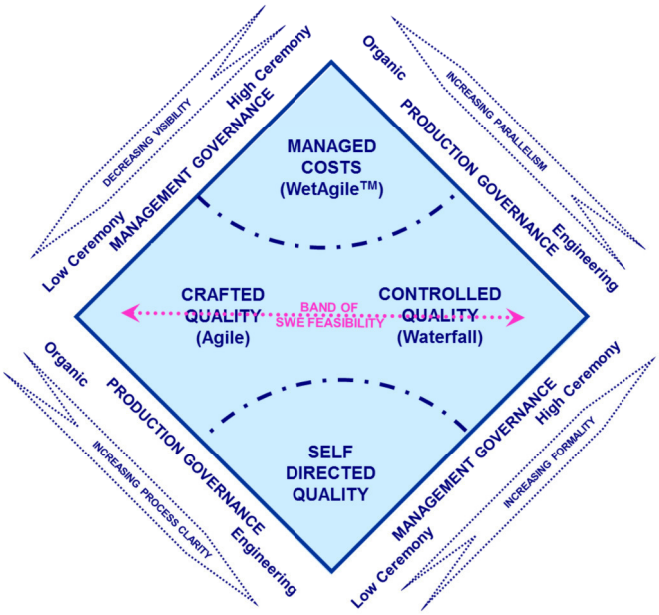


Figure 2: CASM – Insyte's Complex Adaptive Situational Model

competitive environment of the information age organization. Crafted Quality is the consequence of Agile Management meeting Agile Development.

The metaphor chosen for this domain is a curved arrow. It emphasizes the adaptive nature of an agile team that can rapidly respond to change.

But Crafted Quality does not only have benefits. Product rapidly brought to market is often not nearly defect-free, leading to potentially expensive re-work. This outcome is well demonstrated by software product with early releases that are plagued by defects that are only eradicated after a number of upgrades to the product have been implemented.

B. *Controlled Quality (The Cube)*

In this domain the emergent behavior derives from constraints of engineering-style production governance and formal, high ceremony management governance. A well-executed Waterfall approach to software engineering exemplifies Controlled Quality. Such an approach is described in the article *"They Write the Right Stuff"* by Fishman [25]. One of the key benefits of the Controlled Quality approach is that quality requirements are formally addressed at each stage of the life cycle – both in terms of initially specifying the requirements and subsequently verifying fulfillment thereof.

The metaphor chosen for this domain is the cube. It emphasizes the disciplined nature of a team operating under conditions of thorough governance.

But Controlled Quality does not only have benefits. A number of situational characteristics must apply for Controlled Quality to deliver value. These include the ability to drive out and specify requirements and having the time and other resources to analyze, specify and design the full-scope solution. And attempts to drive out full scope requirements, architecture and design can easily lead to a state of "analysis paralysis" which CASM calls "debilitating bureaucracy".

C. *The Band of Software Engineering Feasibility*

It is not by accident that CASM is represented as a diamond-shaped model. This layout places the emphasis on what is called "the band of software engineering feasibility" which stretches from Crafted Quality, Agile at the one end, to Controlled Quality, Waterfall at the other. Depending on circumstances, this implies that the state of dynamic equilibrium of a software engineering team can exist anywhere along the band and still produce value-adding results. An implied characteristic of the software engineering band of feasibility is that it is supported by a culture of "management-by-measurement", meaning that, no matter whether the way of working is Agile or Plan-Driven, management will be enabled by taking, analyzing and responding to relevant measurements.

The evolving, risk-driven approach described by Boehm in May 1988 in *"A Spiral Model of Software Development and Enhancement"* [26] could very well be understood to be a journey across the software engineering band of feasibility, with the first iteration being fully Agile, and the last solidly in Controlled Quality territory.

SEMAT Essence enables practices to define life-cycles, whether Agile or Plan-Driven, by sequencing a number of patterns, one for each phase and/or milestone in the life-cycle The life-cycles are illustrated using the template shown in Fig. 3.

Each Kernel Alpha and its states are shown in a vertical column with their creation at the top and their destruction at the bottom. Milestones are shown as a vertical bar across the grid starting with an inverted triangle to represent the milestone and continuing with a white line over which are shown the states to be achieved to successfully pass the milestone. Where achieving a state is either recommended or optional, the state is shown with a dashed outline and italicized text.

Using the template illustrated in Fig. 3, a sub-clause in the submission to the OMG provides illustrations of a few typical software engineering life-cycles, including an Exploratory Process life-cycle (Crafted Quality) and a Waterfall life-cycle (Controlled Quality). Readers who would like to review these models are urged to access the OMG submission [27].Another useful example is to be found in *"Agile and SEMAT – Perfect Partners"* [28].

This suggests that various instantiations of the SEMAT Life Cycle Model could be placed at various points across the software engineering band of feasibility.

Thus far we have considered Crafted Quality (Agile) and Controlled Quality (Plan-Driven). But what of the other two domains that are not in the band of feasibility?

D. *Managed Costs (The Explosion)*

This condition emerges when high ceremony management governance is applied to a situation that has been given the freedom of organic production governance. This means management expects Controlled Quality behavior while simultaneously giving developers the organic freedom of Agile production. A somewhat dysfunctional expectation as described in *"Corporate Information Systems Management"*

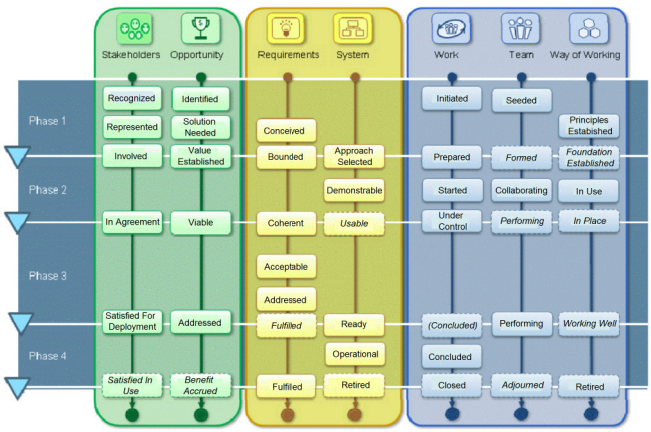


Figure 3: SEMAT Life-cycle Template

(1999) by Applegate, McFarlan and McKenney [29]. The Managed Costs name emphasizes that management will focus on cost and budget control while being quite disconnected from the day-to-day, technical activities of the team. Could this explain why Appelo's survey (referenced above) identified the following factors as contributing to failed agile approaches?

- Management opposed to change – hanging on to high ceremony management governance.
- Loss of management control – that is perceived to happen when moving to a low ceremony approach.
- Lack of engineering discipline – due to organic production approach.
- Organizational need for planning, predictability and documentation – associated with high ceremony management governance.

Steve Pieczko [30] suggests that Managed Costs might be a hybrid condition experienced by a team that is migrating from Controlled to Crafted Quality and, while still "dripping from the waterfall", they're trying to be agile. Hence the name "WetAgile", introduced in 2010 by Pieczko.

A possible explanation, but this author has experienced a number of situations where the somewhat dysfunctional, Managed Costs state seems to be permanent and a breeding ground for "management-by-politics".

The metaphor chosen for this domain is the explosion which emphasizes the often crisis-driven reality of the domain.

E. Self-Directed Quality (The Sphere)

When low ceremony management governance interacts with engineering production governance, the resulting state is Self-Directed Quality (SDQ) (The Sphere). A somewhat surprising situation. Why would practitioners elect to be constrained by engineering production governance when management governance expects no more than low ceremony? Followers of Controlled Quality would see this as an unexpected bonus, while Crafted Quality "agilista" might think of it as madness. This author suggests two possible explanations that need to be tested by means of further research.

The first might be because the tools being used enforce typical engineering production governance. It was for this reason that first generation CASM actually called this domain "Automatic Quality" [31].

A second explanation is that small, (one-person?) software development initiatives might be executed by individual/s who prefer to follow the defined stages of the engineering life-cycle. It might well be that some developers of open source software prefer to adopt this way of working.

The metaphor chosen for this domain is the sphere, emphasizing (from an engineer's point of view), the utopian situation where effective engineering is performed with few management constraints.

F. When Things Go Wrong

We earlier introduced Stephen Wolfram's proposed classification scheme where Class IV corresponds to complexity. Classes I and II correspond to order and Class III to chaos. As illustrated in Fig. 4, CASM's four domains are suggested to correspond to four types of Class IV - Complexity, with Classes I and II lying to the right of the band, and Class III to the left.

When the freedom of Crafted Quality is abused, the situation typically degenerates into a state of chaos (Class III).

Inappropriate responses to governance that desires a Controlled Quality outcome can easily result in creation of Class I or II situations with excessive order –experienced as debilitating bureaucracy.

G. CASM Characteristics

CASM has been introduced as a model of styles of team behavior and in broad terms, any software engineering team could, in response to the governance constraints imposed, be in any one of the four states of dynamic equilibrium. Each state has a set of defining characteristics. Table 1 describes characteristics of the domains. The table derives from published work (Myburgh [32], Boehm and Turner [33]) as well as from experience gained through practical application of the model.

VI. Bringing CASM To Life – Options for Management And Teams

CASM allows management to understand levels of governance that should be applied according to characteristics of the situation. The software engineering team then responds chaordically and achieves a state of dynamic equilibrium that is situationally appropriate. Table 2 identifies a number of these situational characteristics and suggests appropriate governance that should be applied. The table is based on the assumption that we are trying to pin-point the required point of dynamic equilibrium in the band

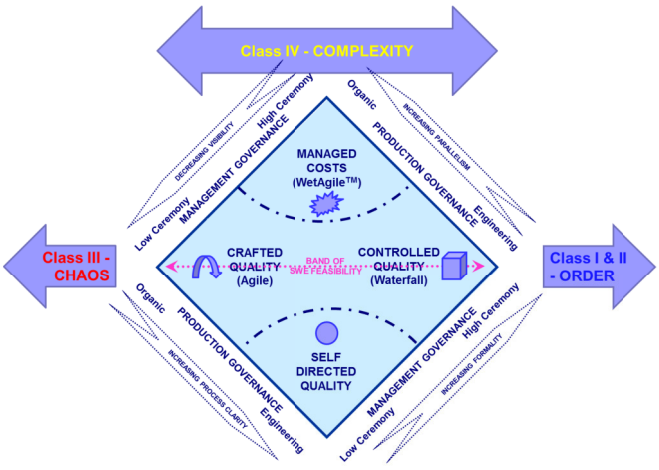


Figure 4: CASM – Modeling States of Complexity

Table 1 – CASM Domain Characteristics

	CRAFTED QUALITY (CrQ) (AGILE)	CONTROLLED QUALITY (CoQ) (WATERFALL)	MANAGED COSTS (MaC) (WETAGILE)	SELF-DIRECTED QUALITY (SDQ)
MANAGEMENT				
CUSTOMER RELATIONS	Dedicated on-site customers	As-needed, formal customer interaction	As-needed, formal customer interaction	As-needed, formal customer interaction
	Focused on prioritised increments	Focus on formal contract provisions	Focus on formal contract provisions	Focused on prioritised increments
	Best Business Practice tends to dominate (Risk-taking)	Best Solution Delivery Practice tends to dominate (Risk-avoiding)	Best Solution Delivery Practice tends to dominate (Risk-avoiding)	Best Solution Delivery Practice tends to dominate (Risk-avoiding)
PLANNING & CONTROL	Internalised plans (low visibility)	Formal, documented architecture and plans	Formal, documented architecture and plans	Internalised plans (low visibility)
	Evolutionary delivery	Incremental or full-scope delivery	Incremental or full-scope delivery	Incremental or full-scope delivery
	Qualitative control	Quantitative control	Quantitative control	Qualitative control
	Classic PMBOK practices less feasible (parallel approach)	Classic PMBOK practices more feasible (sequential approach)	Classic PMBOK practices more feasible (sequential approach)	Classic PMBOK practices less feasible (parallel approach)
	Sometimes "populate first", then plan. Meaning that the team is formed before a plan is available for the work to be done	Often "plan first", then populate. Meaning that the team is formed in response to the needs of a well-defined plan.	Often "plan first", then populate. Meaning that the team is formed in response to the needs of a well-defined plan.	Sometimes "populate first", then plan. Meaning that the team is formed before a plan is available for the work to be done
	Risk contained by time-box	Risk contained with Management Reserve	Risk contained with Management Reserve	Risk absorbed by individual
COMMUNI-CATION	Tacit interpersonal knowledge (low visibility)	Formal, documented architecture & knowledge	Formal, documented architecture & knowledge	Formal, documented architecture & knowledge
PROCESS	Innovative, "black-box", empirical processes	Deterministic, "white-box", defined process sequence	Deterministic, "white-box", defined process sequence	Deterministic, "white-box", defined process sequence
	More often unique initiatives	More often repeatable processes and continuous improvement	More often repeatable processes and continuous improvement	More often repeatable processes and continuous improvement
	Repeated initiatives remain challenging	Repeated projects can become jobs	Repeated projects can become jobs	Repeated projects can become jobs
	Could become chaotic	Usually well organised	Usually well organised	Usually well organised
TECHNICAL				
REQUIRE-MENTS	Prioritised informal stories and test cases	Formalised project capability, interface, and quality. architecture	Formalised project capability, interface, and quality. architecture	Prioritised informal stories and test cases
	Undergoing unforeseeable change	Foreseeable evolution requirements	Foreseeable evolution requirements	Undergoing unforeseeable change
DEVELOP-MENT	Evolving architecture	Guided by full-scope architecture	Guided by full-scope architecture	Evolving or full-scope architecture
	Simple design	Extensive design	Extensive design	Simple or extensive design
	Short increments	Longer increments	Longer increments	Short increments
	Re-work assumed inexpensive	Re-work assumed expensive	Re-work assumed expensive	Re-work assumed expensive
TESTING	Executable test cases define requirements	Documented test plans and procedures	Documented test plans and procedures	Formal test plans and procedures
PERSONNEL				
CUSTO-MERS	Dedicated, collocated CRACK performers	CRACK performers, not always co-located	CRACK performers, not always co-located	Dedicated, collocated CRACK performers
DEVELOP-ERS	Led by those who revise methods to meet situation	Less involvement of those who revise methods	Less involvement of those who revise methods	Less involvement of those who revise methods
	Learn largely by doing	Learn largely by reading	Learn largely by reading	Learn largely by reading
CULTURE	Many degrees of freedom	Framework of policies and procedures	Framework of policies and procedures	Self-limited degrees of freedom
	Thriving on chaos	Thriving on order	Thriving on order	Thriving on order
APPLICATION				
PRIMARY GOALS	Rapid value	Predictability, Stability	Predictability, Stability	Predictability, Stability
	Responding to change	High assurance	High assurance	High assurance
SIZE	Smaller teams and projects	Larger teams and projects	Larger teams and projects	Smaller teams and projects
ENVIRON-MENT	Turbulent	Stable	Stable	Turbulent or stable
	High change	Low change	Low change	High or low change
	Project-focused	Project/organisation focused	Project/organisation focused	Project-focused

CRACK = Collaborative, Representative, Authorised, Committed, Knowledgeable

Table 2 – Appropriate Software Engineering Responses to Situational Characteristics

SITUATIONAL CHARACTERISTIC	YES THIS DESCRIBES THE SITUATION	NO THIS DOES NOT APPLY
Are requirements readily definable?	A CoQ, Waterfall way of working could be adopted on condition that delivery time-scales permit.	The CrQ, Agile way of working is required.
Is there a comprehensive architectural description for the solution?	A CoQ, Waterfall way of working could be adopted. If the scope of delivery is large, an incremental approach will mitigate risk by delivering regular, pre-planned increments.	The CrQ, Agile way of working is required.
Is there pressure to rapidly produce results?	The CrQ, Agile way of working is required.	A CoQ, Waterfall way of working could be adopted on condition that requirements are definable.
Is there pressure to produce accurate schedules, budgets & estimates?	If the accuracy is to be based on schedules, budgets and estimates that are derived from a detailed action plan for the initiative, then a CoQ, Waterfall approach is required. If the accuracy is to be based on the cost per time-box, then a CrQ, Agile approach is indicated.	Schedules, budgets and estimates can be based on the cost per time-box and a CrQ, Agile way of working is suggested.
Does the size of the initiative introduce significant risk?	A CoQ, Waterfall approach is suitable for mitigating this risk – on condition that other characteristics required for CoQ also pertain.	The CrQ, Agile way of working is suggested so that the overhead associated with CoQ, Waterfall can be avoided.
Will implementation of the solution introduce significant change?	A CrQ, Agile way of working allows for resistance to change to be mitigated by limiting the extent of change associated with each iteration. An incremental, CoQ, Waterfall approach could also be used to limit the extent of change introduced during each increment.	Either CoQ, Waterfall or CrQ, Agile approaches could be viable. Other situational characteristics will influence the decision.
Is there significant risk due to technology? (This suggests that unproven, state of the art technology is to be implemented).	A CrQ, Agile approach should be followed by a team that is mandated to experiment with and get to know the new technology.	Either CoQ, Waterfall or CrQ, Agile approaches could be viable. Other situational characteristics will influence the decision.
Does cost-of-failure represent a source of significant risk?	A CoQ, Waterfall way of working should be adopted to allow for product assurance. If the scope of delivery is large, an incremental approach will further mitigate risk by delivering regular, pre-planned increments that could be separately assured.	Either CoQ, Waterfall or CrQ, Agile approaches could be viable. Other situational characteristics will influence the decision.
Does the software engineering team collectively have a high level of competence?	The team should be able to adapt to whatever way of working is situationally appropriate.	This situation represents a significant source of risk, and attempts to adopt a CoQ way of working could easily result in "debilitating bureaucracy", whereas CrQ approaches are likely to evolve into "freelance chaos".

of feasibility. Hence suggestions made refer only to options that lie in this band.

VII. Bringing CASM To Life – Options For The Enterprise

The above analysis of situational factors demonstrates that different ways of working apply to different situations. A small team of software engineers who are working on a focused initiative can be expected to adjust their way of working to be appropriate to the situation. However, in larger organizations where many teams are tackling many initiatives, one could expect different teams to be in different states of dynamic equilibrium at the same time.

To better understand this, we can consider the idea of a hierarchy of people-based complex adaptive systems – a system of complex adaptive systems. Working from the bottom up, we first find an individual person. (Remembering that a single person is already a CAS). If a few people collaborate towards achieving the same goal/s, we discover the next level CAS, viz. a team. Teams could also be contributing to achievement of common goal/s and hence a collection of teams could define the next higher level. For the purpose of this discussion, the highest level CAS will be the enterprise itself. Now, by employing the various metaphors associated with each state of dynamic equilibrium, the diagram in Fig. 5 represents an enterprise as a collection of complex adaptive systems and the metaphor for the Enterprise is suggested to be an amoeba.

VIII. Conclusion

The Complex Adaptive Situational Model described in this paper promotes understanding of what it takes to

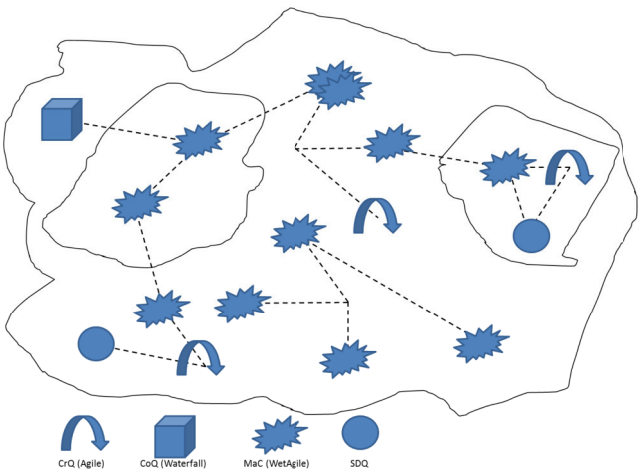


Figure 5: The Enterprise "amoeba" - represented as a System of Complex Adaptive Systems

establish conditions which enable software engineering success, not only with agile approaches, but also traditional, plan-driven software engineering.

Influenced by complexity science, CASM explains aspects of the state of dynamic equilibrium that is achieved by a software engineering team under the constraining influence of management and production governance.

Four states of dynamic equilibrium are defined: Crafted Quality (Agile), Controlled Quality (Waterfall), Managed Costs (WetAgile) and Self-Directed Quality. A band of software engineering feasibility is also described and successful software engineering initiatives require teams to operate in that band which stretches from Crafted Quality to Controlled Quality. Management's challenge is to apply appropriate governance to enable the required state of dynamic equilibrium.

The journey across the band of feasibility is further described by introducing SEMAT, with Crafted Quality amounting to applying SEMAT Essence, and Controlled Quality being achieved by introducing additional practices which satisfy the more stringent governance requirements.

CASM in its four states then allowed introduction of the idea of describing an enterprise as a collection of complex adaptive systems, thereby setting the scene for further research into the complexities of human-driven complex adaptive systems.

ACKNOWLEDGMENT

Many thanks to Dr. Alastair Walker for early support of the model and coining the name of the *Managed Costs* domain. Thanks also to Dr. Barry Dwolatzky for continued support and the opportunity to further develop the model. Dr. Whitey van der Linde is thanked for helping to find the links between the model and complexity science. The substance that SEMAT gives to Situational Software Engineering is primarily thanks to Dr. Ivar Jacobson, one of the founders of SEMAT.

Jurgen Appelo receives special thanks for the overview of complexity and complex adaptive systems that is at times paraphrased from: *"Management 3.0 – Leading Agile Developers and Developing Agile Leaders"* (2011). Readers of this article are encouraged to also read Appelo's publications [http://www.mgt30.com/].

Thanks also to other reviewers of the article including Adrian Schofield, Steve Piezcko, Paul MacMahon. Their feedback improved the final product.

REFERENCES

[1] J. Appelo, Management 3.0 – Leading Agile Developers, Developing Agile Leaders, Addison-Wesley, 2011, p. 2.
[2] J. Appelo, Management 3.0 – Leading Agile Developers, Developing Agile Leaders, Addison-Wesley, 2011, p. 11.
[3] Johnson, P, Ekstedt, M, Jacobson, I, "Where's the Theory for Software Engineering?,"http://dx.doi.org/10.1109/MS.2012.127, pp. 94-96.
[4] P. Kruchten, "Casting Software Design in the Function-Behavior-Structure Framework,"http://dx.doi.org/10.1109/MS.2005.33, pp. 52-58.
[5] A. Abran et al., eds., Guide to the Software Engineering Body of Knowledge, IEEE CS Press, 2004.
[6] D. Hock, Birth of the Chaordic Age, Berrett-Koehler Publishers, Inc., 1999, pp. 205-207.
[7] "Agile Manifesto," [Online]. Available: http://agilemanifesto.org/. [Accessed 15 January 2014].
[8] J. Appelo, Management 3.0 – Leading Agile Developers, Developing Agile Leaders, Addison-Wesley, 2011, p. 28.
[9] OMG, "Essence - Kernel And Language For Software Engineering Methods 1.0 - Beta 1," July 2013. [Online]. Available: http://www.omg.org/spec/Essence/1.0/Beta1/.
[10] I. Jacobson, P.-W. Ng, P. E. McMahon and I. Spence, The Essence of Software Engineering - Applying the SEMAT Kernel, http://dx.doi.org/10.1145/2380656.2380670.
[11] M. M. Waldrop, Complexity: the Emerging Science at the Edge of Order and Chaos, http://dx.doi.org/10.1063/1.2809917, p 145.
[12] D. Hock, Birth of the Chaordic Age, Berrett-Koehler Publishers, Inc., 1999, p. 3
[13] P. Senge, The Fifth Discipline - The Art and Practice of the Learning Organization, http://dx.doi.org/10.1108/eb025496.
[14] F. P. Brooks, Jr., "No Silver Bullet - Essence and Accidents of Software Engineering", http://dx.doi.org/10.1109/MC.1987.1663532.
[15] J. Appelo, Management 3.0 – Leading Agile Developers, Developing Agile Leaders, Addison-Wesley, 2011, pp. 50-51.
[16] J. Appelo, Management 3.0 – Leading Agile Developers, Developing Agile Leaders, Addison-Wesley, 2011, p. 58.
[17] J. Appelo, Management 3.0 – Leading Agile Developers, Developing Agile Leaders, Addison-Wesley, 2011, p. 100.
[18] M. M. Waldrop, Complexity: the Emerging Science at the Edge of Order and Chaos, http://dx.doi.org/10.1063/1.2809917, pp 225-226.
[19] J. Appelo, Management 3.0 – Leading Agile Developers, Developing Agile Leaders, Addison-Wesley, 2011, p. 100, p. 153.
[20] J. Appelo, Management 3.0 – Leading Agile Developers, Developing Agile Leaders, Addison-Wesley, 2011, p. 100, p. 154.
[21] J. Appelo, Management 3.0 – Leading Agile Developers, Developing Agile Leaders, Addison-Wesley, 2011, p. 100, p. 101.
[22] A. J. B. Myburgh, "Successful Combinations of Software Engineering Strategy and Project Management," in Proceedings of the SAIEE Symposium "Professional Issues in Software Project Management - 5 September 1990", Johannesburg, 1992.
[23] A. J. B. Myburgh, "Towards Understanding The Relationship Between Process Capability And Enterprise Flexibility," in Proceedings of "SAATCA 8th International Systems Auditor Convention 24 – 25 August 2005", Johannesburg, 2005.
[24] B. Boehm and R. Turner, Balancing Agility and Discipline – A Guide for the Perplexed, Addison-Wesley, 2004.
[25] C. Fishman, "They Write the Right Stuff" 2007. http://www.fastcompany.com/magazine/06/writestuff.html, Last Accessed 22 June 2014.
[26] B. W. Boehm, "A Spiral Model of Software Development and Enhancement", http://dx.doi.org/10.1109/2.59, pp. 61-72.
[27] OMG, "Essence - Kernel And Language For Software Engineering Methods 1.0 - Beta 1," July 2013. [Online]. Available: http://www.omg.org/spec/Essence/1.0/Beta1/, p 267-271.
[28] I. Jacobson, I. Spence and P. Ng, Agile and SEMAT – Perfect Partners, http://dx.doi.org/http://dx.doi.org/10.1145/2380656.2380670..
[29] L. M. Applegate, F. W. McFarlan and J. L. McKenney, Corporate Information Systems Management - 5th Edition, McGraw-Hill, 1999, p. 184.
[30] S. Pieczko, "Waterfall? Agile? How About WetAgile?", 2010. http://www.WetAgile.com, Last Accessed 18 June 2014.
[31] A. J. B. Myburgh, "Successful Combinations of Software Engineering Strategy and Project Management," in Proceedings of the SAIEE Symposium "Professional Issues in Software Project Management - 5 September 1990", Johannesburg, 1992, p. 94.
[32] A. J. B. Myburgh, "Towards Understanding The Relationship Between Process Capability And Enterprise Flexibility," in Proceedings of "SAATCA 8th International Systems Auditor Convention 24 – 25 August 2005", Johannesburg, 2005.
[33] B. Boehm and R. Turner, Balancing Agility and Discipline – A Guide for the Perplexed, Addison-Wesley, 2004.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 851–860
DOI: 10.15439/2014F98
ACSIS, Vol. 2

Requirement Engineering for Effective Mobile Learning: Modelling Mobile Device Technologies Integration for Alignment with Strategic Policies in Learning Establishments

Remy Olasoji
University of East London
Docklands Campus,
University Way,
London E16 2RD, UK
Email: r.olasoji@uel.ac.uk

David Preston
University of East London
Docklands Campus,
University Way,
London E16 2RD, UK
Email: d.preston@uel.ac.uk

Amin Mousavi
University of East London
Docklands Campus,
University Way,
London E16 2RD, UK
Email: s.a.mousavi@uel.ac.uk

Abstract—In spite of several efforts in the last decade by researchers and educators, expected potential from the use of mobile device technology (MDT) to effect learning transformation is largely unfulfilled. Quantifying benefits of MDTs in learning, either through achievement of learning objectives or enhancement of the process, remains problematic. Rapid changes in development and manufacture also continue to present additional challenges. Most trials typically employ use case approach of evidencing benefits through usage experience. In this paper, application of Requirement Engineering (RE) methodologies to specify goals and requirements for mobile learning (ML) is proposed. Alignment with teaching and learning strategies as well as other institutional goals and strategies will be essential for successful integration of MDTs in learning. RE techniques can be used to achieve this. Finally, a case study is presented to illustrate the use of some of the approaches proposed.

I. Introduction

In recent past, educational communities have appropriated available innovative technologies to enhance learning and transform the system. This has led into the coining of the discipline "educational technology", concerned with the effective facilitation of e-learning and technologies in learning. Technology appropriation has itself become a discipline of sorts, a way of exploring the impact of a given technology on a community or the society at large [1]. In spite of efforts with often mixed results, the community remain on the left foot forward; playing catch-up as advances in technology break moulds, making previously innovative ideas obsolete even before they have begun to take shape.

With increasingly powerful computing capabilities and affordance of convergence between multiple devices such as audio, video, camera etc., MDTs are transforming societal constructs and interactions. Businesses take advantage of advances to streamline their business operations; individuals, groups and communities use them to augment their life styles and coordinate relationships. The society as we know it in this generation is rapidly changing as a result [2], [3].

Within the modern UK educational sector, students admit to using devices to facilitate access to learning and for personal development but see no sustained use in most of their learning sessions and classrooms. Some educators confess they have no clue about precise benefits to learning or how this may be applicable to their teaching practice. Many schools and Further Education (FE) colleges impose an outright ban on the use of mobile devices by students within school premises, believing they are disruptive and problematic for classroom management. Many Higher Educational Institutions (HEIs) have implemented Bring Your Own Device (BYOD) schemes as part of their IT support provision strategies without fully exploring support, privacy and security issues [4], [5], [6].

The UK government is not left out in efforts to facilitate mobile device usage and information mobility in the society. Robust internet and WiFi connectivity is part of the policy strategies of the current government [7]. JANET, the body responsible for providing free public access to WiFi for FE and HE (Higher Education) establish a partnership with BskyB's The Cloud, "one of the UK's leading public WiFi providers" in November 2013, ensuring free and robust service for "over 18 million end-users in UK research and educational sector [8]. Interestingly, some of the staff respondents to a mobile learning research conducted in schools and FE colleges admit they either have zero or very little support to connect or are unsure of how to use them in teaching and learning practices in a mobile learning study conducted recently.

In some respects, BYOD schemes are sometimes little more than strategic ploys to minimise infrastructure costs while ensuring they are competitive in their provisions without really addressing underlying problems. And many support staffs admit they are struggling to support some of the less common or recently released devices. Many also feel they are unequipped or unsupported by the organisation; with no training and / or expert support knowledgebase [9].

The almost 'lightning-speed' pace of advances in MDTs continue to present both potentials and challenges. Previously innovative instructional designs become obsolete almost as soon as implemented. Yet, many remain in use for years well beyond use-by dates. Regardless, some would say mobile learning is here to stay. Others would add, perhaps cynically, that there is no evidence of actual learning involved in some efforts, the devices used primarily for access and delivery only for the most part [10], [11].

978-83-60810-58-3/$25.00

In this paper, it is proposed that application of domain neutral Requirement Engineering (RE) techniques may provide more insights into these problems, and perhaps help align business goals and requirements. Preliminary work relating to the use of RE techniques to explore current issues in mobile learning (ML) and the relationship between MDTs and education is first presented, followed by a case study illustrating some of the approaches proposed.

II. Engineering the current reality

Reference [12, pp. 7] defines Requirement Engineering (RE) as "a set of activities concerned with identifying and communicating the purpose of a software-intensive system". Software-intensive systems are described as systems comprising of some form of hardware or networked components, and involving human interactions and activities. Reference [12, pp. 3] added that RE "provides a framework for understanding the purpose of a system and the contexts in which it will be used", bridging "the gap between an initial vague recognition that there is some problem ... to building a system to address the problem".

Another definition for RE proposed in the year 2000 by [13] stated its suitability for specifying what the authors called "real-world goals" i.e. reflecting the tendency for change in the real world. More recently, [14, pp. 42] agrees, adding that "each RE process starts with an aim to change the current reality". The author stated all software systems are used within a context, adding that while system goals may be clearly defined, quite often variables within the context are not so clear. The latter may explain the rationale for a look to RE methodologies for ML systems. Although not strictly software-intensive applications but many interconnecting systems and technologies; the very nature of the system make it a likely domain for the application of RE.

Mobile devices have progressed from voice communication tools into computerised devices that not only enables easy collaborations between geographically dispersed individuals, but those also creating convergences between multiple media devices. Crucially, they are also providing means of connecting varieties of systems in ever increasingly complex contexts. For the sake of simplicity, these technologies will be referred to in this paper collectively as "mobile device technologies" or MDTs; encompassing mobile devices, convergence affordability, communication channels, remote, local and wireless network connectivity etc. Usage context is that relating to learning establishments, and HEIs in particular.

III. Requirements engineering techniques selection

According to [15], there is no one prescriptive way of applying RE techniques to a system but the authors caution on ensuring techniques are applied early in the system lifecycle. With so many to choose from, techniques will largely depend on the system goal and contexts. A major weakness found in many is their complexity and lack of clarity, making them unusable by anyone but experts in RE or software engineering techniques [16].

Regardless, many authors agree the following core stages are essential in RE [15], [17], [18]:

- Inception and elicitation
- Identification, analysis and negotiation
- System modelling and goal specification
- System validation, risk and change management

Some of the activities involved in each of these stages will be discussed next.

A. Elicitation of needs

The bulk of the fact finding process in RE is usually in the inception and elicitation phase. However, elicitation is also a task that will continue throughout the life of the project and beyond system implementation. For example, whenever changes are made to a system, requirements for those changes have to be re-evaluated [18], [19]. Reference [18] cautioned that not all the information obtained would become requirements. Some needs may not be feasible to implement in the final product.

There are several stakeholders with potential input into the ML system including device manufacturers, educators, students, policy makers and those in the role of learning support and governance. While device manufactures may not be particularly interested in prioritising the needs of the educational community at the exclusion of others, they are likely to be concerned if their product(s) are unusable by members within the community. If the device is overly complicated then consumers, who may also be students and / or educators will not want them. Device manufactures may also be concerned about policies preventing freedom of usage in learning establishments.

Educators are often keen to appropriate technology that would make their practice more effective and achieve learning objectives. They are however unlikely to want to give up too much of their time for pedagogic and instructional design. In the same way, students may have devices but unable to use them effectively for learning. Seamless usage may also be problematic because the necessary connections and support are not adequate or robust enough, or there may be policies prohibiting use [20].

For learning providers, as may be true also for educators and students, running cost is still an issue. Costs may also include provision of ongoing technical support by the institution. Interoperability with other applications on the local network systems will be essential and ensuring the environment is rich enough to support such levels of inter-connectivity may be beyond sustainable budgeting strategies. And while mobile devices include tablets / devices with wide screens, powerful media support features and educational affordances, there are many with less than satisfactory experience still. It is believed this will increasingly become less of an issue [10].

Personal preference and cultural perceptions will also play a key role in intentions to use. For instance, in the past majority of consumers are uncomfortable conducting financial transactions on mobile devices. Today, the number is growing despite persisting security concerns [21]. Possibilities of cyber bullying and abuse are other issues among others. In a research conducted recently, a group of staff and students in an HEI stated of mobile devices: "can cause epilepsy – when it does not work", "too dangerous" and "very dis-humanising". These statements may illustrate extreme opinions held by some stakeholders still. Thus, these and other concerns must be elicited carefully. It is also important to identify sub-groups within each stakeholder groups with potentially differing opinions. For example, educators group may include tutors or pre-service tutors who may also be students themselves in HE. Reference [22] called this purposive sampling; describing the conscious inclusion (and exclusion, as well as the contextual grouping) of certain groups of participants.

Techniques used in elicitation may typically be employed in other RE stages including those for eliciting and analysing goals for the system, which [18] suggest is sometimes overlooked but an important part of fact finding. Establishing goals may in fact aid requirement analysis and can be analysed using goal modelling. One of the more commonly known goal modelling techniques is referred to as KAOS (Keep All Objectives Satisfied) [18], [23] citing [24], [6]. KAOS specify the use of verbs as well as AND / OR operators to link goals to processes [25]. Goal modelling will be discussed and illustrated more in subsequent sections.

Other elicitation techniques include ethnographical research methods [15]. Ethnography is an exploration of the community concerned and the cultural contexts using quantitative methods such as surveys; and qualitative methods such as observations, interviews and focus group studies. In this manner, interests and the emotional appeal of components within the system or the product being developed can be measured [18]. Brainstorming and prototyping may also be employed during the elicitation stage.

B. Identification, analysis and negotiation

This is a logical stage following directly or conducted in parallel with the elicitation of requirements. Information obtained from stakeholders need to be analysed, categorised and ranked. What are the current and new requirements? Who are those involved and where are they located? What are priorities for the business or organisation, and what are the conflicts? Conflicting requirements or potential problems must be identified and resolution decided.

Stakeholder agreement on the goals and requirements could be difficult to obtain without negotiation. Alternative options and acceptable compromises must be presented to resolve complex dissensions and disagreements on requirements and / or goals. Identifying and phrasing the most important goals for the system in terms all stakeholders can agree with and understand, may also be useful [17].

Establishing agreement on root problems can be problematic as in the ML system. Many of the stakeholders may be steeped in blame culture, making buy-in from stakeholders difficult. Even when buy-in is assured, having input from several groups of stakeholders may present a problem for the study. Reference [26] suggest the use of trade-offs adding that it is impossible to satisfy all the requirements by one specification quite often; usually typical of non-functional requirements. An example of trade-off analysis for the ML system can be seen in Table 1. The table shows strengths in opinions and level of importance by doubling or tripling certain symbols.

Reference [18] propose negotiations and brainstorming in several scheduled Quality Attribute Workshops (QAWs). In QAWs, the facilitator creates a Quality Attribute Scenario (QAS) for each of the concerns expressed by a stakeholder. Each stakeholder can express two or more of their most important concerns. The QAS is presented to the group and a handful is selected and debated. Finally, the facilitator supports the group to identify important requirements to be included in the system.

Another potential problem could arise from volatility in functionalities and the increasingly convergence nature of MDTs. Establishing meaning and interpretations of requirements may be difficult, or worse impossible if device features keep changing [18]. Some level of stability may need to be assumed or achieved. Other techniques employed may include prototyping, global analysis, focus group, requirement analysis and release planning [27], [28].

C. System modelling and goal specification

Modelling is an essential RE technique often used to analyse requirements as well as goals at various stages throughout the process lifecycle. Some of the more commonly used modelling techniques are listed below [18]:

- **Artefact modelling:** Used to define the work products and interdependencies and to specify maintenance requirements for processes.
- **Goal-oriented modelling:** Concerning the needs and vision of the business organisation and not necessarily the customers or users of the service(s) or system products.
- **Model-driven RE (MDRE):** Model-driven requirement engineering is typically used for large complex systems and can span the project lifecycle, from inception through to maintenance.

Other modelling techniques used in RE include **feature** and **process** modelling, typically used during the elicitation phase. In Sections IV and V, modelling techniques applicable to the ML system are illustrated in more details.

D. System validation, risk and change management

During this stage system model(s) and specification are evaluated against requirements and agreed. Validation process can often be the most complicated part of RE, resulting in inability to reach a consensus agreement, especially where different stakeholders with conflicting opinions and goals are involved. Risks to the system are identified and measures established to minimise their effect on future optimum performance of the system and to manage changes.

Reference [13, pp. 6] warns, "If stakeholders do not agree with the choice of problem frame, it is unlikely that they will ever agree with any statement of the requirements". The authors suggest a resolution may be to promote an agreement "without necessarily making the goals explicit"; in other words, rephrasing goals and requirements using terms that may be more moderate than specific.

Several RE methods have been suggested for investigating ML and similar systems, and for aligning the goals of the system with learning / business strategies. In the next two sections, categorizing and modelling techniques are explored in more details. A case study using goal modelling to specify some goals for the system is next presented. Information used in the goal model will be extracted from corporate and operational strategies of a UK HEI, demonstrating how alignment may be more easily achieved.

IV. GOALS FOR MOBILE LEARNING (ML)

Information obtained during elicitation needs to be organised, ranked and / or categorised in order to identify them as either goals or requirements of the system. This can often be complicated by the many different classification techniques available in RE. Again, the technique chosen will depend on the objectives for the system and the type of information to be analysed.

Some authors suggest goal analysis and specification is one of the methods that should be used more carefully and prioritised [29], [25]. Both of these authors believe that while many appreciate its importance, it is often side-lined in literature and formal specifications for the system. Goals are well understood to be the objectives or targets to be satisfied by the system under development, and they may often be explicitly presented to system engineers by stakeholders at project inception. The assumption then, that a formal specification for achieving those goals is all that's required may account for the oversight. Reference [25] refer to this as the "top-down" approach [pp. 3].

For [18] and [25], the initial set of goals is just the beginning of goal development process; an important basis on which to continue further analysis and refinement. Reference [25] believes that will require asking the 'HOW' and 'WHY' questions [pp. 3]. Thus, goal elicitation continues alongside establishment and elicitation of needs. Conflicts and problems are identified and resolved. New features or changes in the system will require alterations or modifications. New goals may also arise from validation, risk and change management processes [25], [18].

Goal modelling is sometimes seen as a discipline of sorts and also referred to as Goal-Oriented Requirement Engineering (GORE). This section outlines some strategies used in GORE, which may be employed throughout a project lifecycle during RE stages explored in the previous sections.

A. Classification of goals & requirements

An explicit set of goals or strategies for ML and the integration of MDTs in learning are sometimes missing from teaching and learning strategies. Many institutions would often specify a goal for technology infrastructure provision and support, of which it is assumed technologies supporting ML may be a part. It is proposed in this paper that a specification is necessary to move the agenda forward. This may be explicit or inferred from other goals or strategies. Unfortunately, such considerations have so far been glaringly omitted in past and current ML implementations and literatures.

Goals and requirements for a system may sometimes be classified as **soft** or **hard**. Soft goals describe objectives that are more 'desirable', less precise and therefore subjective; while hard goals are usually specific. Consequently, hard

TABLE I.
A SUB-SET OF TRADE-OFF ANALYSIS TABLE FOR A MOBILE LEARNING SYSTEM

	Functional requirements / stakeholders						Non-functional requirements (NFRs)*		
Goals	Students	Educators	Learning support	Governance	Government / global educational policies / acts	Device manufacturers	Pedagogic	Learning enhancement	Usability
Provide MDT accessible platform interface	✓✓✓	✓	✓	✓✓✓	✓	-	+	+	++
Provide assessments & diagnostics to support ML plans	✓✓✓	-	✓✓✓	✓	✓	-	++	+	+
Provide MDT customisable interface features	✓	✓	✓	-	✓	-			++
Provide MDT lesson planning and management features	-	✓✓✓	-	-	✓✓	-	++	+	
Provide MDT navigation & search functionality in resources	✓✓✓	-	✓✓	-	-	-			++
Provide MDT personalisation learning experience features	✓✓✓	-	-	✓✓	✓✓	-		++	++
Integrate MDT resources and / or activities	✓✓✓	✓✓✓	✓✓✓	✓✓✓	-	✓✓			++
Integrate MDT resource metadata harvesting features	✓✓	✓	✓✓✓	✓✓✓	-	-			++
Prevent integration with social networking sites	--	✓✓✓	-	-	-	-			

** Other non-functional requirements may include security, availability, etc. [14]*
Strengths or level of importance is signified by doubling or tripling of some symbols

Key
Ticks (✓): Prioritisation for stakeholder group
Dash (-): Neutral stand
Double dash (--): Disapproval, conflict or concerns
Plus (+): Relationship between goals and NFRs

goals are sometimes also referred to as functional specifications for the system. For example, specifying requirements for obtaining an educational qualification, 'passing the assessment examination' may be a "hard" goal / requirement but 'passing the assessment examination with distinctions' is not. 'Passing' is required but 'passing with distinction' can only be classified as a 'soft" goal [30], [31]. Therefore, at the top-level, most goals and requirements can be categorised into functional or non-functional. Functional requirements represent functions or actions that the system or part of the system must perform while non-functional requirements are those that measure how well those functions have been performed. While this categorisation is well suited to systems resulting in an end-product, it can be possible to miss other variances within some systems if they are not classified further and the ML system may be an example.

When the root problems in a system have not been established or agreed by stakeholder groups, goals are often unclear and subjective. RE techniques used must therefore be able to not only identify the root problems and specify requirements, but also specify goals for the system. Identifying the factors, issues and strategies within the system may be more relevant in this case. They are also particularly suited for classifying soft goals and requirements, especially those that are subject to many interpretations. It is also possible to develop use cases that can be used in testing the system from developed use case scenarios, which can be generated from the factors [32].

B. Factors, issues and strategies

Factors, issues and strategies are techniques used in global analysis; an RE methodology used to categorise "soft" goals and requirements that may not quite fit well into the functional / non-functional categorisation [30]. Reference [25] defines these as those whose "satisfaction cannot be established in a clear-cut sense"; as opposed to "hard" or requirements "goals whose satisfaction can be established through verification techniques" [pp. 3]. Global analysis is particularly suited to systems that need to be examined from several perspectives and involving many different groups of stakeholders.

Another advantage is that they can help in addressing concerns and barriers within the system when used early in the elicitation process. Classifying all the information gathered during global analysis into factors, issues and strategies may also simplify the ranking process, making it easier to prioritise goals and requirements for the system.

Factors are different from requirements, in that they do not exactly describe the system but may relate to the context or a component of the system. For example, a student stakeholder stated "I have a Blackberry but I can't use it properly and I can't sync it with my MacBook"; relating to the effective working of part of the system and achievement of the goals rather than a requirement of the system. The statement reveals a few factors:

- Synchronisation with a PC / laptop is a desired requirement.
- Some devices (e.g. Blackberry) may not sync properly with some PCs / laptops (e.g. MacBook) … OR … some students may be unaware of how to sync some devices (e.g. Blackberry) with some PCs / laptops (e.g. MacBook)

Factors are sometimes referred to as Quality Attribute Scenarios (QASs) in a general sense which will normally have related use case scenarios defined so that requirements can be linked to them and tested. When there are conflicts in factors, it is classified as an issue and where there are issues there will likely be factors to be identified and strategies to address the issues. These may be indefinite, later to be confirmed within the architectural model for the system. An example of an issue is in the following statement from another student stakeholder.

"I would use my smartphone if I was desperate as in location difficulty; internet access is limited in some places. However due to the small size of the screen I would prefer to use a tablet or a PC."

The above statement technically an issue for the goal of the system can reveal several factors:

- Internet access is limited in some places
- Small size of the screen
- There is a preference for tablet or a PC

The example has also shown how factors inherent within issues can be identified and categorised. The goals of a system can be represented by the factors. Issues can be derived goals that meet the requirement of the factors. Reference [18] refer to these as "issue-goals" and described the dynamic as that of developing a product (solution) that "satisfies a particular combination of factor-goals". Strategies can be decisions contributing the satisfaction of issue and factor goals [pp. 153]. Factors, issues and strategies need to be managed or they might grow into unmanageable levels in the global analysis [18].

C. Quality Attribute Scenarios (QASs)

QAS is another RE technique for categorising information obtained during the elicitation process. The importance of using QAS to further categorise information was mentioned briefly in previous sections. QAS is recommended in architectural requirement engineering in general for collating concerns from stakeholders and categorising them. They provide a "structured textual" way of managing stakeholder concerns and describing how it may respond to the introduction of certain stimulus. A QAS may have the following: stimulus, origin or source of the stimulus, artefact to be stimulated, stimulus context or environment, response to the stimulus and response measure i.e. satisfactory response to the stimulus as its properties [18, pp. 143].

For example, consider the following scenario in an ML system:

"In a BYOD (Bring Your Own Device) scheme in a university, a student requests support for a new type of device after staff training for known systems have been completed. An IT service support personnel was able to figure out how to resolve *the student's problem without any* need for likely costly support required from the device manufacturer nor was there any significant delay in supporting the student. The staff documented the process and trained other staff colleagues to support similar devices within one week."

The above example can be categorised into QAS parts as follows:

Stimulus: Support request for a new type of device.

Stimulus source: A student.

Artefact: The system and the IT service department.

Environment / context: After staff training for known systems has been completed.

Response: An IT service support personnel was able to figure out how to resolve the student's problem.

Response measure: No likely costly support was required from the device manufacturer nor was there any significant delay in supporting the student. The process was also replicable as part of operational strategies in the department within one week.

Not only has a QAS been defined for this scenario, it is also possible to now derive a requirement for the system, based on the QAS process:

- zero device manufacturer support
- no extra delay
- process re-engineering within one week

The following may also be inferred through the QAS process which could form part of the requirement specification:

- Since there is no device manufacturer support, there must be a limit to the types of devices that can be supported. If there is device manufacturer support in place, potentially any type of device may be supported.
- Delay in supporting the student's device may create a negative impression about the department's effectiveness.
- Process re-engineering will require a member of staff with adequate expertise to document the process and train other colleagues to carry on the process in future.
- The staff with the expertise is already a member of the university and part of the system i.e. a stakeholder within the system.

In considering the use of QAS, [18] cautions that it is important to remember there will likely be changes to stakeholders' priorities and to ensure use case scenarios are defined in addition to QAS.

D. Use case analysis and scenarios

Use case analysis is a process modelling technique used to analyse processes so that the relationship of the process within the system to external systems or components can be evaluated and understood fully. Like a QAS, use cases have several parts as follows [18]:

- Actors / users, interacting with the use case.
- Events depicted in the system causing the use case to occur.
- Pre-conditions that must be true for the use case to occur.
- Post-conditions that must be true after the use case has completed successfully.
- Activities within the use case.
- Included use cases for other processes, if any.
- Extended use cases for processes that may take place (optionally) while the use case is occurring.

Use cases are sometimes better defined using scenarios. An activity diagram can also be used to define all possible scenarios within use cases. In a QAS, scenarios involved may include those occurring during normal operations, system-as-objects i.e. passive objects and growth – dealing with changes and exploratory, as well as those dealing with scenarios that are unlikely to occur.

E. Using goal-oriented modeling techniques

Goal-oriented modelling is a useful technique for refining the goals of the business which can be associated with the requirements and needs of a system. They are particularly useful for revealing the relationship between the business goals of the system and functional as well as non-functional requirements of the system.

Review of literature has revealed that one of the problems for the sustenance of ML is the difficulty in quantifying precise benefits when used within a learning process. Defining requirements for the system from business or strategic goals of the learning establishment could be a useful way of establishing relevance to strategic decisions and processes. Goal modelling are often used with Quality Assessment Methods (QAMs), which is a measure of how the defined goals meet the desired quality expected of the system. QAMs can be used as checklist for guiding against the omission of important non-functional requirements. The goal modelling technique presented in this paper illustrates how requirements can be inferred from business goals and strategies.

There are many approaches to goal-oriented modelling, including KAOS, mentioned earlier in this paper, and Non-Functional Requirements (NFR) framework [23] citing [24], [33]. Reference [19] stated that KAOS is "the most formal application of the goal-oriented approach to deriving requirements for computer-based systems" [pp. 15].

V. DERIVING REQUIREMENTS FOR MOBILE LEARNING (ML) FROM GOALS: A CASE STUDY

There could be a disparity in what an organisation define as business goals and what is actually offered in practice. This can sometimes be very costly, leading to losses in revenue and / or goodwill branding as well as inefficiencies. Defining and implementing Quality Attribute Requirements (QARs) may guide against this or minimise the likelihood of devastating differences. Another way may be to develop requirements from the business goals of the organisation [19]. Extracts from the policies and strategies proposed in a white paper by a UK HEI will be used in this section to illustrate this. The HEI is located in London, with campuses in the East. Relevant policies in a strategy document include the following [34]:

- We will ensure that our campuses are an identifiably academic environment with innovative provision for digital mobile learning and spaces for both collaborative and reflective study.
- We will be recognised as a leading university for employability and enterprise, routinely exceeding benchmarks and providing transformational opportunities.
- In all of these areas we will seek to be at the forefront of removing barriers to progression to further study for first-generation undergraduates, supporting access to employment and postgraduate qualifications. In this way, and others, we will facilitate greater UEL student competitiveness in employment markets and subsequently through CPD for promotion and career enrichment.
- In developing a more flexible offer for a more distributed, more mobile and more time-conscious market, we will enhance our distance learning capacity, partnerships and support mechanisms.
- We do not intend to invest significant amounts of capital in these ventures, but will explore a range of collaborative models in partnership with established and new high-quality providers.
- Over the period of this Strategy, when core, full-time undergraduate numbers are likely to remain restricted, there is a greater need than ever for us to deliver our programmes at times and in places which suit the learner. Both teaching and support need to be flexible so that students can access them appropriately.

From the list above, we can identify the following goals:

- Provision for digital mobile learning and spaces for both collaborative and reflective study.
- Provision of transformational opportunities.
- Removal of barriers to progression & facilitation of competitiveness in employment.
- Development of more flexible offer for a more distributed, mobile & and time-conscious market.
- Exploration of a range of collaborative / high-quality partnerships.
- Delivery of programmes at times and in places which suit the learner.

In deriving requirements from goals, [19] suggests a successive decomposition of the goals at the high level. The author suggests using adapted notations to decompose each goal into sub-goals where either all or at least one of the sub-goals will need to be realised for the high-level goal to be satisfied. When all sub-goals must be satisfied, this may be indicated with an arc across the directional arrows. Some goal components may also become sub-goals / requirements for the system. This resulting model is sometimes referred to as goal hierarchy or goal lattice [19]. An illustration can be seen in Fig. 1.

There are several taxonomies in use for defining QARs including ISO 9126 containing 22 quality attributes, including for example the use of ambiguous terminology in definitions [18].

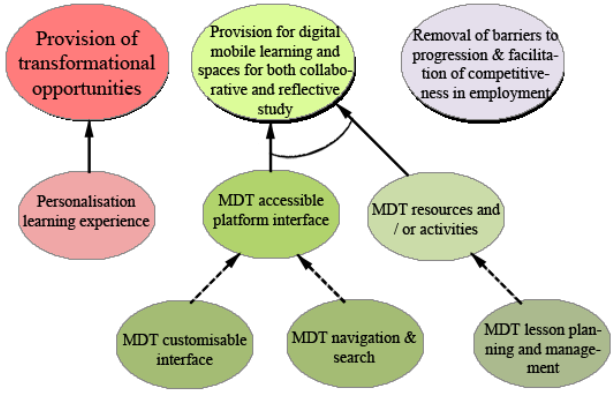


Fig. 1: Goal decomposition from business strategies

Some of the above statements / goals may fall into the category of those needing more clarity and less ambiguity which can be achieved by defining QARs. [18] suggests an integrated approach to defining QAR i.e. defining QAR from an integrated requirements model involving all the functional requirements and architecture of the organisation's operational system. For this, the authors recommend the use of an integrated artefact model (see Fig. 2) as well as goal models to show the artefacts within the system as well as the relationships linking the functional and architectural requirements.

A. Integrated artefact model for ML

Having derived requirements from the goals of the HEI (listed above) using goal-oriented modelling approach (see Fig. 1), an integrated artefact model architecture can also be created to show the relationship between the objects within the system and the attributes, as shown in Fig. 2. Defining the relationships between each of the artefacts within the system will make it possible to define QARs for the system. Relationship of the objects within the system to factors,

issues, strategies and also the placement of test cases are included; as well as how QARs may be applied to use cases, scenarios and functional requirements. An integrated artefact model architecture will also allow for "trace relationships" which are sometimes overlooked to be clearly defined and established (4).

Artefact models are particularly useful for aligning project goals within the broader goal(s) of an organisation. Symbolic notations are often used in some artefact modelling to illustrate relationships between the objects. Some may be defined using predicate logic language involving the use of symbols, quantifiers and logical operators. For example, the predicate equal(A, B) indicate A = B; plus(A, B) indicate A should be added to B and so on [35]. Using techniques such as predicate logic language notation for artefact modelling can however render the model too complex for those without expert knowledge on the subject [16].

Integrated artefact modelling can be simplified by using standardised object relationship notations commonly used in computer system modelling to reveal how the components of a system may be dependent on each other, guiding requirement specification for the system [18]. To illustrate, an integrated artefact model architecture showing how components within the problem statement for mobile learning is shown in Fig. 2. The model shows when QAWs, QASs and test cases may be required for the system. It also reveals when QAR may be needed to guide against extreme differences in opinion among stakeholders. Use cases will need to be established for testing how well the requirements achieve defined goals as well as the functional / non-functional specifications.

VI. Conclusion

In this paper, the use of RE is proposed to explore relationships between MDT and education. Work presented in this paper is part of an ongoing study involving the application of RE techniques in an HE setting, and this work is yet to be completed. Therefore, a full picture of the requirements and goals for the system are yet to be established. The paper has however shown how RE techniques can be of some considerable benefits when used in systems such as ML, in spite of the very challenging prospect of usage within the complex contexts characteristic of ML systems.

A peripheral question in the wider context not addressed in this paper is how a co-evolution relationship between MDT and education may impact these requirements and goals. Arguably, early promises of a technology are often overshadowed by the "hype" accompanying technological adoption in learning establishments. Some technological systems are eventually found to be either badly managed, unfit for purpose and / or mal-aligned with the broader learning and teaching strategies of the organisation; as noted by Gartner, describing this phenomenon as typical Hype Cycle behaviour. Hype Cycle is the graphical representation of the phases of technological adoption and integration into the marketplace. Early adoption often follows rapidly after a trigger period and Research & Development (R&D. This phase is characterised by scores "inflated expectations" and sometimes ill-judged experimentations. The process continues through periods of disillusionment, enlightenment and productivity in a graphical maturity curve. A likely phase at any time could of course be obsolescence, if the technological system implementation is impracticable or unfit for purpose [36].

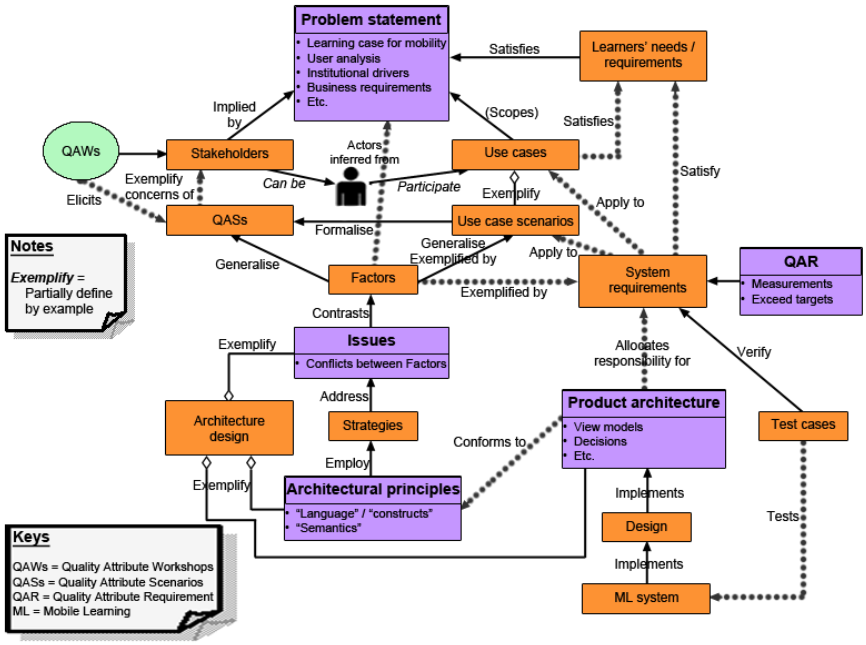


Fig. 2: Integrated artefacts model architecture for Mobile Learning (ML)

Modified from source [18, pp. 130]

Apart from educating a subset of its citizens, two other important functions of HEIs within a nation are "research and innovation", as noted by Professor Paul O'Prey (Vice-Chancellor, University of Roehampton and Chair, Universities UK Longer Term Strategy Network), writing a preface for the 2011/12 UK Higher Education Statistics Agency (HESA) report [37, pp. 2]. It may be a generalised view to state some have found injecting innovation into HE practices including those relating to technological adoption almost impossible. Regardless, and to move the agenda for ML forward, a likely area of further study may be to what extent (if any) HEIs are able to influence products emerging from the R&D phase of technological development and manufacture. Are there any attempts to determine the needs and requirements for learning and advancements and to what extent are these driving the direction and marketisation of advancements? Also related to this track is how quickly and effectively members of educational community innovate and review its learning and teaching strategies / practices to ensure educational technologies such as MDTs are fully integrated; progressing seamless and rapidly towards the productivity phase of the Hype Cycle soon after adoption. Ensuring this occurs will not only save time and effort, but prevent wastage, redundancies and / or inertia that is fast becoming commonplace in many technological adoptions, including MDTs, in learning and teaching.

REFERENCES

[1] J. Fidock and J. Carroll, "Why do users employ the same system in so many different ways?" IEEE Computer Society, pp. 32-39, 2011.

[2] K. Reilly, "Designing research for the emerging field of open development," Information Technologies & International Development, vol. 7(1), pp. 47-60, 2011.

[3] M. Nussbaum and C. Infante, "Guidelines for educational software design that considers the interests and needs of teachers and students," IEEE 13th International Conference on Advanced Learning Technologies, pp. 243-247, 2013.

[4] E. Ackerman, "The Bring-Your-Own-Device dilemma: employees and businesses seek to balance privacy and security," IEEE Spectrum, vol. 50(8), pp. 22, Aug 2013.

[5] D. Jaramillo, N. Katz, B. Bodin, W. Tworek, R. Smart, T. Cook, "Cooperative solutions for Bring Your Own Device (BYOD)," IBM Journal of Research & Development, vol. 57(6), pp. 5:1-5:11, Nov/Dec 2013.

[6] R. G. Lennon, "Changing user attitudes to security in Bring Your Own Device (BYOD) & the cloud," Tier 2 Federation Grid, Cloud & High Performance Computing Science (RO-LCG), 2012 5th Romania, pp. 49-52, 2012.

[7] Department for Work & Pensions (DWP), Digital strategies, pp. 1-32, Dec 2012.

[8] The Cloud, UK leading network infrastructure company JANET partners with The Cloud, [online] Available: http://www.thecloud.net/wifi/news/uk-leading-network-infrastructure-company-janet-partners-with-the-cloud/. [Last accessed: 26 March 2014], 2013.

[9] N. Hopkins, A Sylvester, M. Tate, "Motivations for BYOD: an investigation of the contents of a 21st century school bag," 21st European Conference on Information Systems, pp. 1-12, 2013.

[10] D. Frohberg, C. Göth, G. Schwabe, "Mobile Learning projects – a critical analysis of the state of the art," Journal of Computer Assisted Learning, Blackwell Publishing Ltd, vol. 25, pp. 307-331, 2009.

[11] H. Peng, Y. Su, C. Chou, C. Tsai, "Ubiquitous knowledge construction: mobile learning re-defined and a conceptual framework," Innovations in Education and Teaching International, Routledge Taylor & Francis Group, vol. 46(2), pp. 171-183, 2009.

[12] S. Easterbrook, "What is requirements engineering? (Manuscript in preparation)," [online] Available: http://www.cs.toronto.edu/~sme/papers/2004/FoRE-chapter01-v7.pdf. [Last accessed: 28 February 2014], 2004.

[13] S. Easterbrook and B. Nuseibeh, "Requirements engineering: a roadmap," ICSE 2000 Conference on The Future of Software Engineering, 35-46, 2000.

[14] K. Pohl, "Requirements engineering: fundamentals, principles and techniques," Springer-Verlag Berlin Heidelberg, 2010.

[15] J. K. Ang, S. B. Leong, C. F. Lee, U. K. Yusof , "Requirement engineering techniques in developing expert systems", IEEE Symposium on Computers & Informatics, pp. 640-645, 2011.

[16] A. Ullah and R. Lai, "Modelling business goal for business/IT alignment using requirement engineering," Journal of Computer Information Systems, vol. 7(1), pp. 21-28, 2011.

[17] N. H. Zakaria, A.Haron, S.Sahibuddin, M. Harun, "Requirement engineering critical issues in public sector software project success factor," International Journal of Information and Electronics Engineering, vol. 1(3), pp. 200-209, 2011.

[18] B. Berenbach, D. J. Paulish, J. Kazmeier, A. Rudorfer, Software & systems requirements engineering: in practice, McGraw Hill, 2009, pp. 48-153.

[19] S. Green, "Goal-oriented approaches to requirements engineering," Requirements Engineering, University of the West of England, pp. 1-39, n.d.

[20] The Express, Mobiles face ban in schools, [online] Available: http://www.express.co.uk/news/uk/319250/Mobiles-face-ban-in-schools, [Last accessed: 26 March 2014], 2012.

[21] Oracle Communications, Opportunity Calling: The Future of Mobile Communications – Take Two, Oracle Communications, [online] Available: http://www.oracle.com/us/industries/communications/oracle-communications-future-mobile-521589.pdf. [Last accessed: 6 March 2013], 2011.

[22] Jones & Bartlett Learning, "Quantitative versus qualitative research, or both?" [online] Available: http://samples.jbpub.com/9780763780586/80586_CH03_Keele.pdf. [Last accessed: 28 February 2014], n.d.

[23] A. MacDiarmid and P. Lindsay, "Can system of systems be given self-x requirement engineering capabilities?" Systems Engineering and Test and Evaluation Conference 2010 (SETE 2010), Systems Engineering Society of Australia, 1-15, 2010.

[24] J. Mylopoulos and L. Chung, et al, "From object-oriented to goal-oriented requirements analysis", Communications of the ACM, 42(1), pp. 31-37, 1999.

[25] A. Van Lamsweerde, "Goal-oriented requirements engineering: a guided tour," *RE'01, 5th IEEE International Symposium on* Requirements Engineering, Toronto, pp. 249-263, Aug 2001.

[26] A. G. Sutcliffe, "Requirements engineering," in Soegaard, Mads and Dam, Rikke Friis (eds.), "The Encyclopedia of Human-Computer Interaction, 2nd Ed.," Aarhus, Denmark: The Interaction Design Foundation. [online] Available: at http://www.interaction-design.org/encyclopedia/requirements_engineering.html. [Last accessed: 28 February 2014], 2013.

[27] S. Kausar, S. Tariq, S. Riaz, A. Khanum, "Guidelines for the selection of elicitation techniques," 6th International Conference on Emerging Technologies (ICET), pp. 265-269, 2010.

[28] A. Batool, Y. H. Motla, B. Hamid, S. Asghar, M. Riaz, M. Muktar, M. Ahmed, "Comparative study of traditional requirement engineering and agile requirement engineering," Advanced Communication Technology (ICACT) 15th International Conference, pp. 1006-1014, 2013.

[29] S. Liaskos, S. A. McIlraith, S. Sohrabi, J. Mylopoulos, "Guidelines for the selection of elicitation techniques," 18th IEEE International Requirements Engineering Conference, pp. 135-144, 2010.

[30] P. Donzelli and P. Bresciani, "Goal-oriented requirements engineering: a case study in e-government," Springer-Verlag Berlin Heidelberg, pp. 601-616, 2003.

[31] E. Kavakli and P. Loucopoulos, "Goal modelling in requirements engineering: analysis and critique of current methods," Information

Modeling Methods and Methodologies: Advanced Topics in Database Research, pp. 102-124, 2005.

[32] X. Zheng, X. Liu, S. Liu, "Use case and non-functional scenario template-based approach to identify aspects, *Second International Conference on Computer Engineering and Applications,* pp. 89-93, 2010.

[33] A. Dardenne, A. Van Lamsweerde, et al, "Goal-directed requirements acquisition." *Science of Computer Programming,* vol. 20(1-2), pp. 3-50, 1993.

[34] University of East London, [online] Available: http://www.uel.ac.uk/planning/, [Last accessed: 26 March 2014], n.d.

[35] S. Lakmazaheri and W. J. Rasdorf, "An artifact modelling approach for developing integrated engineering systems." [online] Available: http://etd.lib.ncsu.edu/publications/bitstream/1840.2/652/1/95+an+artifact+modeling+approach+for+developing+ [Last accessed: 18 June 2014], n.d.

[36] Gartner, Hype cycles, [online] Available: http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp. [Last accessed: 11 June 2014], n.d.

[37] Universities UK in collaboration with Higher Education Statistics Agency (HESA), "Higher Education: A diverse and changing sector: Patterns and trends in UK Higher Education." [online] Available: http://www.universitiesuk.ac.uk/highereducation/Documents/2013/PatternsAndTrendsinUKHigherEducation2013.pdf. [Last accessed: 11 June 2014], pp. 2, 2013.

Emerging Aspects in Information Security

ADMITTEDLY, information security works as a backbone for protecting both user data and electronic transactions. Protecting the communication and data infrastructure of an increasingly inter-connected world has become vital nowadays. Security has emerged as an important scientific discipline whose many multifaceted complexities deserve the attention and synergy of the computer science, engineering, and information systems communities. Information security has some well-founded technical research directions which encompass access level (user authentication and authorization), protocol security, software security, and data cryptography. Moreover, some other emerging topics related to organizational security aspects have appeared beyond the long-standing research directions.

The Emerging Aspects in Information Security (EAIS'14) workshop focuses on the diversity of the information security developments and deployments in order to highlight the most recent challenges and report the most recent researches. The workshop is an umbrella for all information security technical aspects. In addition, it goes beyond the technicalities and covers some emerging topics like social and organizational security research directions. EAIS'14 is intended to attract researchers and practitioners from academia and industry, and provides an international discussion forum in order to share their experiences and their ideas concerning emerging aspects in information security met in different application domains. This opens doors for highlighting unknown research directions and tackling modern research challenges. The objectives of the EAIS'14 workshop can be summarized as follows:

- To review and conclude researches in information security and other security domains, focused on the protection of different kinds of assets and processes, and to identify approaches that may be useful in the application domains of information security
- To find synergy between different approaches, allowing to elaborate integrated security solutions, e.g. integrate different risk-based management systems
- To exchange security-related knowledge and experience between experts to improve existing methods and tools and adopt them to new application areas
- To present latest security challenges, especially with respect to EC Horizon 2020

Topics

Topics of interest include but are not limited to:

- Biometric technologies
- Human factor in security
- Cryptography and cryptanalysis
- Critical infrastructure protection
- Hardware-oriented information security
- Social theories in information security
- Organization- related information security
- Pedagogical approaches for information security
- Individual identification and privacy protection
- Information security and business continuity management
- Decision support systems for information security
- Digital right management and data protection
- Cyber and physical security infrastructures
- Risk assessment and risk management in different application domains
- Tools supporting security management and development
- Emerging technologies and applications
- Digital forensics and crime science
- Misuse and intrusion detection
- Security knowledge management
- Data hide and watermarking
- Cloud and big data security
- Computer network security
- Security and safety
- Assurance methods
- Security statistics

Event Chairs

Awad, Ali Ismail, Luleå University of Technology, Sweden

Bialas, Andrzej, Institute of Innovative Technologies EMAG, Poland

Program Committee

Banerjee, Soumya, Birla Institute of Technology

Bun, Rostyslav, Lviv Polytechnic National University

Clarke, Nathan, Plymouth University, United Kingdom

Cyra, Łukasz, European Commission - Joint Research Centre Institute for the Protection & Security of the Citizen

Dworzecki, Jacek, Police Academy in Szczytno

Fernandez, Eduardo B., Florida Atlantic University, United States

Furnell, Steven, Plymouth University, United Kingdom

Furtak, Janusz, Military University of Technology, Poland

Geiger, Gebhard, Technical University of Munich, Faculty of Economics

Grzenda, Maciej, Orange Labs Poland and Warsaw University of Technology, Poland

Hämmerli, Bernhard M., Hochschule für Technik+Architektur (HTA), Switzerland

Harnesk, Dan, Luleå University of Technology

Hasssaballah, M., South Valley University, Egypt

Kalbarczyk, Zbigniew, University of Illinois at Urbana-Champaign

Kapczynski, Adrian, Silesian University of Technology, Poland

Klamka, Jerzy, Polish Academy of Sciences

Kosmowski, Kazimierz, Gdansk University of Technology

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 863–870

DOI: 10.15439/2014F38
ACSIS, Vol. 2

Enterprise-oriented Cybersecurity Management

Tomasz Chmielecki, Piotr Chołda, Piotr Pacyna, Paweł Potrawka, Norbert Rapacz, Rafał Stankiewicz, and Piotr Wydrych
AGH University of Science and Technology, Kraków, Poland
Department of Telecommunications
al. Mickiewicza 30, 30-059 Kraków, Poland
Email: pacyna@kt.agh.edu.pl

***Abstract*—Information technology is widely used in processes vital to enterprises. Therefore, IT systems must meet at least the same level of security as required from the business processes supported by these systems. In this paper, we present a view on cybersecurity management as an enterprise-centered process, and we advocate the use of enterprise architecture in security management. Activities such as risk assessment, selection of security controls, as well as their deployment and monitoring should be carried out as a part of enterprise architecture activity. A set of useful frameworks and tools is presented and discussed.**

I. Introduction

CYBERSECURITY has been recognized as a business concern and declared an enterprise-wide activity. There is a growing understanding that cybersecurity requirements for the confidentiality, integrity and availability of services provided by the IT infrastructure in an enterprise must be elevated to the same, or higher, level, as the security requirements for the elements of the enterprise that deliver a business function. In consequence, cybersecurity should not be associated with IT technology alone and should no longer be regarded as purely an IT domain. In essence, IT departments are not able to conduct proper risk assessment and mitigation on their own. The information necessary to conduct risk analysis properly is available to business management. When decisions and actions are taken in a process in which IT and business management work together to assess risks and determine priorities in risk mitigation, we can speak about *enterprise-oriented cybersecurity management*. Current practice shows, however, that cybersecurity is still based on technical rules of thumb. The use of formalized methodologies like risk management is not common. The perception of business goals in the process is fragmentary; so many aspects are omitted in cybersecurity. In consequence, the process is incomplete. In this paper, we promote the usage of enterprise architecture-based tools and methodologies to deal with cybersecurity in enterprises which rely on IT infrastructures to deliver products and services.

The proposed approach calls for a paradigm shift in cybersecurity. It requires management personnel to share essential data with IT people to enable business impact analysis and to rely on outcomes that define security priorities. The knowledge of risks in IT departments (likelihood and impact of various threats) and countermeasures should complement the knowledge in business departments. A common workspace for business and IT is an enterprise architecture. It enables collaboration, owing to an improved awareness of the business processes that support the company's mission on one side, and their realization through operational activities, supported by IT, on the other side. The decisions pertaining to security are based on a proper assessment of vulnerabilities and threats and provide options for a response (e.g., continuity and recovery plans, security controls).

Enterprise-oriented cybersecurity management is not a state but a persistent process, with the ability to adapt continuously to a changing environment. Cybersecurity must not be considered an isolated activity—merely a domain-specific precaution against isolated hacking or sabotage activity. Attackers will tend to affect business by targeting general, enterprise-level goals by impairing applications and supporting infrastructure (e.g. platform systems). A vulnerability at one level impacts other levels. Consequently, loss expectancy tends to magnify through cross-layer dependencies. To understand vulnerabilities, risks need to be studied and evaluated top-down: from business principles, through business objectives, and business functions, down to security controls, and also bottom-up for traceability and evaluation. Such analysis is enabled by a thorough description of the enterprise architecture along with an aligned *risk assessment*. Afterwards, the main goal of the *risk response* is to select countermeasures dealing with the risks recognized. The effects of the deployed measures are *continuously monitored*. The enterprise architecture should also drive transition with change management, including major upgrades in security policies and their implementations. One of the critical methods for achieving the goal is risk management [1]–[3]. This should employ enterprise architecture as a valuable source of information about the enterprise. While this may seem engaging too much overhead and may seem counterproductive, even the first exercise will provide value in a reasonable time. In the course of the paper we discuss a collection of tools (e.g., frameworks and software applications) supporting change or risk management.

In our paper, we elaborate on the pillars fundamental to organizing cybersecurity management (enterprise architecture, threat meta-models, risk assessment and response, risk moni-

This scientific research was partially financed by the Polish Ministry of Science and Higher Education from the research budget for 2013-2015, Project No. IP2012 022972 and partially supported by the Polish Ministry of Science and Higher Education under Grant O R00 0119 12. Part of this work was also funded by the Polish Ministry of Science and Higher Education under project 1310/7.PR UE/2010/7.

978-83-60810-58-3/$25.00

TABLE I
SELECTED ENTERPRISE ARCHITECTURE FRAMEWORKS

Framework	Context	Description	Advantages	Drawbacks
TOGAF 9.1 (2012)	Open, universal	Provides a process lifecycle to build and manage architecture transitions within an enterprise—Architecture Development Method (ADM) and a set of models.	• ADM is the central point • Ensures a controlled environment for change • Substantially aimed at transitional architectures	Lack of precise model guidance (Archimate 2.0 fills that gap)
DoDAF 2.0 (2009)	Military	Defines a set of views and models for visualizing the complexities in an architecture description and reasoning for various stakeholders. The architecture data gathered becomes central, and the data schemes provided define its structure. There is no obligatory method of development	• Provides data schemes and a precise meta-model • Aimed at transitional architectures • Supports SOA • Tailored for large and complex systems	• No single obligatory method of development • Military-oriented • Limited support for non-functional requirements (like cybersecurity)
The Zachman Framework (2008)	Business	Is best described as a scheme or taxonomy of EA. It classifies views based on six interrogative questions (why, how, what, who, where, when) and five abstraction layers (contextual, conceptual, logical, physical, detailed). No methodology is defined for developing an architecture	• Compact and easy to follow • Well defined viewpoints	• No methodology for building EA • No transitional architectures • Limited support for non-functional requirements (like cybersecurity)

toring) and then summarize how they are integrated. Section II introduces enterprise architectures. Section III deals with the main processes in cybersecurity provisioning, that is, risk management. Section IV summarizes the ideas presented in a unified view. Afterwards, we shortly conclude.

II. ENTERPRISE ARCHITECTURE

Enterprise architecture (EA) is used for the description of complex enterprises. The description includes business processes and their mapping to operational activities for the key processes. It serves as a blueprint for the enterprise structure and operations. Enterprise architecture is a set of models that depict how various business and technical elements work together [4]. Along with ontologies or meta-models, it describes the terminology, the composition of enterprise components, and their relationships with the surrounding environment, as well as the guiding principles for eliciting requirements, design and evolution. The enterprise architecture frameworks (see examples in Table I) are templates for development of instances of EA. A set of languages used to describe the enterprise architectures has been developed and a few of the popular options are sketched in Table II.

A. Role of EA in Security Management

Technically speaking, cybersecurity activity is about establishing a linkage between secured objects and vulnerabilities, threats and countermeasures, as well as monitoring them. Risk is the perception of a relation between these and business objectives. A balance is required between these elements for three essential, interdependent objectives: confidentiality, integrity, and availability. The enterprise approach to cybersecurity requires that risk management should be carried out simultaneously at the business, application, data and technology layers, and combined. Business impact analysis, as a basic step in risk assessment, and business continuity planning, the main concern of risk response, requires precise data about the enterprise. Such knowledge should embrace at least simplified principles defining the enterprise's mission and the manner in which this is accomplished.

Security management should be organized as a process of continuous improvement. Activities such as, for example, risk monitoring, risk assessment, selection of security controls and their deployment need to be carried out repeatedly. Short iterations lend themselves to rapid response to risks that require prompt response.

The security management process causes modifications to the enterprise. These changes can be considerable. As such, they should be staged in transitions describing the change of enterprise architecture.

B. Sample Case Study of EA

To illustrate various EA-related aspects, we have developed a sample view of EA presented in Fig. 1. It shows an architecture for an IT infrastructure supporting a gas transportation process using a networked SCADA control system. As can be seen on the right, EA describes the structure of enterprise organization, business processes, applications and technology that allow the enterprise's goals to be achieved. The notation uses the Archimate 2.0 language, which allows for linking the elements of the architecture together and tracing the relationships among elements. Here, the main business process is gas transportation. This is supported by four subprocesses at the application layer (agreement management, etc.). Those subprocesses are supported by software applications

TABLE II
SELECTED ARCHITECTURE DESCRIPTION LANGUAGES

Framework	Context	Description	Advantages	Drawbacks
Archimate 2.0 (2012)	Enterprise-oriented	Archimate is an architecture description language. The main part covers business, application, and technology layers. There are two extensions: motivation and implementation which makes it compatible with the TOGAF framework. Archimate defines multiple views, but it is possible to define other views, too	• Allows for modeling dependencies • In line with the newest version of TOGAF	• Suitable only for modeling on the enterprise level, lower levels need another notation (like BPMN) • Thus far, a limited set of the supporting software tools
UML 2.1.4 (2013)	Software	UML is a universal language, but is usually perceived as software-oriented and is used for the solution architecture description	Wide modeling software support	Seldom used for business purposes
BPMN 2.0 (2011)	Business	Standard for business process modeling. Provides a graphical notation and model elements focused on business processes and roles. Flow diagrams are similar to UML activity diagrams	Widely used in business analytics	Not possible to map business processes to applications or technologies
UPDM 2.1 (2013)	Military	UML profiles and graphic notation supporting the models and views taken from DoDAF framework	Full enterprise architecture support	Military-oriented
SysML1.2 (2010)	System engineering	Extension of a subset of UML	• Compact set of diagrams • System-of-systems support	No relationships modeled to business

(like CRM system) and cyberinfrastructure (file management system, databases, etc.). After adding security knowledge, it becomes possible, for instance, to trace the impact of a file server fault (induced by DDoS attacks) on two systems: CRM and capacity planning, which as a consequence influence (via the information service) the SCADA control system and impair business processes. A real EA will contain much more information for use by stakeholders (like clients, owner, or governmental administration) and formulated with multiple views. The data can be stored in a repository, where a formal representation of the structure along with the related threat models enables reasoning and reporting on the likelihood or impact of various incidents, thus supporting risk assessment.

C. Vulnerability and Threat Meta-Model

Cybersecurity management requires deep knowledge of vulnerabilities and threats. This knowledge is maintained in respective databases and needs to be incorporated intothe enterprise architecture. To make this possible, efficient meta-model of cybersecurity-related data is necessary. This introduces a vocabulary, syntax, and constraints as well as enables cybersecurity modeling. The enterprise architecture description is enriched by risk assessment with contextual information on cybersecurity issues.

A fragment of an example cybersecurity meta-model is shown in Fig. 2 (see for instance [7] for an alternative model). *Secured objects* span many categories: humans, physical resources, and immaterial assets. All in all, these fall into two classes, being an *asset* or a *process*. They have their own *security attributes* (like a predefined value of availability, for instance). *Vulnerabilities* are attached to security objects during risk assessment. Vulnerabilities will manifest as incidents in the event of a *threat* materialization, which will exploit them. *Risk* is a measure of *likelihood* and *impact* of threat realizations. After the vulnerabilities and threats are identified, it is possible to produce countermeasures using *security controls*, which are a technique for risk response. A control can be accomplished with an organizational procedure (like authentication enforcement) or with an asset protecting other assets (e.g. IPS/IDS systems) or a combination of the two.

III. RISK MANAGEMENT

As a formalized process, risk management aims at dealing with all the threats and related countermeasures in a cyclic manner as shown in Fig. 3. Risk serves as an explicit interface between the business and IT. The following three aspects are taken into account during risk assessment: *exposure* of a secured object to selected threats; and two quantifiable aspects—the *likelihood* of those events, and the *impact* on the enterprise, if they occur. While threat analysis and likelihood evaluation are evaluated by IT experts, the evaluation of impact on business processes is of a non-technical character only, related to financial measures (for instance, penalties for outages), or public safety and liability issues. The business impact is assessed either in qualitative terms (high-medium-low), or preferably in quantitative ways, as this allows for finding a risk response based on optimization methods. Risk assessment has been studied for a long time and commercial frameworks to perform it are also present [8], see Table III. Typically, frameworks suggest what should be done, but not exactly how to carry it out.

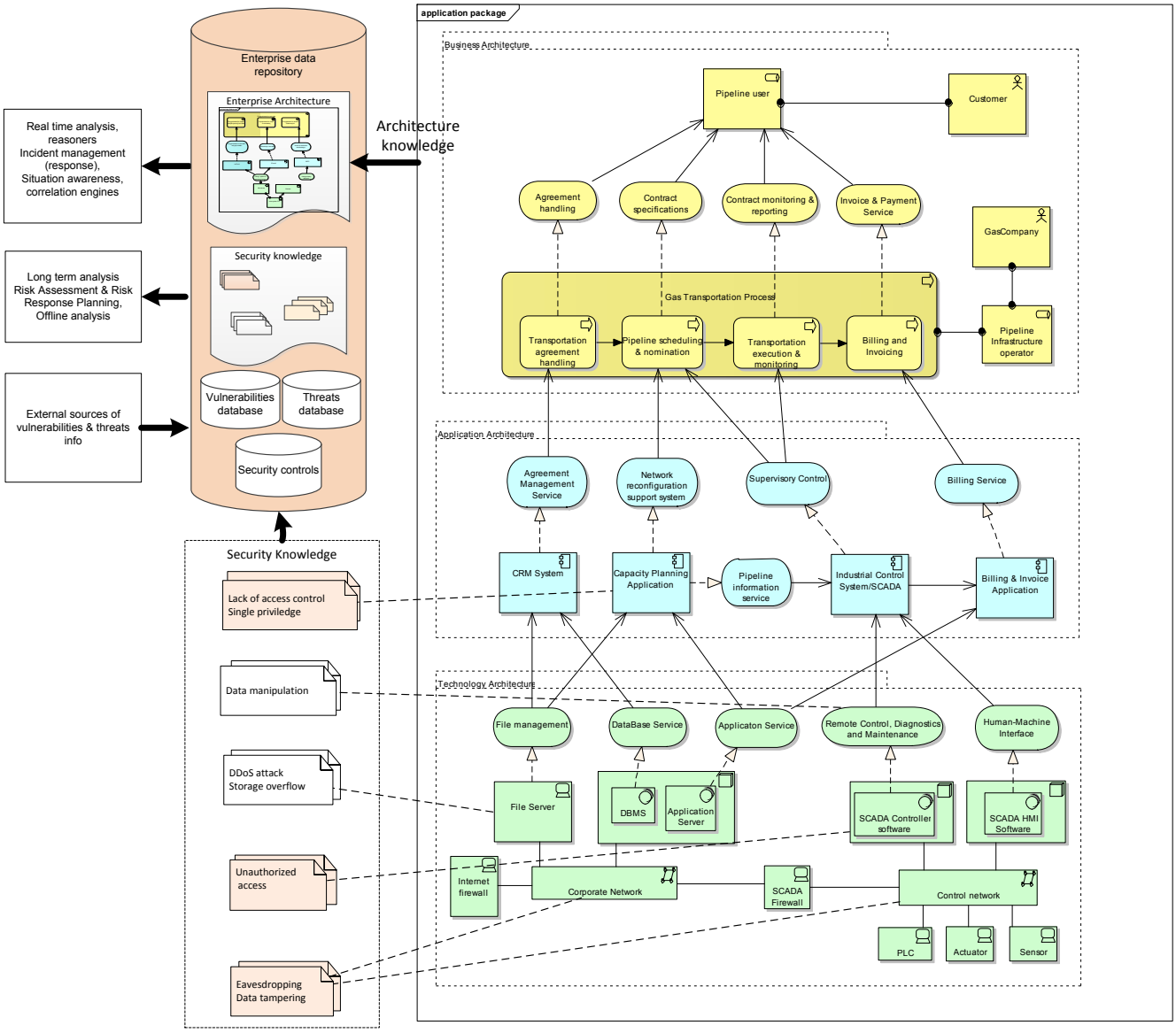


Fig. 1. Role of an example instance of an enterprise architecture in cybersecurity management [5], [6].

Risk management has become overwhelming in information technology as it covers a broad range of issues, as shown in Fig. 4. Its usage in the context of network resilience against attacks is covered in [10] and against random failures in [2]. Sometimes, it is even postulated to engage the end user in this [11], despite some concerns: no clear goals from customers, a low level of considering their actions, a lack of interest in security, a pure lack of technical expertise, or even a slowdown in the adoption of new technologies. Enterprises have better knowledge of their goals, actions and technology to be able to effectively combine the data provided by the enterprise architecture and use it with risk management techniques to improve its operations.

A. *Risk Assessment*

Risk assessment analyzes the enterprise operation from various domain viewpoints: public safety (against threats of massive human injuries); business logic (like checking for process deadlocks); IT cybersecurity in relation to a specific industry field (e.g. SCADA concerns in oil transportation systems); The system-of-systems analysis encompasses methodologies for analyzing multi-scale, interconnected and interdependent systems with emergent behaviors [12]. The following three types of failures are characteristic of interdependent infrastructures [13], but can also be observed in Fig. 1.

- Cascading: when a failure in one infrastructure causes the failure of other infrastructures (note the propagation of technology failures all the way up to business process

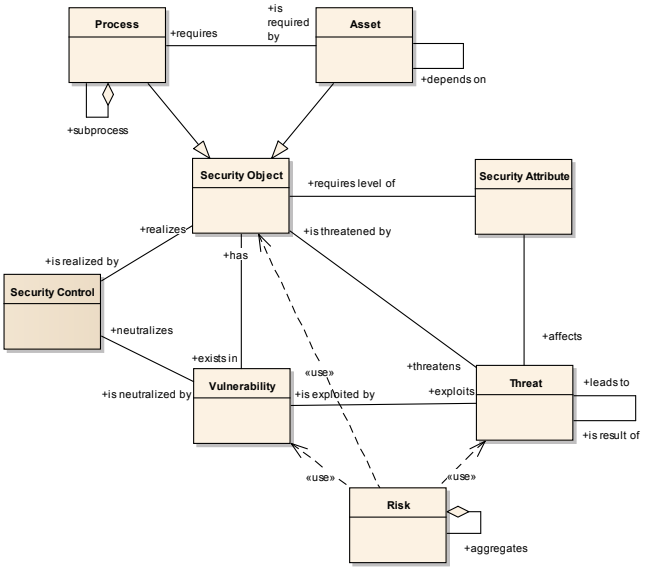


Fig. 2. Cybersecurity meta-model [5], [6].

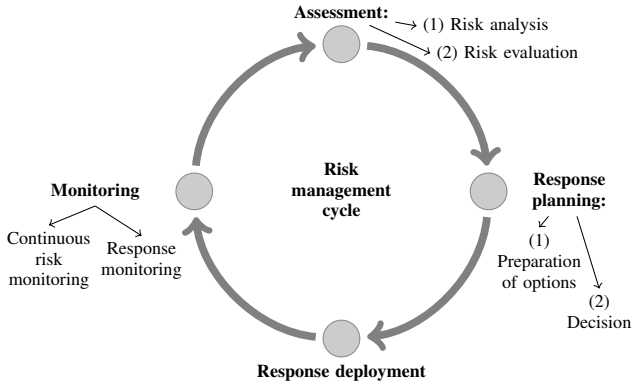


Fig. 3. Simplified risk management cycle [9].

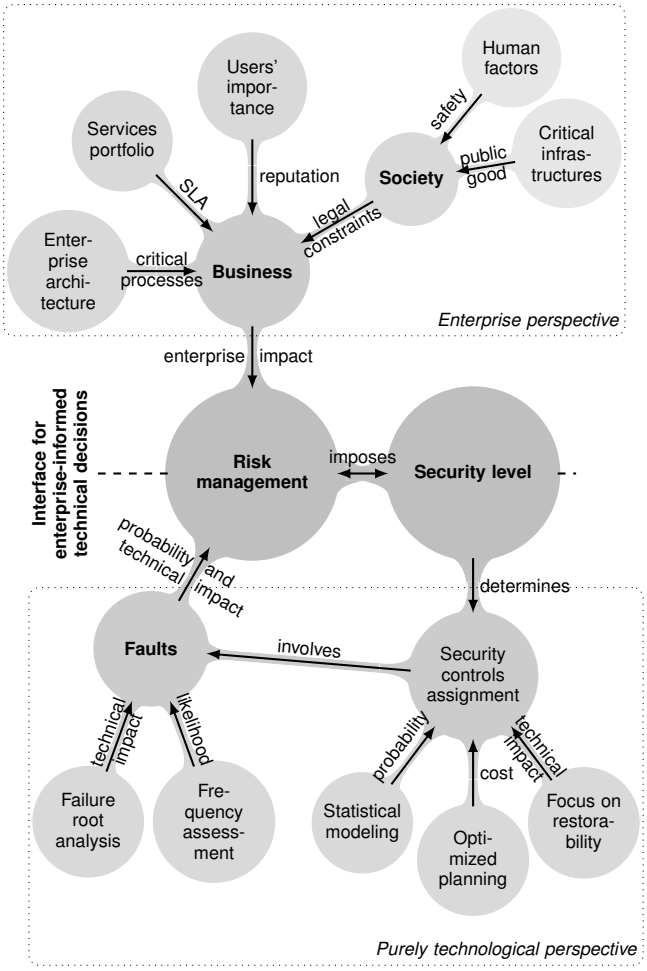


Fig. 4. Different aspects of risk management.

discontinuity).

- Escalating: when an existing failure in one infrastructure exacerbates an independent failure in another infrastructure, increasing its impact (see the earlier example of a file system failure in combination with SCADA control faults).
- Common cause: when two or more infrastructures are affected simultaneously because of a common event (see the destructive impact of corporate network outages on all subprocesses).

Basically, risk assessment consists of risk analysis (identifying vulnerabilities, threats, and related risks) and risk evaluation, determining their probability and impact on the business goals. Although risk can be assessed qualitatively, where probability and impact are assessed using ordinal scales (e.g. small-medium-high) and their combinations, a more sophisticated approach, known as probabilistic risk assessment (PRA), is more meaningful. It is a strictly quantitative approach during which both impact and probability are assessed and expressed mathematically [14]. In the best case scenario, the full probability distribution function (PDF) of the impact expressed in monetary units can be found. In this case, business-related measures like Value-at-Risk (VaR) can be applied. These are easily understood in the investing sector, and therefore can be useful in communicating risk to management. Although such measures were invented for the banking sector to assess the obligatory level of savings, it is suggested that they be used in the telecommunications sector to assess the level of cybersecurity-related investments in network design [2], [10], [15], [16]. The usage of such metrics is especially useful as there is a large toolbox of optimization methods elaborated in the modern portfolio theory, for which VaR is the basic quantitative risk measure and can be used during risk response planning.

B. Risk Response

After analyzing threats and evaluating the related risks, it is necessary to prepare a risk response proposal to be decided and accepted by business management. One of the

TABLE III
SELECTED FRAMEWORKS SUPPORTING RISK MANAGEMENT FOR IT CYBERSECURITY

Framework	Scope	Advantages	Drawbacks
COBIT 5.0, **Risk IT**, **Val IT** (2012)	IT governance (COBIT) combines a business perspective and IT control model approach. Risk IT focuses on IT-enabled risk management and Val IT covers financial IT governance	• Emphasizes relationships between business and IT processes • Includes aspects of control, risk, cost efficiency and maturity • Compatible with audit procedures • Uses RACI (Responsible-Accountable-Consulted-Informed) charts presenting a detailed allocation of responsibilities	• Lack of technical details and low level practices • No description of methods to transition
SABSA (2009)	Framework for the development of a security architecture in an enterprise	• Intuitive and understandable distribution of layers • Well planned and described risk management processes and their succession, interfaces and attributes	• Lack of coverage of all aspects of IT cybersecurity at an equal detail level • The concepts covered without the required explanation, which makes it difficult to properly implement
ITIL 3.0, **M_o_R** (2011)	A set of practices for IT service management, combining IT services with a business perspective	• Popular and widely used description language • Recommendations based on best practices • IT service management considered in a systematic and consistent manner	• Expensive to implement • Long time to implement correctly • Neither generic nor self-sufficient, should be combined with another risk management framework
CC-ISO 15408 ver. 3.1 (2009)	International technical standard for IT cybersecurity certification of products related to IT	• Facilitates risk assessment in relation to particular assets (systems, applications, devices) • Defines different levels of cybersecurity and quality requirements	• Expensive to implement • Used mainly at the development stage • Does not support a holistic approach to the organization, but focuses only on the evaluation of a particular resource or product

most important parts of the risk response is to ensure continuity in the business process operation. This is performed by *business continuity planning* and *disaster recovery* [17], where continuity may be defined as a state in which a system is operational again after disruption at a well-defined level after a certain time, bounded by the maximum tolerable downtime parameter. While it is possible that this goal can be realized by various methods, scenarios are prepared. Each scenario should contain a set of countermeasures (*security controls*), their manner of implementation, the resulting risk change and the cost involved. The first decision is how to deal with recognized risks. Typical decisions that are relevant in the technical context are as follows: *acceptance* when nothing is done about the recognized risks (no changes in comparison to the actual state are necessary); *avoidance* of situations where threats take place (elimination of a problematic information system with many vulnerabilities); *reduction* of the likelihood (addition of a firewall decreasing the number of successful attacks); *mitigation* of the impact, the most popular decision (encryption of data so that it cannot be used even if stolen). Three most common strategies apply to mitigation [18], [19], see Fig. 5:

- *Risk minimization*: choice of the minimum impact possible; can be very costly but might be the first choice especially in critical infrastructures, where the public good is most important.
- *Total (benefit) coverage*: a strategy where the cost of scenario implementation is balanced with the reduction in impact; the strategy is imposed by NIST series of recommendations [20] for US federal institutions.
- *Profit maximization*: a choice of scenario where the marginal impact reduction is balanced by the marginal cost of this reduction, ensuring minimization of the total cost of incidents and risk mitigation.

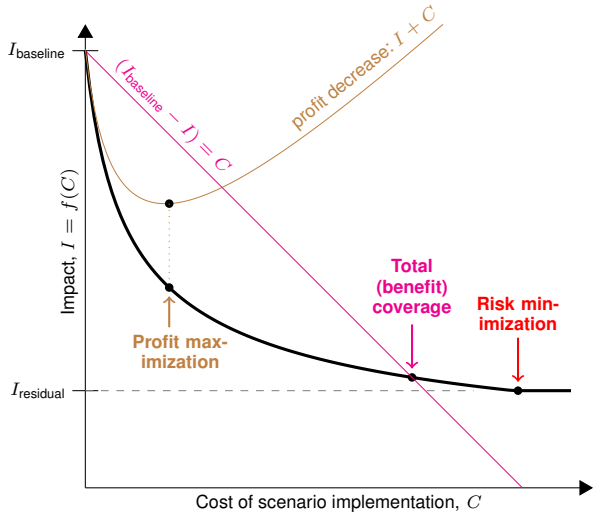


Fig. 5. Illustration of the three basic risk mitigation strategies.

C. Examples of Security Controls

Security controls are countermeasures aimed at avoiding, reducing, or mitigating risk. Two popular catalogues of security controls are ISO 27002 [21] and NIST 800-53 [22] (see Fig. 6). ISO 27002 provides 176 cybersecurity controls organized into twelve sections. The controls contain well defined objectives as well as implementation guidance. Its broad scope makes it applicable to any industry and business of arbitrary size. NIST SP 800-53 groups 317 controls into 18 families organized under three classes:

- *Technical* (e.g. AC-11: session locks).
- *Organizational*: for managing information cybersecurity programs (e.g. MP: media marking).
- *Management*: pertaining to business governance (e.g. PM-10: cybersecurity authorization process).

D. Continuous Monitoring

The main goal of the continuous monitoring process is to maintain up-to-date knowledge about the effectiveness of the risk response scenarios implemented in the enterprise. The monitoring has the following goals:

- to deliver the information about the state of the processes and assets to appropriately assess the current risk;
- to discover changes in processes and assets state that may influence the level of security and effectiveness of the implemented security controls, resulting in new threats;
- to recognize cybersecurity incidents.

Continuous monitoring is responsible for ensuring consistency between the implemented security controls and standards, recommendations and regulations. Collection of cybersecurity related data is a discrete process triggered by incidents, changes, etc. Some of the monitoring tasks can be called repeatedly or on schedule. However, monitoring is a continuous process.

IV. Consolidated View on Enterprise-oriented Cybersecurity Development

So far, we have defined various elements of an enterprise-oriented approach to security deployment. Here, we show how they are integrated. The cybersecurity management process will consist of at least four repetitive steps: risk assessment (adjust), risk response (plan), implementation of security controls (do), and continuous monitoring (check).

Each of the four phases of the cybersecurity management process consists of several tasks. The scope, granularity and time frame depend on the enterprise. The activities of the process are carried out with a focus on various aspects pertaining to different layers of the EA: business, application, or technology. Security principles, requirements, goals and constraints are thus formulated at various levels of enterprise description. The implementation process should be coordinated at various levels, in accordance with the four steps. The scope is tailored to the enterprise's needs, priority and the available level of funding. The activities may have different durations, but they complement each other. For example, an implementation of a security control protecting a server against a specific attack at the technology level supports a security implementation process at the application level dealing with the classification of confidential information which, in turn, protects a well defined business goal (e.g. compliance with the regulations on personal data security).

The processes organizing the cybersecurity management cycle operate on various time scales. The incidents require a rapid response. Also, new vulnerabilities should be addressed without delay. In such cases, primarily risk assessment, response and implementation of security controls must be performed rapidly. Then, these are based on common IT cybersecurity practices, not always optimal from the cost viewpoint. Immediate solutions are called 'quick wins.' On the other hand, a security implementation related to a new project run in the enterprise, the deployment of new assets, or the creation of novel operational processes result in triggering a long-term process. Each phase will then require very careful analyses and involve much more time.

The EA transition process should be closely related to the security implementation process. While defining current state and intermediate state (transitional) enterprise architectures, all recognized assets and processes will require risk assessment and the preparation of a risk response. The security controls should be employed together with the implementation of the new enterprise architecture.

A new instance of security implementation process may be triggered in various cases:

- Continuous monitoring recognizes that an implemented security control has become ineffective or inadequate (e.g. due to a change in the surrounding environment).
- A new vulnerability in an asset or a new threat exploiting a recognized vulnerability has been announced.
- New assets have been deployed in the enterprise: all dependent systems must at least be assessed from the risk viewpoint.
- The process of enterprise architecture transition has started (e.g. due to a business management decision).

V. Conclusions

Cybersecurity is crucial to the contemporary enterprise. We describe a business view of cybersecurity by showing recognized frameworks known thus far in enterprise governance. Enterprise architecture frameworks allow the development of an EA, which is crucial to properly address risk, but differ in the extent to which they guide through the cybersecurity aspects. Given the vast number of incidents, machine-assisted decision support becomes a decisive factor and this is the main issue to be solved in the future.

References

[1] M. E. Johnson *et al.*, "Security through Information Risk Management," *IEEE Security & Privacy*, vol. 7, no. 3, pp. 45–52, May/Jun. 2009. [Online]. Available: http://dx.doi.org/10.1109/MSP.2009.77

[2] P. Chołda *et al.*, "Towards Risk-aware Communications Networking," *Reliability Engineering & System Safety*, vol. 109, pp. 160–174, Jan. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.ress.2012.08.009

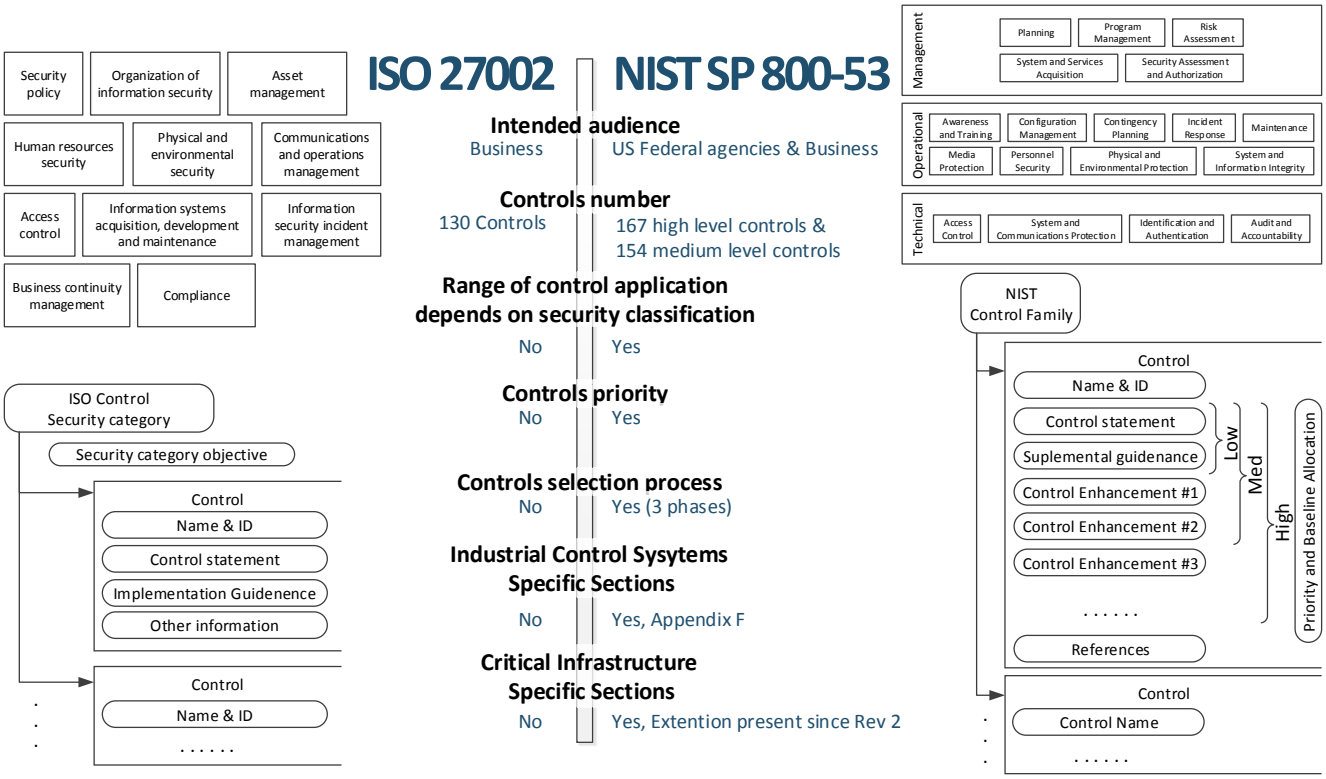


Fig. 6. Comparison of security controls defined by ISO and NIST.

[3] P. Pacyna *et al.*, "Założenia i cele metodyki OKIT do wdrażania systemu bezpieczeństwa teleinformacyjnego w infrastrukturach krytycznych," in *Nowoczesne systemy łączności i transmisji danych na rzecz bezpieczeństwa. Szanse i zagrożenia*, A. R. Pach *et al.*, Eds. Warszawa, Poland: Wolters Kluwer Polska SA, 2013, pp. 442–457, (in Polish).

[4] J. Barateiro *et al.*, "Manage Risks through the Enterprise Architecture," in *Proc. 45rd Hawaii International Conference on System Sciences HICSS-45*, Grand Wailea, Maui, HI, Jan. 4-7, 2012. [Online]. Available: http://dx.doi.org/10.1109/HICSS.2012.419

[5] N. Rapacz *et al.*, "Elementy skutecznego zarządzania bezpieczeństwem w przedsiębiorstwach obsługujących infrastruktury krytyczne," in *Nowoczesne systemy łączności i transmisji danych na rzecz bezpieczeństwa. Szanse i zagrożenia*, A. R. Pach *et al.*, Eds. Warszawa, Poland: Wolters Kluwer Polska SA, 2013, pp. 458–475, (in Polish).

[6] P. Pacyna *et al.*, *OKIT. Metodyka ochrony teleinformacyjnych infrastruktur krytycznych*. Warszawa, Poland: Wydawnictwo Naukowe PWN, 2013, (in Polish).

[7] S. Fenz *et al.*, "Information Security Risk Management: In Which Security Solutions Is It Worth Investing?" *Communications of the Association for Information Systems*, vol. 28, no. 22, pp. 329–356, May 2011.

[8] J. Araujo Wickboldt *et al.*, "A Framework for Risk Assessment based on Analysis of Historical Information of Workflow Execution in IT Systems," *Computer Networks*, vol. 55, no. 13, pp. 2954–2975, Sep. 15, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2011.05.025

[9] P. Chołda and B. E. Helvik, "Reliable Network-based Services," *Computer Communications*, vol. 36, no. 6, pp. 607–610, Mar. 15, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.comcom.2013.01.003

[10] T. Ackermann, *IT Security Risk Management. Perceived IT Security Risks in the Context of Cloud Computing*. Wiesbaden, Germany: Springer Fachmedien, 2013.

[11] A. van Cleeff, "A Risk Management Process for Consumers: The Next Step in Information Security," in *Proc. New Security Paradigms Workshop NSPW'10*, Concord, MA, Sep. 21-23, 2010. [Online]. Available: http://dx.doi.org/10.1145/1900546.1900561

[12] Y. Y. Haimes, "Models for Risk Management of Systems of Systems," *International Journal of System of Systems Engineering*, vol. 1, no. 1/2, pp. 222–236, 2008. [Online]. Available: http://dx.doi.org/10.1504/IJSSE.2008.018138

[13] S. M. Rinaldi *et al.*, "Identifying, Understanding, and Analyzing Critical Infrastructure Interdependencies," *IEEE Control Systems Magazine*, vol. 21, no. 6, pp. 11–25, Dec. 2001. [Online]. Available: http://dx.doi.org/10.1109/37.969131

[14] M. Todinov, *Risk-Based Reliability Analysis and Generic Principles for Risk Reduction*. Amsterdam, The Netherlands: Elsevier Science & Technology Books, 2006.

[15] L. Mastroeni and M. Naldi, "Violation of Service Availability Targets in Service Level Agreements," in *Proc. Federated Conference on Computer Science and Information Systems FedCSIS 2011*, Szczecin, Poland, Sep. 18-21, 2011.

[16] A. J. Gonzalez and B. E. Helvik, "SLA Success Probability Assessment in Networks with Correlated Failures," *Computer Communications*, vol. 36, no. 6, pp. 708–717, Mar. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.comcom.2012.08.007

[17] T. Costello, "Business Continuity: Beyond Disaster Recovery," *IT Professional*, vol. 14, no. 5, pp. 62–64, Sep./Oct. 2012. [Online]. Available: http://dx.doi.org/10.1109/MITP.2012.92

[18] E. E. Anderson, "Firm Objectives, IT Alignment, and Information Security," *IBM Journal of Research and Development*, vol. 54, no. 3, May/Jun. 2010, paper 5. [Online]. Available: http://dx.doi.org/10.1147/JRD.2010.2044256

[19] P. Chołda, "Risk-Aware Design and Management of Resilient Networks," in *Proc. 4th International Workshop on Resilience and IT-Risk in Social Infrastructures RISI 2014*, Fribourg, Switzerland, Sep. 8, 2014.

[20] "Managing Information Security Risk. Organization, Mission, and Information System View," NIST SP 800-39, Mar. 2011.

[21] "Information Technology—Security Techniques—Code of Practice for Information Security Management," ISO/IEC 27002, Oct. 2005.

[22] "Security and Privacy Controls for Federal Information Systems and Organizations," NIST SP 800-53, Feb. 2012.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 871–878
DOI: 10.15439/2014F140
ACSIS, Vol. 2

A New Mode of Operation for Arbiter PUF to Improve Uniqueness on FPGA

Takanori Machida*, Dai Yamamoto†, Mitsugu Iwamoto* and Kazuo Sakiyama*
*The University of Electro-Communications
1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan
Email: {machida, mitsugu, sakiyama}@uec.ac.jp
†Fujitsu Laboratories Ltd.
4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki-shi, Kanagawa, 211-8588, Japan
Email: yamamoto.dai@jp.fujitsu.com

***Abstract*—Arbiter-based Physically Unclonable Function (PUF) is one kind of the delay-based PUFs that use the time difference of two delay-line signals. One of the previous work suggests that Arbiter PUFs implemented on Xilinx Virtex-5 FPGAs generate responses with almost no difference, *i.e.* with low uniqueness. In order to overcome this problem, *Double Arbiter PUF* was proposed, which is based on a novel technique for generating responses with high uniqueness from duplicated Arbiter PUFs on FPGAs. It needs the same costs as 2-XOR Arbiter PUF that XORs outputs of two Arbiter PUFs. Double Arbiter PUF is different from 2-XOR Arbiter PUF in terms of *mode of operation for Arbiter PUF*: the wire assignment between an arbiter and output signals from the final selectors located just before the arbiter. In this paper, we evaluate these PUFs as for uniqueness, randomness, and steadiness. We consider finding a new mode of operation for Arbiter PUF that can be realized on FPGA. In order to improve the uniqueness of responses, we propose *3-1 Double Arbiter PUF* that has another duplicated Arbiter PUF, *i.e.* having 3 Arbiter PUFs and output 1-bit response. We compare 3-1 Double Arbiter PUF to 3-XOR Arbiter PUF according to the uniqueness, randomness, and steadiness, and show the difference between these PUFs by considering the mode of operation for Arbiter PUF. From our experimental results, the uniqueness of responses from 3-1 Double Arbiter PUF is approximately 50%, which is better than that from 3-XOR Arbiter PUF. We show that we can improve the uniqueness by using a new mode of operation for Arbiter PUF.**

I. Introduction

RECENTLY, counterfeit products have been a problem in commercial market. The security for existing anti-counterfeit technologies relies on the technical difficulty to create a duplicate. However, future developments in counterfeit technologies might affect the technical difficulty. Physically Unclonable Function (PUF) [1] is being focused as a future solution [2].

PUF is a function in which an input (challenge) is related to one unique output (response) based on physical units such as semiconductor circuits. It is difficult to duplicate PUFs because the response values of PUFs depend on a physical variation. This difficulty to duplicate PUFs can be used device authentication against counterfeiting [3][4]. For example, a server as a verifier stores challenge–response pairs for a device as a prover. The verifier can authenticate the device by using the challenge–response pairs since they are unique for the PUF implemented in the device. PUFs are also used for a more secure method of storing secret keys than non-volatile memories. A secret key stored on internal memories will be revealed if an attacker can open the package of a device. In contrast, the secret key on PUFs cannot be read out accurately because physical variation and the values of responses have been changed once the package of the device is opened. Therefore, it is expected that PUFs are used for secure key generation [5][6].

PUFs are implemented not only on ASIC (Application Specific Integrated Circuit) [7] but also on FPGA (Field Programmable Gate Array) [8][9]. FPGA implementations have an advantage that their design and implementations are easy to change. Therefore, FPGAs are widely used in commercial products in the real world [10].

Some evaluation results on FPGAs of Arbiter PUF (APUF) [11], one of the delay-based PUFs, have been reported [12][13]. Previous work of [12][13] suggests that the APUFs implemented on Xilinx Virtex-5/Kintex-7 FPGAs generate responses with quite low uniqueness. The authors of [14] claim that one of the reasons for the low-unique responses obtained from APUFs on Virtex-5 FPGAs is based on the problem of SLICEs on the FPGAs. The problem is mentioned in general FPGAs. In a conventional APUF, a response is generated by comparing signals through two wires. The length of the two wires for any challenges in APUFs is expected to be equal. However, the layout of logic elements (*i.e.* SLICE) on FPGAs is completely fixed, so the length of wires among the logic elements cannot be controlled by designers. Because the difference between delay times arisen from physical variation is much smaller than that from the wire length, the responses obtained from different APUFs on different devices have small difference against a lot of challenges, *i.e.* low uniqueness.

In order to generate responses with high uniqueness on FPGAs, a novel technique that is called *Double Arbiter PUF* (DAPUF) [14] is proposed. The authors of [14] duplicate another APUF on neighboring SLICEs where the original APUF is implemented. They assume that a wire of duplicated APUF has almost the same length as the wire of the original APUF. 2-XOR APUF whose response is obtained by XORing 2-bit responses from two APUFs on the same FPGA are proposed in [3]. It has the same circuit costs as DAPUF, *i.e.*

978-83-60810-58-3/$25.00

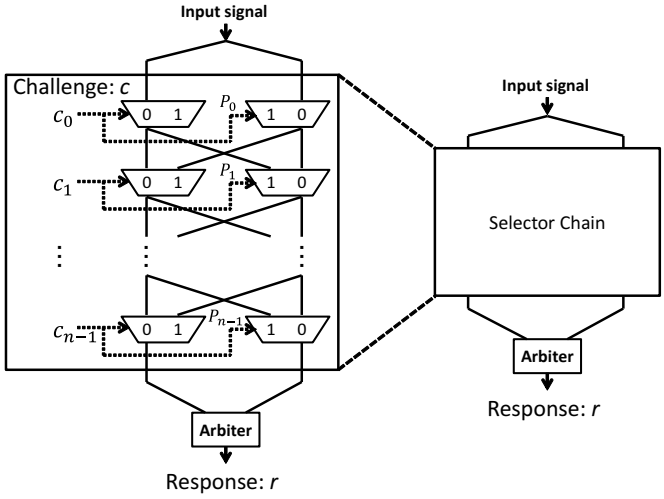


Fig. 1. Structure of conventional APUF

having two selector chains. In this paper, we call m-n APUF or DAPUF which has m selector chains and generates n-bit responses. For example, 2-XOR APUF is called *2-1 APUF*, and 2-bit output DAPUF is called *2-2 DAPUF*.

A. Our Contributions

In order to compare the structure of PUFs, we define a selector chain of conventional APUF as a building block as shown in Fig. 1. 2-2 DAPUF is different from 2-1 APUF in the wire assignment between the arbiter and output signals from the final selectors as shown in Figs. 2(a) and 3(b). Our two contributions in this paper are as follows:

- We introduce a new concept: *mode of operation for APUF* that is determined by a choice of the wire assignment. We compare PUFs that have two selector chains such as 2-2 DAPUF and 2-1 APUF and evaluate these PUFs on Virtex-5 FPGA regarding the uniqueness, randomness, and steadiness.
- We propose 3-1 DAPUF by using three selector chains, which is an improved version of 2-2 DAPUF. We compare it to conventional 3-XOR APUF, which have three selector chain. The evaluation results of these PUFs on Virtex-5 FPGA regarding the uniqueness show that 3-1 DAPUF generates responses with high uniqueness.

First, we evaluate four PUFs that have two selector chains. For a fair comparison, each PUF with the same-length response is compared. Therefore, 2-2 DAPUF is compared to *2-2 APUF*: two conventional APUFs as shown in Fig. 2(b). From our experimental results, we show that the uniqueness of responses from 2-2 DAPUF is higher than that from 2-2 APUF. We propose *2-1 DAPUF* by XORing 2-bit responses of 2-2 DAPUF as shown in Fig. 3(a), and compare it to 2-1 APUF. Our experimental results show that the uniqueness of responses from 2-1 DAPUF is approximately 41%, which is superior to 2-1 APUF.

One pair of the 2-2 DAPUFs has comparatively low uniqueness of responses because the proportion of 0s and 1s in responses (randomness) is still biased [14]. In order to eliminate the influence of the biased responses, we use another duplicated APUF, *i.e.* having three selector chains. In this paper, we propose *3-1 DAPUF* whose response is generated by XORing 6-bit responses of DAPUFs as shown in Fig. 4(a), for details to Sect. VI. Then, we compare 3-1 DAPUF to 3-XOR APUF that have three selector chains and generate 1-bit response. In this paper, we denote 3-XOR APUF as *3-1 APUF* as shown in Fig.4(b). Our experimental results show that the uniqueness of responses from 3-1 APUF is approximately 6%, which is still low. In contrast, the uniqueness of responses from 3-1 DAPUF is approximately 50%, which is much superior to that from 3-1 APUF.

We show that we can improve the uniqueness by using the new mode of operation for APUF and using responses obtained by XORing responses from more duplicated arbiter on Virtex-5 FPGAs.

B. Organization of This Paper

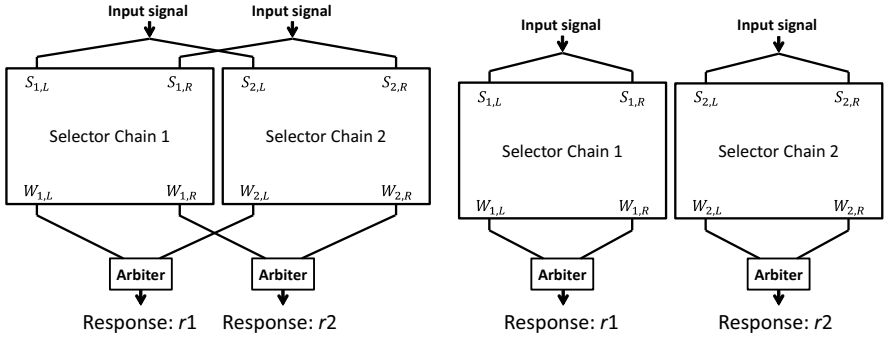
Organization of this paper is following. Section II gives previous work. Section III mentions the motivation of this work. Section IV shows experimental setup such as environment, and evaluation indicators. Moreover, we introduce the experimental results of conventional APUF evaluated by these indicators. Section V compares DAPUF to other APUFs which have two selector chains by using the indicators. Section VI proposes 3-1 DAPUF and compares it to 3-1 APUF, which have three selector chains by using the indicators. Finally, Sect. VII concludes this work.

II. ARBTIER PUF

Arbiter PUF (APUF) is one of the delay-based PUFs that use the difference between delay times of two signals. APUF has left and right selector pairs connected in series as shown in Fig. 1. Each bit of n-bit challenge c corresponds to a selection input c_i to the selector pair P_i ($0 \leq i < n$). After the challenge is determined, an input signal is supplied to the first selector pair P_0 at the same timing. For the case of $c_i = 1$, the left (right) selector in P_i is cross-connected to the right (left) selector in P_{i+1}, respectively. For the case of $c_i = 0$, the left (right) selector in P_i is straightly connected to the left (right) selector in P_{i+1}. This means that an input signal reaches through various paths depending on the value of the challenge. A 1-bit response is determined by which signal reaches an arbiter faster than the other. The response is strongly affected by the propagation path of the input signal, *i.e.* the challenge.

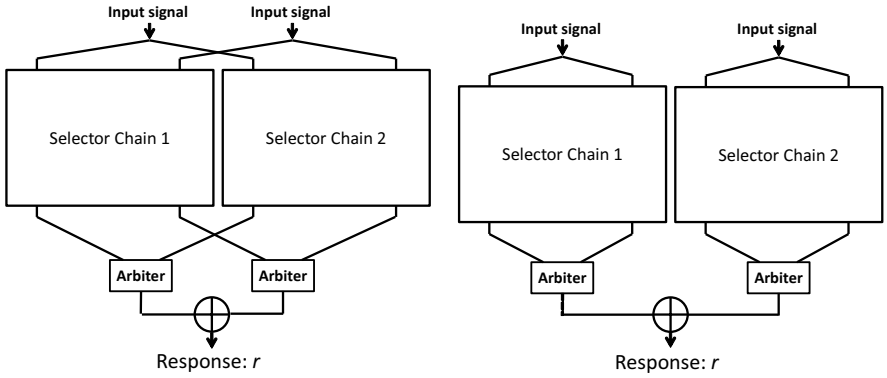
In APUFs, it is possible to increase the number of challenge bits easily by using more selector pairs [3]. APUF with n selector pairs has 2^n challenges. It is known that APUF can be modeled and simulated by building software models and programs based on the relation between challenges and responses [11]. This modeling against APUFs makes it possible for an attacker to predict responses for almost all challenges [15].

In order to prevent this modeling prediction, N-XOR APUFs have been proposed, where N-bit responses obtained from N APUFs are XORed into 1-bit response [3].



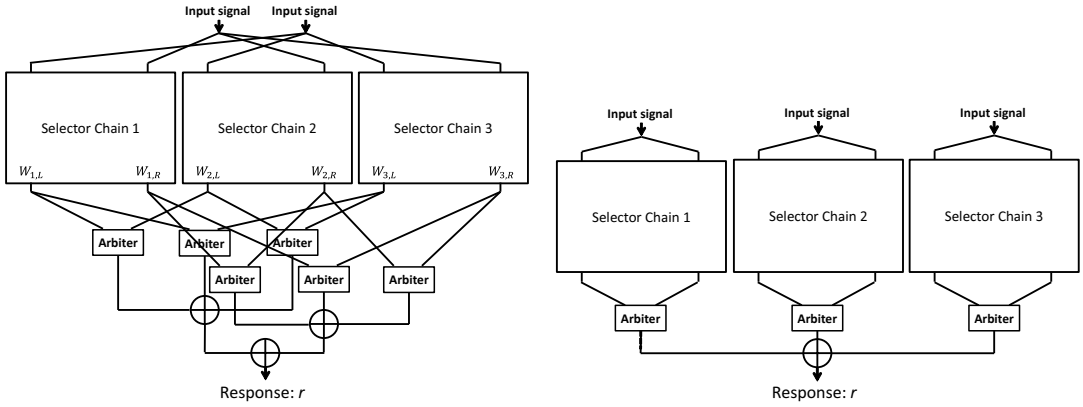
(a) 2-2 DAPUF (DAPUF) (b) 2-2 APUF

Fig. 2. Structure of 2-2 PUFs



(a) 2-1 DAPUF (b) 2-1 APUF (2-XOR APUF)

Fig. 3. Structure of 2-1 PUFs



(a) 3-1 DAPUF (b) 3-1 APUF

Fig. 4. Structure of 3-1 PUFs

III. MOTIVATION

Previous work of [12][13] reports that APUFs implemented on Xilinx Virtex-5 FPGAs generate responses with low uniqueness. DAPUF has been proposed in order to generate responses with high uniqueness even on such FPGAs [14]. Although almost all of DAPUFs improve the uniqueness of responses, the authors of [14] clarify that one pair of the DAPUFs has comparatively low-unique responses, which should be solved.

In this paper, we divide APUF into three components as shown in Fig. 1.

- An input signal part

- A selector chain part (building block)
- An arbiter part

Two wires are input to the top of the building block, and two wires are output from the bottom of it as shown in Fig. 1. The two wires from the top connect the input wires to 1-bit selector pair P_0, and that from the bottom connect output wires from n-bit selector pair P_{n-1} to the arbiter. The input signal is provided into the selector pair P_0 at the same time through the two wires from the top. Depending on the challenge value, the right input signal reaches the arbiter as the left or right signals.

To divide APUF into three parts enables us to clarify the difference between 2-2 DAPUF and 2-2 APUF: assignment of the input signal part and the arbiter part to the selector chain part, *i.e.* the mode of operation for APUF. We evaluate both PUFs with the same circuit costs as for the uniqueness, randomness, and steadiness of the PUFs and make comparison of these.

IV. Preliminary for Our Experiment

A. Experimental Environment

In this paper, APUF was implemented on a Xilinx Virtex-5 FPGA (XC5VLX30) [16] on SASEBO-GII (Side-channel Attack Standard Evaluation Board) [17]. Xilinx ISE 13.2 and Xilinx PlanAhead 13.2 were used for logic synthesis and floorplanning, respectively. We designed the APUF with 64-bit challenges so that wire length difference between two paths could be minimized by using Static Timing Analysis (STA), according to [9]. The selector pair (P_i) is located on the SLICE pair (X14Y(76-i), X15Y(76-i)). An input signal is supplied from the register with the equal distances from 1-bit selector pair P_0. A response is generated from another register (*i.e.* the arbiter in Fig. 1) with the equal distance from selector pair P_{n-1}.

B. Evaluation Indicators

Several performance indicators of PUFs are introduced in [18].

1) Uniqueness: When the same challenge is given to different PUFs, the responses should be completely different from one another. We use the value of SC Inter-HD (Same-Challenge Inter-Hamming Distance) [13] divided by the response bit length as the indicator of uniqueness. SC Inter-HD is calculated as the average of the Hamming distances between two responses obtained from different two PUFs for the same challenge. If the value of the uniqueness is close to 50%, we regard the uniqueness of the responses to be high in this paper.

2) Randomness: The proportion of 0s and 1s in responses should be equal. In this paper, we define the randomness of responses as the average number of 1s in responses (for randomly chosen challenges) divided by the responses bit length. The randomness of responses is 50% ideally.

3) Steadiness: When the same challenges are given to a PUF for repeated measurements, all of the responses should be the same. We use the value of SC Intra-HD (Same-Challenge Intra-Hamming Distance) [13] divided by the response bit length as the indicator of steadiness. SC Intra-HD is calculated as the average of the Hamming distances between arbitrary two responses for the same challenge. If the steadiness is close to 0%, we regard the steadiness of responses to be ideal.

TABLE I. Uniqueness of Conventional 1-1 APUF

Pair of FPGAs	uniqueness[%]
A with B	4.72
B with C	4.96
C with A	4.44

TABLE II. Randomness and Steadiness of Conventional 1-1 APUF

FPGA	randomness[%]	steadiness[%]
A	53.81	0.76
B	56.53	0.83
C	54.00	0.45

C. Results of Conventional Arbiter PUFs on FPGAs

Previous work [12][13] shows that APUFs implemented on Xilinx Virtex-5 and Kintex-7 FPGAs generate responses with low uniqueness, experimentally. In this section, we implement APUFs on Virtex-5 FPGAs and evaluate these PUFs for preliminary experiments.

First, we evaluate the uniqueness of 5000-bit responses obtained from APUFs on FPGA-A, FPGA-B, and FPGA-C. Table I shows the uniqueness of responses obtained from conventional APUFs on Virtex-5 FPGAs. The uniqueness is less than 5%, while 50% ideally. This means that physical variation of each PUF cannot be extracted as the uniqueness of responses.

Second, we evaluate the randomness of 2^{16} responses. The results are shown in the left part of Table II. The randomness is around 50%, which is the ideal. However, the authors in [19] report that the most of responses from APUFs on Virtex-5 FPGAs become either 1 or 0 with particular challenges. They evaluated 2^{16} responses from 64-bit APUFs for the challenges where $c_0 = c_1 = \cdots = c_7 = 1$ are fixed and $c_8, c_9, \ldots, c_{63}$ are randomly chosen. Under the condition that the Hamming weight is odd, the proportion of 1s in responses is approximately 80%, and under the condition that the Hamming weight is even, the proportion of 1s in responses is approximately 30% [19]. If the difference between the delay times of the two signals in selector chains is critically large, the responses are determined by whether the signal having larger delay than the other reaches right or left wire input to the arbiter. The two signals of conventional APUFs are crossed when $c_i = 1$ $(0 < i < 64)$. Therefore, whether Hamming weight of c_i is odd or even determines whether the signal having larger delay is supplied to the right or left wire. Under the condition of randomly chosen challenge, the proportion of 1s and 0s in responses become approximately 50%. In the result, the randomness of the responses from conventional APUFs on FPGAs comes to 50% regardless of the proportion of 1s and 0s in responses for particular challenges.

Finally, we evaluate the steadiness of 128-bit responses. The results are shown in the right part of Table II. The steadiness is calculated with 128-bit response for fixed challenges for 128 repeated measurements. The challenges are

randomly chosen. The steadiness is less than 1% among all pairs of FPGAs, which is nearly ideal as PUF. It shows that conventional APUFs generate the same responses for the same challenges. However, it is based on low uniqueness of the responses.

We reconfirm that the conventional APUFs on Virtex-5 FPGAs have low unique responses, which are not enough to perform as ideal PUFs.

V. DAPUFs *v.s.* APUFs OF $m = 2$

A. Modes of Arbiter Operation

1) Double Arbiter PUF: It is discussed that the length of the two wires in APUF are not equal at all [19]. In [14], it is suggested that the reason why conventional APUFs on Virtex-5 FPGAs generate low-unique responses is the unequal length of the two wires. Since the difference between delay times arisen from physical variations is much smaller than that from the signal propagation on the wire, the physical variations of each PUF cannot be found in responses, *i.e.* the uniqueness of responses become low. In order to generate responses with high uniqueness, it is proposed that a novel technique called *Double Arbiter PUF* (2-2 DAPUF) [14]. 2-2 DAPUF is designed for the purpose of equalizing the length of the two wires. Figure 5 shows the floorplanning for Xilinx PlanAhead. As illustrated in Fig. 5, the length of wires (1) and (2) seems equal but different precisely. Therefore, the authors duplicate another APUF on neighboring SLICEs where the original APUF is implemented. The authors expect that wire (1) has the almost the same length as wire (3) because both cell-pairs (1a,1b) and (3a,3b) are symmetrically located on the neighboring SLICEs.

Figure 2(a) shows the mode of operation for APUF of 2-2 DAPUF. Let $S_{i,L}$ and $S_{i,R}$ be the left and right wires which are inputs to the first selector pairs P_0 in Selector Chain i $(1 \leq i \leq 2)$, respectively, as shown in Fig. 2(a). The signals on $S_{1,L}$ and $S_{2,L}$ are supplied to Selector Chain 1 and 2 at the same time. Let $W_{i,L}$ and $W_{i,R}$ be the left and right wires, respectively. They are outputs from the n-th selector pairs P_{n-1} in Selector Chain i $(1 \leq i \leq 2)$, respectively. The signal on $S_{1,L}$ reaches $W_{1,*}$ $(* \in \{L, R\})$ and the signal on $S_{2,L}$ reaches $W_{2,*}$ regardless of the value of the challenge. Similarly, the signals on $S_{1,R}$ and $S_{2,R}$ are supplied to the Selector Chain 1 and 2 at the same time. Two 1-bit responses $r1$ and $r2$ are generated from two pairs of wires ($W_{1,L}$, $W_{2,L}$) and ($W_{1,R}$, $W_{2,R}$), respectively. Therefore, the signals on $S_{1,L}$ and $S_{2,L}$ are not crossed even if $c_i = 1$ $(0 < i < 64)$. The signals on $S_{1,R}$ and $S_{2,R}$ are also not crossed even if $c_i = 1$.

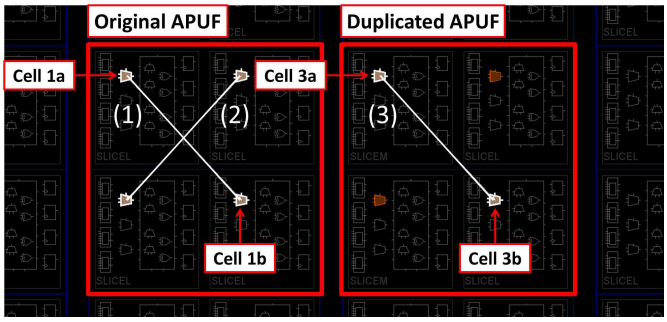


Fig. 5. Xilinx PlanAhead floorplanning

2-2 DAPUF generates 2-bit responses. We propose *2-1 DAPUF* that generates 1-bit responses obtained by XORing the 2-bit responses as shown in Fig. 3(a).

2) 2-XOR Arbiter PUF: We consider two PUFs that have the same circuit costs as DAPUFs of $m = 2$, *i.e.* having two selector chains. A straight forward example of APUFs of $m = 2$ is just two APUFs generating responses, *i.e. 2-2 APUF* as shown in Fig. 2(b). The signals on $S_{1,L}$ and $S_{1,R}$ are supplied to Selector Chain 1 at the same time. It depends on the value of challenges whether signal on $S_{1,L}$ or $S_{1,R}$ reaches $W_{1,L}$ or $W_{1,R}$. Similarly, the signals on $S_{2,L}$ and $S_{2,R}$ are supplied to the Selector Chain 2 at the same time. Two 1-bit responses $r1$ and $r2$ are generated from two pairs of wires ($W_{1,L}$, $W_{1,R}$) and ($W_{2,L}$, $W_{2,R}$), respectively. Therefore, the signals on $S_{1,L}$ and $S_{1,R}$ are crossed when $c_i = 1$ $(0 < i < 64)$. The signals on $S_{2,L}$ and $S_{2,R}$ are also crossed when $c_i = 1$. This is different from 2-2 DAPUFs in which the two signals are not crossed. We compare 2-2 APUF to 2-2 DAPUF and discuss the uniqueness, randomness, and steadiness. The difference of the wire connections has influence on its results, as mentioned in next section.

These 2-bit responses can be XORed into 1-bit responses: 2-1 APUF (2-XOR APUF) as shown in Fig. 3(b). We compare 2-1 APUF to 2-1 DAPUF according to the uniqueness, randomness, and steadiness.

B. Results

The results of the uniqueness and randomness of 2-2 DAPUF are from [14].

First, we evaluate the uniqueness of 5000-bit responses obtained from DAPUFs and APUFs of $m = 2$. The results are shown in Table III. The uniqueness of responses from 2-2 DAPUFs introduced in [14] is higher than that from 2-2 APUFs. However, the uniqueness of responses $r1$ between FPGA-A and FPGA-B is approximately 9%, which is comparatively lower than that of others. The reason for this is discussed along with the results of the randomness, as mentioned in the next paragraph. The uniqueness of responses from 2-1 DAPUFs is approximately 42%, which is much higher than that from 2-1 APUFs.

Second, we evaluate the randomness of 2^{16} responses. The results are shown in Table IV. In the following, we discuss the reason why 2-1 APUFs have low randomness although conventional APUFs have high randomness. From Table III, two conventional APUFs on Virtex-5 FPGAs generate low-unique responses: one PUF generates the same responses as the other PUF for many challenges. Therefore, 2-1 APUFs whose 1-bit response is obtained by XORing the responses of the two PUFs have low randomness obviously because the response becomes 0 when the same values are XORed. However, it is worth mentioning that the high randomness value is just superficial as mentioned in Sect. IV C. The reason why one pair of the 2-2 DAPUFs has comparatively low uniqueness of responses can be explained by low randomness of responses $r1$ on FPGA-A and FPGA-B. Here, we discuss one of the reasons of this low randomness.

TABLE III. UNIQUENESS [%] OF APUFS AND DAPUFS OF $m = 2$

	2-2 APUF		2-2 DAPUF		2-1 APUF	2-1 DAPUF
Pair of FPGAs	$r1$	$r2$	$r1$ [14]	$r2$ [14]	r	r
A with B	4.72	4.40	8.76	37.62	4.96	41.36
B with C	4.96	5.94	61.26	56.64	5.62	49.70
C with A	4.44	5.58	66.90	36.98	5.58	48.06

TABLE IV. RANDOMNESS [%] OF APUFS AND DAPUFS OF $m = 2$

	2-2 APUF		2-2 DAPUFs		2-1 APUF	2-1 DAPUF
FPGA	$r1$	$r2$	$r1$ [14]	$r2$ [14]	r	r
A	53.81	56.92	1.72	54.20	6.32	55.19
B	56.53	56.25	7.68	25.62	4.72	31.40
C	54.00	54.04	68.50	80.22	4.93	50.63

TABLE V. STEADINESS [%] OF APUFS AND DAPUFS OF $m = 2$

	2-2 APUF		2-2 DAPUF		2-1 APUF	2-1 DAPUF
FPGA	$r1$	$r2$	$r1$	$r2$	r	r
A	0.76	0.67	0.67	7.11	1.43	7.79
B	0.83	0.73	4.70	6.52	1.36	11.22
C	0.45	0.08	2.96	7.24	0.52	10.05

2-2 APUFs
: Even if the deterministic difference between delay times of the two signals is produced, the randomness seems high. Because it depends on Hamming weight of challenge whether the signal having lager delay than the other is supplied to left or right wire input to the arbiter.

2-2 DAPUFs
: In contrast, if the deterministic difference between delay times of the two signals is produced, the randomness can become low. Because it does not depend on Hamming weight of challenge whether the signal having larger delay is supplied to left or right input to the arbiter. This is caused by the signals on $S_{1,L}$ and $S_{2,L}$ ($S_{1,R}$ and $S_{2,R}$) are not crossed, as mentioned above.

Finally, we evaluate the steadiness of 128-bit responses. The results are shown in Table V. Almost all of the implemented 2-1 DAPUFs generate responses with lower steadiness than 2-1 APUFs. We consider that the large difference of delay times arisen from the imbalance wire length, as mentioned in Sect. IV C, results in high steadiness, but in contrast, low uniqueness. There is a trade-off between steadiness and uniqueness.

VI. 3-1 DAPUF *v.s.* 3-1 APUF

A. Modes of Arbiter Operation

1) 3-1 Double Arbiter PUF: One pair of the 2-2 DAPUFs has comparatively low uniqueness of responses because they have still biased responses [14]. In order to eliminate the influence of the biased responses, we use the following technique. We duplicate another APUF, *i.e.* having three selector chains, and generates multiple responses. Even if each of these responses is biased, we can obtain a less-biased response by XORing these responses. Let $W_{i,L}$ and $W_{i,R}$ be the left and right wires which are outputs from the n-th selector pairs P_{n-1} in Selector Chain i $(1 \leq i \leq 3)$, respectively. We use two wires chosen from three left wires: $W_{1,L}$, $W_{2,L}$, $W_{3,L}$ to generate three 1-bit responses as shown in Fig. 4(a). Similarly, we use two out of three right wires: $W_{1,R}$, $W_{2,R}$, $W_{3,R}$. Therefore, the left and right wires can generate six 1-bit responses in total. In this paper, we consider *3-1 DAPUF*: having three selector chains and generating a 1-bit response by XORing the six 1-bit responses.

2) 3-XOR Arbiter PUF: 3-1 APUF (3-XOR APUF) generates 1-bit responses obtained by XORing 3-bit responses from three conventional APUFs as shown in Fig. 4(b). The circuit costs of a 3-1 APUF are the same as that of a 3-1 DAPUF. They have three selector chains and generate 1-bit responses. We compare 3-1 DAPUF to 3-1 APUF according to the uniqueness, randomness, and steadiness.

B. Results

First, we evaluate the uniqueness of 5000-bit responses obtained from APUFs and DAPUFs of $m = 3$. The results are shown in Table VI. The uniqueness of responses from 3-1 DAPUFs is $50\pm1\%$, which is very close to the ideal results. In contrast, the uniqueness of responses from 3-1 APUFs is approximately 6%, which is much inferior to that from 3-1 DAPUFs. Further, 3-1 DAPUFs generate responses with high uniqueness among all pairs of FPGAs although one pair of the 2-2 DAPUFs has comparatively low uniqueness of responses. This means that we can eliminate the influence of the biased responses from the DAPUFs. The uniqueness of responses from 3-1 APUFs does not improve similarly to 2-1 APUFs. We consider that this is caused by the low uniqueness of each response from three conventional APUFs.

Second, we evaluate the randomness of 2^{16} responses. The results are shown in Table VII. The randomness of responses generated from 3-1 DAPUFs is around 50%, which is almost ideal. Further, the randomness of responses from 3-1 DAPUFs is more improved than that from 2-1 DAPUFs. The randomness value of responses from 3-1 APUFs seems high since the number of XORing responses are three (odd number).

Finally, we evaluate the steadiness of 128-bit responses. The results are shown in Table VIII. The steadiness of responses from 3-1 DAPUFs is approximately 12%, which is

TABLE VI. UNIQUENESS [%] OF APUF AND DAPUF OF $m = 3$

Pair of FPGAs	3-1 APUF	3-1 DAPUF
A with B	5.96	50.60
B with C	6.76	51.34
C with A	6.32	48.78

TABLE VII. RANDOMNESS [%] OF APUF AND DAPUF OF $m = 3$

FPGA	3-1 APUF	3-1 DAPUF
A	54.88	55.68
B	55.05	52.54
C	54.96	53.59

TABLE VIII. STEADINESS [%] OF APUF AND DAPUF OF $m = 3$

FPGA	3-1 APUF	3-1 DAPUF
A	1.43	14.11
B	1.36	10.93
C	0.74	10.35

inferior to that from 3-1 APUFs. We consider that one of the reasons is a trade-off between the steadiness and uniqueness.

We show the summary of the uniqueness, randomness, and steadiness for 1-1, 2-1, and 3-1 PUFs in Figs. 6, 7, and 8, respectively. The uniqueness of responses from APUFs and DAPUFs is improved with the increasing the number of selector chains. It is clear that the uniqueness of responses from DAPUFs using the new mode of operation is superior to APUFs. The randomness of responses from DAPUFs is also improved with that. That from only 2-1 APUFs having the even selector chains is lower than the other. However, the high randomness of 1-1 and 3-1 APUFs is just superficial as mentioned above section. In contrast, the steadiness of responses become low, *i.e.* the value of steadiness is high, with that.

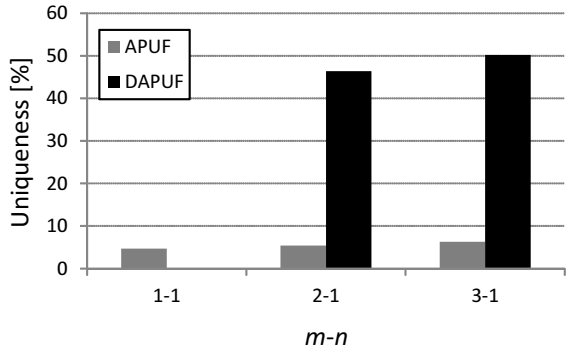


Fig. 6. Summary of uniqueness for m-n APUFs and DAPUFs

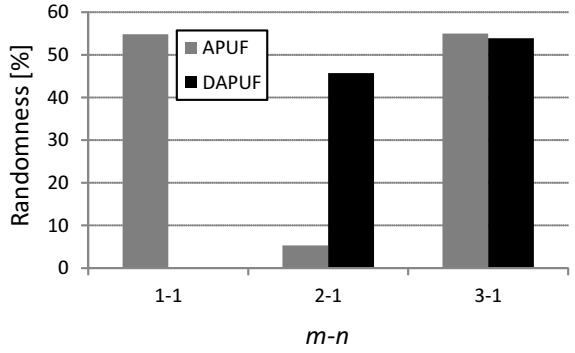


Fig. 7. Summary of randomness for m-n APUFs and DAPUFs

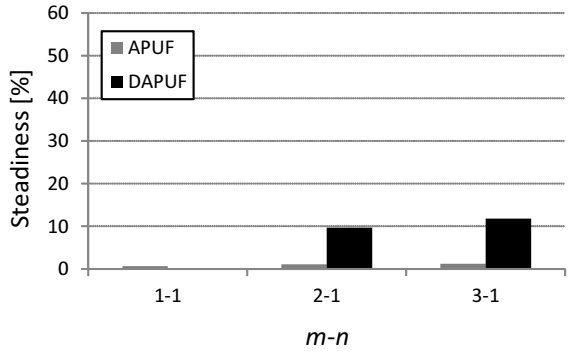


Fig. 8. Summary of steadiness for m-n APUFs and DAPUFs

We show that we can improve the uniqueness and randomness by using the new mode of operation for APUF and using responses obtained by XORing responses from more duplicated selector chains on Virtex-5 FPGAs.

VII. Conclusion and Future Work

2-2 DAPUF was proposed in order to generate responses with high uniqueness in previous work. In this paper, we introduced new concept: *mode of operation for APUF* that was determined by the connection method of the wires to arbiter. We compared DAPUFs and APUFs of $m = 2$ that have two selector chains such as 2-2 DAPUF and 2-1 APUF and evaluated these PUFs regarding the uniqueness, randomness, and steadiness. Further, we proposed 3-1 DAPUF by using three selector chains, which was improved version of DAPUF. We compare 3-1 DAPUF to 3-1 APUF, which have three selector chains, and evaluate these PUFs. From our experimental results, the uniqueness of responses from 3-1 DAPUFs was approximately 50%, which was much superior to that from 3-1 APUFs. On general FPGAs, we showed that we could improve the uniqueness and randomness by using the new mode of operation for APUF and using responses obtained by XORing responses from more duplicated selector chains.

The future work of this study is to implement *4-1* or *5-1 DAPUF* and to evaluate the responses from these PUFs according to the uniqueness, randomness, and steadiness. Further, we compare these PUFs to 2-1 and 3-1 DAPUFs by using the results.

References

[1] P. S. Ravikanth, "Physical one-way functions," Ph.D. dissertation, 2001, http://dx.doi.org/10.1126/science.1074376.

[2] P. Tuyls, B. Skoric, and T. Kevenaar, *Security with Noisy Data: Private Biometrics, Secure Key Storage and Anti-Counterfeiting*. Springer-Verlag New York, Inc., 2007, http://dx.doi.org/10.1007/978-1-84628-984-2.

[3] G. E. Suh and S. Devadas, "Physical Unclonable Functions for Device Authentication and Secret Key Generation," in *Proceedings of DAC*, 2007, pp. 9–14, http://dx.doi.org/10.1145/1278480.1278484.

[4] B. Gassend, D. Lim, D. E. Clarke, M. van Dijk, and S. Devadas, "Identification and authentication of integrated circuits." *Concurrency and Computation: Practice and Experience*, pp. 1077–1098, 2004, http://dx.doi.org/10.1002/cpe.805.

[5] Z. S. Paral and S. Devadas, "Reliable and efficient PUF-based key generation using pattern matching." in *in Proceedings of HOST*. IEEE Computer Society, 2011, pp. 128–133, http://dx.doi.org/10.1109/HST.2011.5955010.

[6] H. Handschuh, G. J. Schrijen, and P. Tuyls, "Hardware intrinsic security from physically unclonable functions." in *Towards Hardware-Intrinsic Security*, 2010, pp. 39–53, http://dx.doi.org/10.1007/978-3-642-14452-3_2.

[7] R. Maes, "Physically Unclonable Functions - Constructions, Properties and Applications," in *Springer*, 2013, http://dx.doi.org/10.1007/978-3-642-41395-7.

[8] J. H. Anderson, "A PUF design for secure FPGA-based embedded systems," in *Proceedings of ASP-DAC*, 2010, pp. 1–6, http://dx.doi.org/10.1109/ASPDAC.2010.5419927.

[9] K. Seki, Y. Hori, and H. Imai, "Implementation and Evaluation of Physical Unclonable Function on SASEBO-GII (in Japanese)," in *Symposium record of SCIS*, 2010.

[10] S. Saifullah, A. Khawaja, Hamza, Arsalan, Maryam, and Anum, "Keyless car entry through face recognition using FPGA," in *Proceedings of FITME*, 2010, pp. 224–227, http://dx.doi.org/10.1109/FITME.2010.5654862.

[11] D. Lim, J. W. Lee, B. Gassend, G. E. Suh, M. van Dijk, and S. Devadas, "Extracting secret keys from integrated circuits," *IEEE Trans. Very Large Scale Integr. Syst.*, pp. 1200–1205, 2005, http://dx.doi.org/10.1109/TVLSI.2005.859470.

[12] A. Maiti, V. Gunreddy, and P. Schaumont, "A Systematic Method to Evaluate and Compare the Performance of Physical Unclonable Functions," in *Embedded Systems Design with FPGAs*, 2013, pp. 245–267, http://dx.doi.org/10.1007/978-1-4614-1362-2_11.

[13] Y. Hori, T. Katashita, and K. Kobara, "Performance Evaluation of Physical Unclonable Functions on Kintex-7 FPGA (in Japanese)," in *IEICE Technical report of RECONF*, 2013.

[14] T. Machida, D. Yamamoto, M. Iwamoto, and K. Sakiyama, "A Study on Uniqueness of Arbiter PUF Implemented on FPGA (in Japanese)," in *Symposium record of SCIS*, 2014.

[15] U. Rührmair, F. Sehnke, J. Sölter, G. Dror, S. Devadas, and J. Schmidhuber, "Modeling Attacks on Physical Unclonable Functions," in *Proceedings of CCS*, 2010, pp. 237–249, http://dx.doi.org/10.1145/1866307.1866335.

[16] XILINX, "Virtex-5 FPGA User Guide," http//www.xilinx.com/support/documentation/user_guides/ug190.pdf.

[17] National Institute of Advanced Industrial Science and Technology, "Side-channel Attack Standard Evaluation Board (SASEBO)," http://www.risec.aist.go.jp/project/sasebo/.

[18] Y. Hori, T. Yoshida, T. Katashita, and A. Satoh, "Quantitative and Statistical Performance Evaluation of Arbiter Physical Unclonable Functions on FPGAs," in *Proceedings of ReConFig*, 2010, pp. 298–303, http://dx.doi.org/10.1109/ReConFig.2010.24.

[19] T. Machida, T. Nakasone, and K. Sakiyama, "Evaluation Method for Arbiter PUF on FPGA and Its Vulnerability (in Japanese)," in *IEICE Technical report of ISEC*, 2013.

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 879–884

DOI: 10.15439/2014F48
ACSIS, Vol. 2

Evaluation of highly available and fault-tolerant middleware clustered architectures using RabbitMQ

Maciej Rostanski
Academy of Business in Dabrowa Gornicza
ul. Cieplaka 1C, 41-300 Dabrowa Gornicza, Poland
Email: mrostanski@wsb.edu.pl

Krzysztof Grochla, Aleksander Seman
Proximetry Poland, Sp. z o.o.
Al. Rozdzienskiego 91 40-203 Katowice, Poland
Email: {kgrochla, aseman}@proximetry.pl

***Abstract*—The paper presents a performance evaluation of message broker system, Rabbit MQ in high availability - enabling and redundant configurations. Rabbit MQ is a message queuing system realizing the middleware for distributed systems that implements the Advanced Message Queuing Protocol. The scalability and high availability design issues are discussed. Since HA and performance scalability requirements are in conflict, scenarios for using clustered RabbitMQ nodes and mirrored queues are presented. The results of performance measurements are reported.**

I. Introduction

MODERN distributed systems have modular architecture. The applications, devices or appliances which are distributed parts of the whole solution, need to connect and scale. The applications need to connect to one another as components of a larger application, or to user devices and data. Nowadays, messaging, understood as an information flow or a network rather than a stack, needs to be supported by the system. As Richardson writes in [21]: "Future applications (..) [will be] always on, cloud hosted, and accessible anywhere. Input and processing are continuous and automatic, and deliver a filtered stream of information that the user wants, as it happens."

The middleware layer, often referred to as a 'glue' between different system components, allows communication between them. The modern requirement is to overpower the limits of point-to-point communication, and, moreover, to do it in a non-synchronous fashion. This is also referred to as a time- , space- and synchronisation-decoupling [6], and is especially important, given the fact, that the distributed systems now involve thousands of entities, which may be distributed throughout vast geographical distances, and whose behaviour and even location may vary in time. Message queuing also called message-oriented middleware is an architectural pattern. It is based on a message broker, an intermediary program which realizes message validation, message transformation and message routing functions. Message broker provides common infrastructure for interactions between elements of the distributed systems, which interact by sending or receiving messages. It is a recent alternative for distributed interaction between components, entities of an information processing system. Message queuing is thoroughly described, for example, in [2], [3] and [6]. It is often based on a publish/subscribe-like interaction [6]. The message queuing is an alternative to Classifications, which are complementary to the publish/subscribe model of a distributed information system [9]. Classifications involve techniques such as message passing, shared spaces or remote invocations and constitute solutions to the middleware layer challenges. Middleware systems are also subject to numerous studies, concentrating on networking and concurrent design. There is a concept of using patterns in overall software architecture, with [5] as a main example, or for security related applications ([7], [8]).

This work was supported by Proximetry Poland

This paper describes design considerations of scalability and high availability (HA) improving solutions using RabbitMQ software, an open source message broker and queuing server that is becoming more and more popular as a middleware. The balance between HA and scalability is challenging because of contrary requirements, the scalability and performance-optimisation mechanisms are in principle hindered by high availability or fault tolerance solutions, which prefer stability and durability over performance.

This paper presents possible configuration scenarios for a RabbitMQ cluster of servers, which combine scalability with high availability / fault tolerance (HA/FT) requirements. For this purpose, RabbitMQ is described as middleware and clustering options are presented as well as HA possibilities. The scenarios were implemented for test-field studies, whose results are presented. Most of the available literature or reports such as [1], [19] or [13] concentrates on the scalability issues and performance results, or, from a different perspective, strictly on high availability / fault-tolerance solutions for queuing [21]. This paper aims to bring a novelty in discussing solutions that combine both requirements, as it is a probable industry scenario.

The paper is organized as follows: the design requirements for middleware system are presented and briefly explained—specifically, scalability and high availability concerns are discussed. The next part includes a short summary of RabbitMQ, the message broker used in research. The main part includes message broker configuration scenarios for scalability and high availability; the experimental results of constructed systems are presented for comparison. Finally, conclusions are revealed.

978-83-60810-58-3/$25.00

II. RABBITMQ AS A MIDDLEWARE

From designer's perspective, message-oriented middleware can be seen as a (1) queuing system, where messages are concurrently pulled by consumers, as well as (2) subscription-based exchange solution, allowing groups of consumers to subscribe to groups of publishers, resulting in a communication network or platform, or a message bus [6]. Such bus or queuing system has to be able to scale in terms of geographical distance as well as in terms of devices or applications served. Quoting Jones et al. [13], "the distribution of information sent from the publishers to the hub to be distributed to the necessary subscribers allows for applications to run while relying on data from other locations, wherever they may be."

RabbitMQ is an open source message broker and queuing server that can be used to let disparate applications share data via a common protocol or to simply queue jobs for processing by distributed workers. RabbitMQ middleware supports many messaging protocols [17], among which the most important are STOMP: Streaming Text Oriented Messaging Protocol [20] and AMQP: Advanced Messaging Queuing Protocol [11].

Within this paper the AMQP-defined messaging architecture is used. Within the core of the message broker architecture are queues; every message received by the RabbitMQ always is placed in a queue. Messages in queues can be stored in memory (memory-based) or on a disk (disk-based). Second important elements of the RabbitMQ are *exchanges* - the delivery service for messages. The exchange used by a publish operation determines if the delivery will be direct or publish-and-subscribe, for example. A client chooses the exchange used to deliver each message as it is published. The exchange looks at the information in the headers of a message and selects where they should be transferred to. This is how AMQP brings the various messaging idioms together - clients can select which exchange should route their messages [15].

A. Specific system design requirements for middleware

1) Scalability: Scalability is an architectural characteristic, which can be defined as a capability to cope and perform under an increased or expanding workload. A system that scales well will be able to maintain or even increase its level of performance or efficiency when tested by larger operational demands. In terms of message-queuing, or even publisher/-consumer exchange system, this would mean the possibility of increasing processing speed or message throughput, user capacity, etc.

2) Resiliency: In order to be resilient (which means to be able to deal with internal failures), the system needs to implement some forms of high availability (HA) or fault tolerance (FT). In general, HA and FT systems are designed with two different design principles in mind. Given the availability (A) formula (eq. 1),

$$A = \frac{MTBF}{MTBF + MTTR}$$

HA aims to minimize downtime and IT service disruption; so the common goal in HA is to increase Mean Time Between Failure (MTBF) and decrease Mean Time to Repair (MTTR). HA applications are designed to have a high level of service uptime. HA solutions may feature many elements, e.g: system management, live replacement (hot-swap), component redundancy and failover mechanisms. Common strategy is to avoid single points of failure in the system. This can be difficult, because demands on such systems include not only ensuring the availability of important data, but also efficient resource sharing of the relatively expensive components.

Typical HA solution involves clustering; symmetrical (all nodes have similar capabilities) or asymmetrical (nodes have different possibilities and inventory). Clustering in this context can be described as the use of two or more systems loosely coupled to provide system level redundancy. Because these systems are not directly coupled, they utilize standard network connections to communicate failovers. This can cause failover latencies to take several seconds to complete. Typically, there is a middleware software solution to provide a failover mechanism between the two systems. But this middleware has to be protected with HA in mind as well, possibly with the use of clustering.

Contrary to HA, which implies a service level in which both planned and unplanned outages do not exceed a small stated value [18], fault-tolerant (FT) systems tend to implement as much component redundancy and mirroring techniques as possible, in order to eliminate system failures completely (this is of course from client's perspective, in fact introducing redundant components will make component failures occur faster) [4].

But FT has its problems; performance degradation is another concern. As an example, let's discuss mirroring a single server. Besides handling all of the file transfer work for network users, the primary server may have to process additional I/O as it passes information along to the mirror server. This can also add substantial processor overhead if system usage is heavy. In effect, RAM, CPU and network performance is degraded.

B. Scalable and fault-tolerant middleware

Message broker, being one of the most crucial components of distributed system, is supposed to be fault-tolerant. That means, typical HA configuration (as described in II-A1) is not the best option; restarting message broker on another node in case of failure would introduce a timeout span, as the service is being restarted and prepared for operation, but, what is worse, the message queue of failed message broker would be lost entirely. For message broker, both HA and FT solutions were considered:

1) HA (Active/Passive solution): in which the downtime of message broker service is expected in case of planned or unplanned unavailability of primary server. Queues and messages have to be persistent (disk-based), and message broker can be restarted elsewhere in the system. It is possible to base such solution on virtualization, where MB running host can be virtualized and rely on hypervisor built-in HA mechanism. This would cause hypervisor to run another instance of VM in case of a failure of primary MB guest or

even virtualization host. Another active/passive solution is to deploy clustering HA solution like pacemaker [16] in order to manage message broker and restart it (or migrate) when necessary, using available resources.

2) FT (Active/Active solution): means that the planned or unplanned downtime of message broker doesn't have any effect on queuing system. Typically it is implemented by MB leveraging clustering mechanism built-in RabbitMQ, which is developed strictly for such situations, and replicates queues on every RabbitMQ node in the cluster. RabbitMQ nodes failure monitoring, and IP load-balancing techniques are explained further in detail. Active/active solution can also be based on virtualization, where MB running host can be virtualized and, for example, marked as FT-demanding in VMware vCenter virtualisation environment. This would create a VM mirror image called "replica", updated in real-time, ready to be run in case of a failure of primary MB host.

The second solution is a typical Active/Active topology and is recommended as more reliable and scalable at the same time. Additionally, virtualization was not considered, because it would introduce additional conditions and variables to the experiments and is a subject for another study. This paper's research is concentrated on RabbitMQ message broker clusters and its characteristics.

C. RabbitMQ cluster setup and operation

The clustering built into RabbitMQ was designed with two goals in mind: allowing consumers and producers to keep running in the event of node failure, and linearly scaling messaging throughput by adding more nodes [21]. With clustering, a client can connect as normal to any node within a cluster. If that node should fail, and the rest of the cluster survives, then the client should notice the closed connection, and should be able to reconnect to some surviving member of the cluster. [17]

The design decision that had to be made was an IP addressing of the cluster. As RabbitMQ documentation describes, it's not generally advisable to hardcode node hostnames or IP addresses into client applications: this introduces inflexibility and will require client applications to be edited, recompiled and redeployed should the configuration of the cluster change or the number of nodes in the cluster change. As in general, this aspect of managing the connection to nodes within a cluster is beyond the scope of RabbitMQ itself. RabbitMQ's authors recommend a more abstracted approach, including a dynamic DNS service which has a very short TTL configuration, or a plain TCP load balancer (for example HAproxy [12]), or some sort of mobile IP achieved with pacemaker or similar technologies [16]. For this study, HAproxy was chosen as a load balancer between clients and cluster nodes.

III. Clustering scenarios

The maximization of the systems performance suggests that content of the queues should not be replicated throughout the cluster. The queue owner node has full information about it; other nodes in the cluster only know the queue's metadata and a pointer to the node where the queue actually is stored.

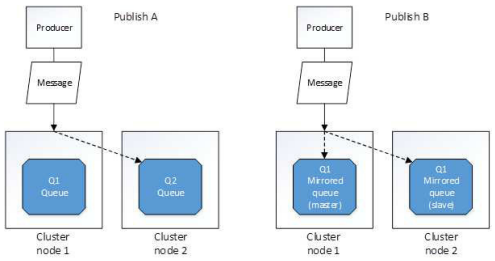


Fig. 1. Publishing to queues: a) on another node, b) to the mirrored queue.

This solution allows to limit storage space requirements and increase performance—replicating messages to every node would result in increase of network and disk load for every node, keeping the performance of the cluster the same (or worse) [21]. Regardless where publish is made, message will end up on the queue owner node. This leads to main performance optimization technique: to increase performance for every added node by spreading queues across nodes. On the contrary to performance-driven requirements for queues, there is a need for queue to be redundant when the main goal is to achieve high availability and fault tolerance. If a queue owner node fails, all of the messages within a queue are gone. An active-active redundancy option is possible; any queue can be mirrored. The mirrored queue is achieved by creating slave copies of the queue on other nodes in the cluster. It can be copied on every node, but the designer is able to specify a subset of nodes in the cluster for a queue to live on. Both situations are presented on Fig. 1.

The design of the cluster and its queues can support the following:

1) Creating fully mirrored queues on every node in order to achieve HA; create very efficient connection between nodes and create RAM nodes for quick distribution of messages,
2) Creating spread queues, but configure mirrored queues for at least one master and one slave (allowing for one node failure),
3) Creating fully spread queues and do not mirror them, but make them durable instead—all of the nodes are disk based, and in the event of failure, message broker is restarted elsewhere.

Within above listed possibilities, 1) is a scenario for maximum fault-tolerance, 3) is a scenario allowing some downtime for maximum performance (which is HA scenario) and 2) is a compromise between those two.

A. Cluster and queues configuration

Considering a three-node cluster, one can come up for specific testing scenarios that can provide comparable results for performance assessment [14]. Those results may provide an answer, whether given configuration is useful for a specific real-world scenario [10]. For testing purposes, there were following implications made: a) cluster may include up to three nodes, b) queues are created in the cluster as a single

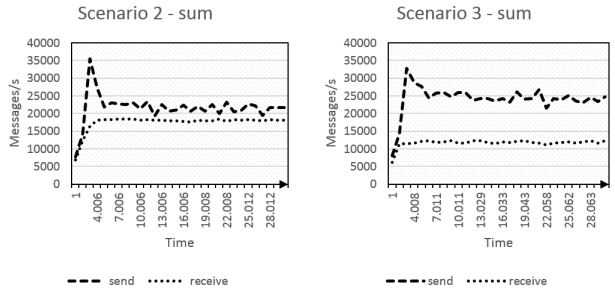


Fig. 3. The comparison of two-queue, one-client-per-queue publishes and consumes using scenario 2 and scenario 3.

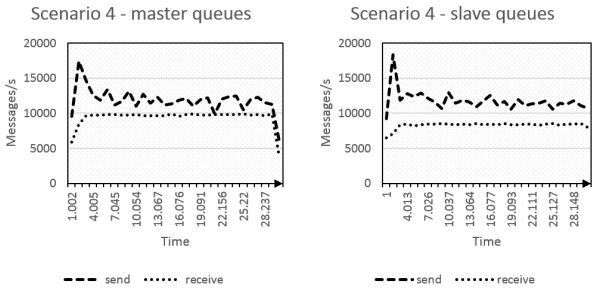


Fig. 4. Mirrored queues performance results example.

(non-mirrored), fully mirrored, and spread (mirrored to one node) queue, c) all of configuration scenarios are put to the three tests:

1) single publishes and consumes: there are single publishers and consumers for both queues,
2) balanced conversations: there are three publishers and three consumers for every queue (as many as cluster nodes),
3) many conversations: there are six publishers and six consumers.

Possible configuration possibilities for two queues are presented on Fig. 2. On the left, there is no fault-tolerance; queues are not mirrored and the cluster is configured for performance scalability; on the right, queues are mirrored for resiliency—this is possible only using two nodes minimum; adding the third node creates two possibilities—mirroring queues and adding one node for scalability (scenario 5) or spreading mirrored queues on available nodes (scenario 6).

B. Initial testing and results

All of scenarios were tested with commodity-equipped virtual machines (single core, 4GB RAM, 8GB HDD) which eliminated any possible networking issues. The hypervisor host was equipped with Intel i7 CPU and 32GB RAM. There was no resource overload. The most interesting observations were as follows. In conclusion A, there is practically no difference between the performance of publishing to single or multiple queues on one node. Scenario 1 is viable and does not introduce any performance problems. This question doesn't need any more evaluation.

The results of scenario 2 and scenario 3 (more than one node in the cluster) show significant improvement of performance over scenario 1. Fig. 3 presents exemplary results of single publisher and consumer for both queues, summarized for comparison. These results are expected, however designer has to keep in mind such configuration is not fault-tolerant—if a node fails, the queue is no longer available for publishing or consuming. The difference between scenario 2 and scenario 3 is interesting and should be a subject for another study—adding supplementary node allowed faster publishing, but the consuming rate dropped, as the cluster nodes communication introduced an overhead. In effect, whole system performance was kept on the same level. This design could be more appropriate with large number of publishers and consumers, but such study is out of scope for this paper.

The performance of sending and receiving to mirrored queues, as in scenarios 5 and 6, is significantly worse. Fig. 4 shows typical results (scenario 4 is shown). The difference between publishing and consuming to the master (owner) node compared to publishing and consuming from the slave node is as expected - publishers are unaffected, but the consumers suffer from intra-cluster traffic (master-to-slave) overhead.

C. Detailed evaluation

The initial tests results show the importance of load balancing the traffic between the clients and cluster nodes. The message publishing or consuming rate depends whether the client was redirected to the:

1) "master" node (the node which is the master for the specific queue being used),
2) "slave" node (the node which specific queue is being replicated onto),
3) "empty" node (the node which is part of the cluster but the queue resides on other nodes).

If the client is redirected onto "master" or "slave" nodes, the published messages do not need to be communicated to every node in the cluster, which has good effect on performance. Otherwise, message sending/receiving rates drop.

For detailed information on this impact, the cluster was set up with three nodes - two disk-based and one RAM-based. Such configuration assures that if queue is mirrored, it always resides on at least one disk-based node, and messages are written to disk and can be retrieved even after power failure. For such cluster, a single queue was tested for performance when configured as:

1) "single" queue (not mirrored at all, for performance comparison and load balancing issues evaluation described before),
2) "spread" queue (mirrored to one other node in the cluster, as suggested in scenario 6),
3) "mirrored" queue (mirrored to every other node in the cluster).

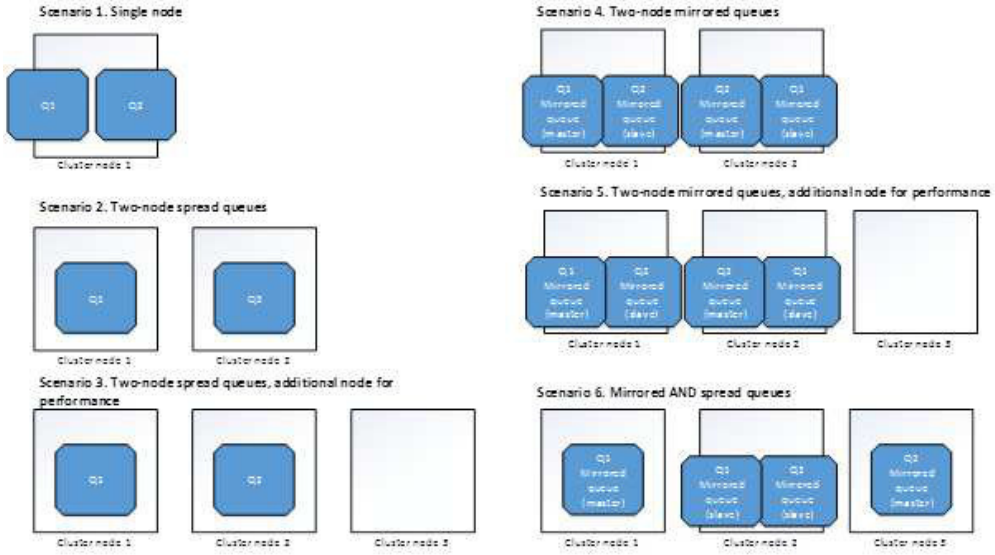


Fig. 2. Performance-driven vs. Fault-Tolerant-driven testing scenarios.

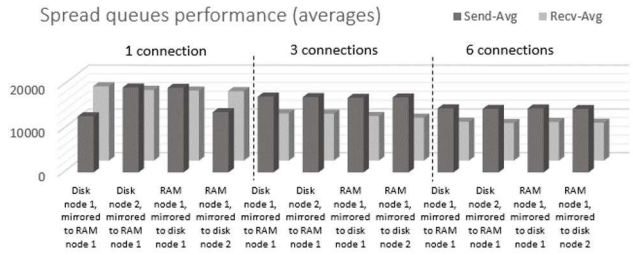


Fig. 5. Spread queues performance extensive testing results comparison

The results are presented in Table 1. For every queue configuration average message publication/consummation rates (for 9 consecutive tests) are shown, as well as standard deviation. The effect of load-balancing is visible mostly for single client consuming from mirrored queues—if the client's request is redirected to an empty node, the average message receive rate is significantly lower. Also, the effect is visible when sending to single-hosted queue, as the two other nodes are empty, therefore they need to redirect a request.

Fig. 5 presents exemplary detailed extensive testing run results in quantitative form, but gathered for comparison and conclusions. Every queue configuration was created on master disk or RAM node, then tested for a minute-long run ten consecutive times. Such tests were conducted for one, three and six simultaneous connections.

Most important observations are, as indicated by mutliple tests (ca. 50 re-runs), the performance of single queue drops significantly when this queue is mirrored throughout entire cluster for fault-tolerance. Full mirrored queue is therefore not as good architectural choice as it would seem, especially if there are frequent moments of only one producer active. The performance of spread queues is about 10%-20% minimum better than fullmirrored queues on a three-node cluster. There is no significant difference in RAM / disk node effect on mirroring. Spread queues are stable; performance degradation is however visible when receiving by many clients at once.

IV. Conclusions

This paper shows there are many considerations for building clustered middleware and implementing scalable yet fault-tolerant system. Queues need to be distributed evenly, or internal transfers within the cluster will cause performance to drop, especially for receiving clients. There is, however a way to mirror queues asymmetrically, which is shown by experimental results in this paper.

Relevant studies show many more aspects have to be taken under consideration—for example, the disk-based nodes compared to RAM-based nodes performance, or the expected distribution of the clients (publishers and consumers) but results show there is a possibility to create a design principles for specific clients count and message rates requirements, which can be a subject of next authors' study. This is authors' contribution in discussing solutions that combine both requirements, as it is a real industry scenario.

To summarize results, study shows that while typical single queues on clustered nodes are key to performance, if the requirements include fault-tolerance, performance can still be improved by "spreading" queues to be mirrored only by one more node, as N+1 rule dictates.

Acknowledgement

This work is partially supported by NCBIR INNOTECH Project K2/HI2/21/1 84126/NCB R/13, The Effective Management of Telecommunication Networks Consist of Millions of Devices.

TABLE I
SUMMARIES FOR MOST IMPORTANT SCENARIOS (10 PUBLISHERS, 10 CONSUMERS)

Scenario	Scenario 3 (performance)	Scenario 5 (mirrored)	Scenario 6 (spread mirrors)
Average publish rate [msg/s]	33296.59	8553.28	12668.86
Average consume rate [msg/s]	16162.00	5087.00	8231.55

REFERENCES

[1] M. Altherr, M. Erzberger and S. Maffeis, "iBus - a software bus middleware for the Javaplatform," in: *Proceedings of the International Workshop on Reliable Middleware Systems*, 1999, pp. 43–53.

[2] G. Banavar, T. Chandra, R. Strom, and D. Sturman, "A case for message oriented middle-ware", in: *Proceedings of the 13th International Symposium on Distributed Computing (DISC99)*, 1999, pp. 1–18

[3] B. Blakeley, H. Harris, and J. Lewis, *Messaging and Queuing Using the MQI*. McGraw-Hill, New York, NY, 1995.

[4] P. Buchwald, "The Example of IT System with Fault Tolerance in a Small Business Organization", in: *Internet—Technical Development and Applications 2*, Springer 2012, pp. 179–187

[5] F. Buschmann et al., it Pattern-oriented software architecture: a system of patterns, John Wiley and Sons, Inc. New York, NY, USA ÂŠ1996 ISBN:0-471-95869-7

[6] Eugster et al., "The Many Faces of Publish/Subscribe", in: *ACM Computing Surveys, Vol. 35, No. 2*, June 2003, pp. 114–131.

[7] X. Yuan and E. B. Fernandez, "Patterns for Business-to-Consumer E-Commerce Applications", accepted for the International Journal of Software Engineering and Applications (IJSEA)

[8] M. VanHilst, E. B. Fernandez and F. Braz, "A Multidimensional Classification for Users of Security Patterns", in *Journal of Research and Practice in Information Technology, vol. 41, No 2*, May 2009, pp. 87–97

[9] M. Franklin and S. Zdonik, "A framework for scalable dissemination-based systems",in: *Proceedings of the 12th ACM Conference on Object-Oriented Programming Systems, Languages and Applications (OOPSLA'97)*.ACM Press, New York, NY, 1997, pp. 94–105.

[10] K. Grochla, L. Naruszewicz, "Testing and Scalability Analysis of Network Management Systems Using Device Emulation", in: *Computer Networks*, Springer 2012, pp. 91-100

[11] P. Houston, "Building distributed applications with message queuing middleware" (Whitepaper). Available online at http://msdn.microsoft.com/library/en-us/dnmqqc/html/bldappmq.asp, 1998

[12] "HAProxy. The Reliable, High Performance TCP/HTTP Load Balancer". Website: http://haproxy.1wt.eu/, accessed: 21.01.2014

[13] B. Jones, S. Luxenberg, D. McGrath, P. Trampert and J. Weldon, "RabbitMQ Performance and Scalability Analysis", project on CS 4284: Systems and Networking Capstone, Virginia Tech 2011

[14] S. Nowak, M. Nowak and M. Foremski, "New Synchronization Method for the Parallel Simulations of Wireless Networks", in *11th International Conference, NEW2AN 2011, and 4th Conference on Smart Spaces, ruSMART 2011, St. Petersburg, Russia, August 22-25, 2011. Proceedings, LNCS 6869*, Springer Berlin Heidelberg, pp. 405–415

[15] J. O'Hara, "Toward a Commodity Enterprise Middleware", *ACM Queue 5 (4)*, June 2007, pp. 48–55

[16] "Pacemaker. A scalable High Availability cluster resource manager". Website: http://clusterlabs.org/, accessed: 18.01.2014

[17] RabbitMQ documentation [online], http://www.rabbitmq.com/documentation.html, accessed 21.01.2014

[18] M. Rostanski, "High Availability Methods for Routing in Soho Networks", in *Internet - Technical Developments and Applications 2*, Springer 2011, pp. 154–152

[19] Salvan, M., *A quick message queue benchmark: ActiveMQ, RabbitMQ, HornetQ, QPID, ApolloâĂŞ* [online: http://bit.ly/1b1UGTa], April 2013

[20] The Simple Text Oriented Messaging Protocol website [online], http://stomp.github.io/, accessed 20.01.2014

[21] A. Videla and J. Williams, *RabbitMQ in action. Distributed messaging for everyone*. Manning, April 2012

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 885–889
DOI: 10.15439/2014F43
ACSIS, Vol. 2

Solution for Secure Private Data Storage in a Cloud

Kirill Shatilov, Vladislav Boiko, Sergey Krendelev, Diana Anisutina, Artem Sumaneev

Department of Information Technology, Novosibirsk State University, Novosibirsk, Russia

shatilov@ccfit.nsu.ru, boikovladislav@gmail.com, s.f.krendelev@gmail.com, diana.anisutina@gmail.com, sumaneevartem@gmail.com

***Abstract*—Cloud computing and, more particularly, cloud databases, is a great technology for remote centralized data managing. However, there are some drawbacks including privacy issues, insider threats and potential database thefts. Full encryption of remote database does solve the problem, but disables many operations that can be held on DBMS side; therefore problem requires much more complex solution and specific encryptions. In this paper, we propose a solution for secure private data storage that protects confidentiality of user's data, stored in cloud. Solution uses order preserving and homomorphic proprietary developed encryptions. Proposed approach includes analysis of user's SQL queries, encryption of vulnerable data and decryption of data selection, returned from DBMS. We have validated our approach through the implementation of SQL queries and DBMS replies processor, which will be discussed in this paper. Secure cloud database architecture and used encryptions also will be covered.**

I. Introduction

RAPID growth and development of cloud technologies has leaded them to popularity and widespread usage. Although customers are excited by cloud features and benefits, they are very concerned about confidentiality of data stored and processed in a cloud. Insider threats combined with a general lack of transparency into provider process and procedure has dropped confidence in security of cloud data storage [1].

Data confidentiality is highly important for Cloud Databases (Database as a Service, DbaaS), and there are threats of disclosure of vulnerable user's data to unauthorized parties. First of all, curious and malicious database administrators may capture or leak data [3]. Also a theft of database may possibly occur, leaving data in hands of malefactor [4].

Listed problems are still actual [2], and as a result multiple solutions to problem of trusting clouds have been developed. Encryption of all data in remote database was offered as a method of providing provable confidentiality [5, 6]. But such an approach demands all operations will be held on a client side after decryption of database content. Other solutions, such as MIT CryptDB [7], lack fully homomorphic encryptions and use third-party encryptions with relatively low cryptostrength and known vulnerabilities [8].

To address listed cloud security issues we designed secure cloud database architecture, several encryption algorithms and a SQL data encoding component. Proposed solution addresses mentioned challenges using following key ideas:

- The first is to parse SQL queries and encrypt selected user data on client side. None of encryption keys are passed to server or to any proxy; all confidential data passed to DBMS are secure.
- The second idea is usage of wide variety of order preserving and fully homomorphic encryptions. All encryptions that are used in proposed solution are proprietary developed encryptions with relatively high speed and provable cryptostrength.
- The third technique is a combination of various encryptions in single table. Encrypting different columns with different encryptions within one table greatly reduce chances of successful cryptanalysis.

In this paper we present designed architecture, description of main components and overview of used encryptions.

The next section of paper features a brief explanation of basic principles giving the main idea of a handling of SQL queries used in proposed solution. After it, in Section 3, we describe the general architecture of proposed secure cloud database and client's component. Section 4 gives some insight into encryptions, which are used in secure SQL queries processing. Finally, the last section of this paper summarizes our achievements, also exposing some possible future development and additional security features improving protection of cloud database.

II. Basic Principles

In this Section basic principles are discussed illustrating main idea of secure cloud database. Core component of proposed secure cloud database is SQL queries processor. All user's SQL queries are analyzed and transformed by program components on client's side before sending to DBMS (see Fig. 1). Similarly, all replies from database are processed.

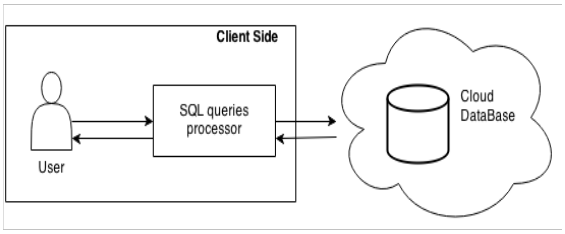


Fig. 1 Overview

Also, encryption of data, extracted from user's queries, is a responsibility of SQL query processing component. Proposed solution doesn't require any modifications on DBMS side.

User's SQL query handling is shown on Fig. 2. Information about encrypted columns such as encryption keys and encrypted column name(s) is needed in order to parse, decrypt and reconstruct user queries.

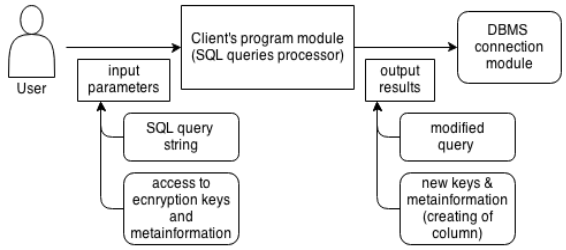


Fig. 2 Query processing

When user retrieves data from cloud database, selection of column comes as a reply from DBMS. Information about encrypted columns is needed to decrypt and present selected data to user in suitable form (see Fig. 3)

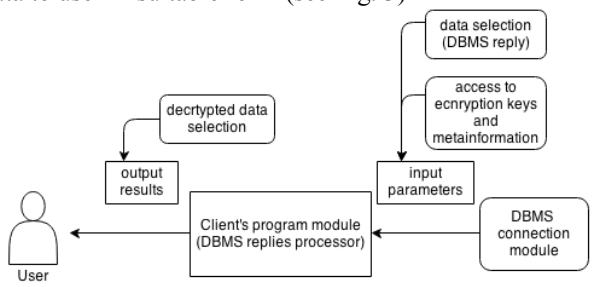


Fig. 3 Data selection processing

Basic idea of SQL queries processing is explained on different types of SQL expressions below.

"CREATE" statement is the only statement in terms of proposed secure cloud database, which may contain additional keywords, which are not included in SQL language. These keywords are markers of different encryptions applied to table columns. To indicate that column in currently created table should be encrypted, user should add a constraint corresponding to encryption's marker (identification string).

If SQL queries processing component encounters encryption marker, following steps will be performed. First, there encryption's keys are generated or chosen. Next, based on encryption information, number (can be more than one, in case when the result of encryption is a vector of multiple values), names, types and constraints of output columns are determined. Correct SQL string is created according to determined information.

After that modified statement is sent to DBMS. Output names of encrypted columns are anonymized, while anonymization of table name is optional (anonymization means changing real names of column to generated ones).

Processing of SQL statements such as "INSERT", "SELECT", "DELETE" uses common principles. All data from query are extracted and data from encrypted columns are encrypted. Also names of columns, which values are encrypted and used in processed statement, are modified according to their anonymized names. In some cases (for example, homomorphic encryption) changes in mathematical operation can be made. Responses for "SELECT" queries from database are decrypted if needed.

Correctness of performing "JOIN" operation inside DBMS depends on encryption properties. If encryption is deterministic, output column is single and both columns from each joining tables, have same key, no additional mechanisms are needed to perform "JOIN" operation.

It is very important to understand that some restrictions may apply to using full functionality of SQL language and different DBMS specific structures due to fact that proposed solution targets multi DBMS support, also various constraints can be caused by using order preserving or fully homomorphic encryption.

One of the restrictions is limited usage of "ALTER TABLE" construction is. As long as table altering doesn't affect encrypted columns, it can be performed, but adding or removing encryption from already existing table is unsupported. Another restriction is incompatibility of encryptions with several column constraints (e.g. "FOREIGN KEY").

This concludes basics principles and mechanisms of SQL queries processing in discussed approach. Main idea is to perform query analysis and modification, which include encryption of vulnerable user's data on client side, without affecting DBMS or adding any intermediate components.

III. ARCHITECTURE

Section 3 gives insight into secure cloud database architecture. As it was declared in previous sections, proposed solution does not use any DBMS components or any proxies in process of processing user's SQL queries and encrypting or decrypting data. All description of architecture applies to client's program module.

Client's program module consists of 4 basic components:

- Encryptions interfaces and encryption modules.
- Cryptographic metadata storage
- SQL queries processing component
- Database response's processor

Encryptions interfaces module provides two interfaces – "Key" and "Encoder". These interfaces define set of properties required for encryptions' correct work and interaction with other components. Due to "Encoder" and "Key" interfaces architecture and realization of entire solution does not depends on specific encryptions and is open to integration with other crypto algorithms.

Cryptographic metadata storage is responsible for storing information supporting SQL queries and DBMS replies processors. Among service information, the following values are kept in this storage:

- crypto keys for encryption of data in column
- map of real name of the column to anonymized names
- types of encryption used for column
- names of tables, where encrypted columns are located

Cryptographic metadata storage is an interface for retrieving and adding information about encrypted columns. Current realization means that all data are stored in file. File handling is a subject of future development. Expected solution is to store file encrypted on user's removable drives.

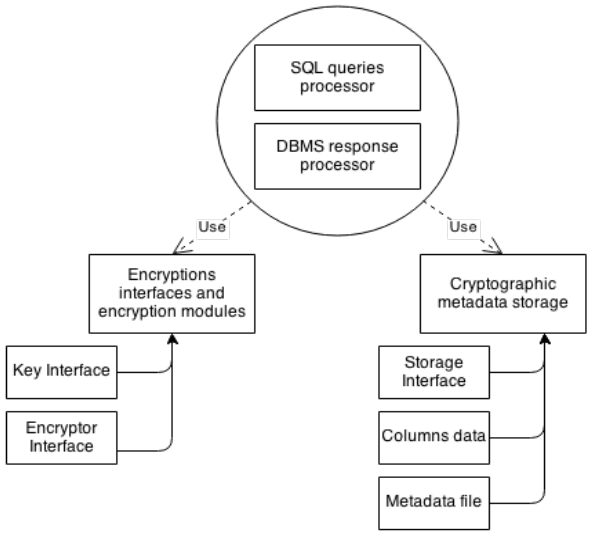


Fig. 4 Architectural overview

Core component of proposed solution is SQL queries processor; other components one way or another support its main function – analyze SQL queries and encrypt/decrypt data. SQL queries processing module consists of sub modules, each responsible for parsing exact SQL statement: "CREATE", "SELECT", "INSERT" and so on. These sub modules follow statements grammar to extract data, intended to be encrypted by user.

Last observed basic component is Database response's processor. Its main purpose is to detect encrypted columns in response, combine them (when multiple output columns correspond to single initial column), and to decrypt data to display them in suitable form to user. This module actively interacts with Cryptographic metadata storage, in order to correctly decrypt and modify response.

Architectural overview is summarized in Fig. 4. Two service components, Encryption and Crypto Storage modules, support SQL queries and database response processors in their main purpose – to manage all encryption and arithmetical transformations, while leaving secure cloud database's user with illusion of work with ordinary database.

IV. Encryptions

This section features description of encryptions that secure cloud database uses. We use three types of encryption: deterministic, order preserving and homomorphic encryptions.

Deterministic. Deterministic encryption provides strong security, it leaks only which encrypted values correspond to the same data value. In secure cloud database it can be used for storing password hashes, when no operations are conduct over data, but confidentiality is very important.

In proposed solution, we use proprietary developed deterministic block encryption [10].

Order Preserving. Order preserving (OP) encryption allows order relations between encrypted data items to be established, without revealing data itself. Such encryption can be used to protect salaries or other economical information inside secure database with possibility of performing order operations.

There are various OP encryptions, used in solution. For different types of data we can use different OP encryptions in single table; this provides extra resistance to crypto attacks. Only two encryption schemes will be discussed in this paper.

A. Arithmetic coding encryption

The first scheme, *based on arithmetic coding*, builds a representation of integer in an appropriate form. Assuming that integers are non-negative and do not require more than n bits, Then each number c is mapped to bit string $b=(a_1,a_2,\dots a_n)$ with first most significant bit. Let us define f as order-preserving mapping which maps string to some real number from interval $[0,1)$.

The simplest way to represent number s:

$$s=a_1\frac{1}{2}+a_2\frac{1}{2^2}+\dots+a_n\frac{1}{2^n}$$

$$s=\frac{c}{2^n}$$

One more way of representation:

We will seek for another representation for the number s.

We will use the arithmetic coding for representation f. Note, that s satisfies the equation:

$$2^n s=c$$

We will solve equation:

$$G(x)=2^n x-c=0 \quad (1)$$

Equation (1) has only one solution on the interval $[0,1)$. In case of bisection method for seeking solution, the source number s will be found in n steps. The main idea of arithmetic coding is that intervals can be split at random. In this case, the approximate solution of (1) can be found in fewer steps which allow arithmetic coding to compress data.

Let us describe splitting process. Let define

$$k=\frac{p}{p+q},l=\frac{q}{p+q},k+l=1,$$

where p and q are random natural numbers. Next, let us split interval $(0,1)$, in two pieces:

$$\left[0,\frac{p}{p+q}\right],\ \left[\frac{p}{p+q},1\right]$$

and calculate $G\left(\frac{p}{p+q}\right)$. If $G\left(\frac{p}{p+q}\right)>0$, we choose interval $\left[0,\frac{p}{p+q}\right]$ and return 0. Else if $G\left(\frac{p}{p+q}\right)<0$ then

interval $\left[\frac{p}{p+q},1\right]$ is chosen and *l* is returned. Chosen interval will be marked with $\left[a_1,b_1\right]$.

New interval is split into two pieces at ratio $\frac{k}{l}$. After that following steps are performed: calculating value of *G(x)* in new point and choosing one of new intervals according to sign of *G(x)*; marking new interval with $\left[a_2,b_{12}\right]$. Proceeding by induction, we compute the interval $\left[a_n,b_n\right]$. Its length is $k^r l^{n-r}$, where *r* is the number of zeros in *b*. b_n is rational, so it can be expanded in powers of up to m degree, where m is the smallest number satisfying the equation:

$$\frac{1}{2^m} < k^r l^{n-r}$$

This equation can be rewritten in

$$-m < r\log_2 k + (n-r)\log_2 l$$

form, or

$$m > -r\log_2\frac{p}{p+q} - (n-r)\log_2\frac{q}{p+q} =$$
$$= r\log_2\frac{p+q}{q} + (n-r)\log_2\frac{p+q}{q} =$$
$$n\log_2 p+q - r\log_2 p - (n-r)\log_2 q$$

Therefore *m* can be calculated if *b*, *k* and *l* are known. Estimation of *m* does not have depended on *r* for the general case. There is very rough approximation:

$$m > n\log_2(p+q)$$

If for b_n there is an appropriate estimate which takes *m* bits, then bit string $f(b)=(b_1,b_2,\ldots,b_m)$. Designed mapping preserves the order by construction.

So, $f(b)$ is a value, that is sent to DBMS, the key is set of intervals (p_i,q_i) where value i is in interval $[1;n]$. Value *n* depends on size of input data.

B. Radix encryption

The second scheme's, *based on different number systems*, basic idea is conversion of numbers from notation with one radix to another.

For the first step is necessary to obtain the vector of coefficients from number in first-radix representation. Next step is replacing in the current representation first radix with second chosen radix. At the last step performed when a subsidiary vector of nonnegative numbers is added to the vector of coefficients from the current number representation. Note that values of the sum of these vectors must be less than second radix and second radix is greater than first.

Having final representation with new radix and modified coefficients the result can be calculated as second-radix – decimal conversion. The secret key is consist of first, second radixes and subsidiary vector of nonnegative numbers.

To illustrate this idea, let us consider one iteration of encryption. There can be made several iterations.

Original number is $s \in N$.

Secret key is $p,q \in N$, $\langle b_0, b_1, \ldots b_{n-1}\rangle$, $b_i < q, b_i \in N$.

Steps of encryption:

1. Obtaining S's p-radix representation:
$$s=\alpha_0+\alpha_1*p+\alpha_2*p^2+\ldots+\alpha_{n-1}*p^{n-1}$$
2. Replacing p-radix with q-radix:
$$s'=\alpha_0+\alpha_1*q+\alpha_2*q^2+\ldots+\alpha_{n-1}*q^{n-1}$$
3. Adding subsidiary vector $\langle b_0, b_1, \ldots b_{n-1}\rangle$ to the vector of coefficients $\langle \alpha_0,\alpha_1,\alpha_2,\ldots,\alpha_{n-1}\rangle$.
$$s''=(\alpha_0+b_0)+(\alpha_1+b_1)*q+(\alpha_2+b_2)*q^2+\ldots$$
$$+(\alpha_{n-1}+b_{n-1})*q^{n-1}$$

where $\forall i: \alpha_i + b_i < q$.

The result of encryption is $w=s''$.

Process of decryption consists of following steps:

1. Obtaining w 's p-radix representation:
$$w=y_0+y_1*q+y_2*q^2+\ldots+y_{n-1}*q^{n-1}$$
2. Replacing q-radix with p-radix:
$$w'=y_0+y_1*p+y_2*p^2+\ldots+y_{n-1}*p^{n-1}$$
3. Subtracting vector $\langle b_0, b_1, \ldots b_{n-1}\rangle$ from the vector of coefficients $\langle y_0, y_1, \cdots y_{n-1}\rangle$:
$$w''=(y_0-b_0)+(y_1-b_1)*q+$$
$$+(y_2-b_2)*q^2+\ldots+(y_{n-1}-b_{n-1})*q^{n-1}$$

The result of decryption is w'', which is equal to s

The algorithm of encryption is correct and order preserving.

Modification of the considered scheme was used in the implementation. There are several iterations; also number of bits for values in the key can be specified in encryption module configuration.

This encryption has passed multiple tests and following results were measured:

Speed of encryption 125 Mbit / s

Speed of decryption 111 Mbit / s

(PC's configuration: Mobile Dual Core Intel Atom N570, 1666 MHz, 4 GB RAM, OS Windows 7).

Homomorphic. An encryption scheme is called fully homomorphic if it's able to evaluate an arbitrary function over ciphertexts. In this case decrypted value must match to a calculation result of the same function over plaintexts. The main feature of scheme [11] that is used in proposed secure cloud database is ability to define a strict upper bound of ci-

phertext size when performing calculations on it for both addition and multiplication.

V. Achievements and future Work

Proposed solution is not theoretical work only. We have implemented described crypto algorithms, measured encryption and decryption speed, optimized realization. Furthermore, prototype of client's program module has been successfully coded and tested. We used C++ and Boost's regular expressions to perform all described operations on SQL strings and databases' data selection. During tests on simple data scheme, with different "SELECT", "UPDATE", "DELETE" queries, slight (around 10-15%) overhead was detected, because of time elapsed for encryption/decryption.

Future development of proposed solution aims three main targets. First, is to develop and improve existing encryptions. Speed optimization is one of the most significant goals. Second target is to further develop of client's program module. This target includes functionality expansion, support of different DBMS, development of supporting modules (metafile encryption, authorization module).

The third aim is security improvement. Interest in analysis of encryption weaknesses and vulnerabilities [8, 9] is escalating, thus several measures can be taken to minimize risk of successful security breach. For example, to complicate frequency analysis, subsystem of phantom "SELECT" queries can be made in order to average number of queries to each column. Another idea for improving system's resistance to attacks is to add to columns garbage data that will be detected and ignored during decryption on client side. This method can change distribution of encrypted data massive inside DBMS, and as a result can make more difficult crypto attack on OP encrypted columns.

References

[1] Cloud Security Alliance. Top Threats to Cloud Computing V1.0 Cloud Security Alliance 2010.

[2] Cloud Security Alliance. The Notorious Nine. Cloud Computing Top Threats in 2013. Available: https://downloads.cloudsecurityalliance.org/initiatives/top_threats/The_Notorious_Nine_Cloud_Computing_Top_Threats_in_2013.pdf

[3] William R Claycomb, Alex Nicoll: Insider Threats to Cloud Computing: Directions for New Research Challenges CERT 2012.

[4] Privacy Rights Clearinghouse. Chronology of data breaches. Available: http://www.privacyrights.org/data-breach

[5] A. J. Feldman, W. P. Zeller, M. J. Freedman, and E. W. Felten. SPORC: Group collaboration using untrusted cloud resources. In Proceedings of the 9th Symposium on Operating Systems Design and Implementation, Vancouver, Canada, October 2010.

[6] P. Mahajan, S. Setty, S. Lee, A. Clement, L. Alvisi, M. Dahlin and M. Walfish. Depot: Cloud storage with minimal trust. In Proceedings of the 9th Symposium on Operating Systems Design and Implementation, Vancouver, Canada, October 2010.

[7] R. A.Popa, C.M.S.Redeld, N.Zeldovich, and H.Balakrishnan: CryptDB: Protecting Condentiality with Encrypted Query Processing proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles, 2011.

[8] L. Xiao, O. Bastani, I-Ling Yen: Security Analisys for Order Preserving Encryption Schemes, January, 10, 2012.

[9] R. Steinwandt, W. Geiselmann, and R. Endsuleit, "Attacking a polynomial-based cryptosystem: Polly Cracker," International Journal of Information Security, vol. 1, no. 3, pp. 143–148, 2002.

[10] Egorova V., Chechulina D., &Krendelev S. F. (2013) New View on Block Encryption (Unpublished) Available: https://db.tt/vnE9wfgj

[11] A A Zhirov, A., Zhirova, O., & Krendelev, S. F. (2013). Practical Fully Homomorphic Encryption over Polynomial Quotient Rings. In WorldCIS'13. London, UK.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 891–899
DOI: 10.15439/2014F186
ACSIS, Vol. 2

Order-preserving encryption schemes based on arithmetic coding and matrices

Sergey F. Krendelev
Lab of Modern Computer Technologies, Novosibirsk State University, Novosibirsk, Russia
Email: s.f.krendelev@gmail.com

Mikhail Yakovlev
Dept. of Information Technology, Novosibirsk State University, Novosibirsk, Russia
Email: m.o.yakovlev@gmail.com

Maria Usoltseva
Dept. of Information Technology, Novosibirsk State University, Novosibirsk, Russia
Email: m.a.usoltseva@gmail.com

Abstract—In this article we describe two alternative order-preserving encryption schemes. First scheme is based on arithmetic coding and the second scheme uses sequence of matrices for data encrypting. In the beginning of this paper we briefly describe previous related work published in recent time. Then we propose alternative variants of OPE and consider them in details. We examine drawbacks of these schemes and suggest possible ways of their improvement. Finally we present statistical results of implemented prototypes and discuss further work.

I. Introduction

Security is a fundamental issue solved by DBMS and cloud service designers. Using cryptographic algorithms to store confidential data in encrypted from isn't always the best solution. In case of relational (SQL) database it is impossible to process encrypted data on DBMS side. Generally two options are possible. The first one is to decrypt and process data on client side, which leads to significant traffic increase between DBMS and application due to general inability to single out the necessary data. Second option is to decrypt data on DBMS side, which is unsafe in case DBMS or cloud service isn't reliable.

The research speculates on ability to use special types of encrypting, which allow not only to safely store data, but also to perform a set of operations on it. Particularly, the research concerns order-preserving cryptosystems.

Definition (order-preserving function). Let A, B be sets with given order relation on each set. Function

$$F : A \to B$$

is said to preserve order if

$$x < y\,(x, y \in A) \Leftrightarrow F(x) < F(y).$$

Encryption based on using such functions is called order-preserving encryption (OPE). Assume there is a set of unique plaintexts $P = p_1, p_2, \dots p_{|p|}, p_i < p_{i+1}$. The corresponding encrypted values are represented as $C = c_1, c_2, \dots c_{|c|}, c_i < c_{i+1}$. Such encryption enables subset of SQL operations on encrypted data including, e.g. selection of encrypted values intervals.

II. Related work

In the last 10 years several scientific papers has published, which introduced different schemes of order-preserving encryption.

A. OPE based on pseudo-random number generator

One of the first approaches, represented in the research [1], suggests that integer number p is encrypted as follows:

$$c = \sum_{j=0}^{p} R_j,$$

where R_j is jth number generated by reliable pseudo-random generator. The drawbacks of this scheme are the memory footprint of encrypted values and possible overflow, resulting from calculation of ciphertexts for large plaintexts while working with built-in types. There is also the complexity of adding new plaintexts: for adding p_i where $p_i < p < p_{i+1}$ we need to re-encrypt values $p_j, j > i$.

That's why this method is ineffective for big numbers and, in some cases, the encryption result can be predicted. For instance, suppose μ average distribution of numbers, generated by the pseudo-random number generator. For uniform distribution on the interval $[1 \dots Max]$, $\mu = \frac{Max + 1}{2}$, then $f(x) = \mu x$ will approximate encryption function.

B. OPE based on polynomial functions

In research [2] a sequence of strictly increasing polynomial functions is used for encrypting integer values. These polynomials may be of first or second order. The secret key is polynomial coefficients. Sequence of polynomials is applied to initial number in a way that one function's output value is another function's input value. Decryption is done by applying inverse functions in reverse order. Sometimes it might be impossible to find the inverse function for a specific polynomial. Authors suggest using simple polynomials $f(x) = ax^b + c$ as they all have inverse

This research was performed in Novosibirsk State University under support of Ministry of education and science of Russia (contract no. 02.G25.31.0054).

functions $f(x)=\sqrt[b]{\frac{x-c}{a}}$. Besides, maximum degree of such polynomials, according to the authors, does not exceed 2, and a set of possible coefficients – $\{1\ldots32\}$. In this case decryption is a lot more complicated than encryption due to square root operation. As built-in integer types are implemented with fixed length, inevitable overflow errors appear. In this work authors suggest using logarithmic functions $f(x)=\log_2 x+c$. This implies working with non-integer types, therefore accuracy errors should be considered. To avoid accuracy and overflow errors this scheme requires rather complex selection of parameters.

C. Research by R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu

Research [3] speculates on encryption of data from subset of integers, authors also think it possible to use non-integer types represented as integers of the same size. This method encrypts data so that ciphertexts follow a certain distribution selected by the user. In order to generate encryption function they use all data that needs to be encrypted (if the database doesn't contain records, administrator has to add several possible records) and the list of possible sample distributions. Key is generated from all these samples. Besides, in order to simulate distribution data has to be split into buckets, in which linear interpolation is used. One of the main drawbacks is that the time of key generation time is linear in the size of database. And if the key has already been generated with adding of new records it might need to be regenerated along with re-encryption of data.

D. Research by A. Boldyreva N. Chenette, Y. Lee and A. O'Neill

A. Boldyreva in [5], [6] shows the connection between OPE schemes and hypergeometric (HG) и and negative hypergeometric (NHG) distributions. The connection allows to simulate OPE scheme through HG or NHG generator. There are effective algorithms of accurate random value generation of these distributions. One of order-preserving encryption features is that range of encryption function is always larger than input argument set (two different numbers can't be encrypted to the same numbers). Suppose encryption function maps a set $[1\ldots M]$ to a set $[1\ldots N]$, $N>M$. In order to encrypt the set of numbers $[1\ldots M]$ we need M random numbers from the set $[1\ldots N]$. Let us denote the set of these numbers by N_M. They have to be ordered and associated correspondingly with initial numbers. I.e. function $f:[1\ldots M]\to N_M$ maps the ith element of the set $[1\ldots M]$ to the ith element of the set N_M. Since we don't know, how many plaintexts have to be encrypted, we cannot determine the size of the set $[1\ldots N]$. Authors offer to generate the next element of the ordered set N_M only when it is necessary for ciphertext generation. This approach is called lazily sampling the function.

Let m be a plaintext from the set $[1\ldots M]$, g – a function, generating the set $[1\ldots N]$, x – number of elements selected into the set N_M after y steps. The number x is characterized by the hypergeometric distribution. Encryption starts with entire domain $[1\ldots M]$ and range $[1\ldots N]$. Let $y=\frac{\max(N)}{2}$ be a range gap. With a certain key and initial parameters the number x can be calculated. If $x<m$, we need to consider the points of the domain greater x and y, and less than or equal to x and less than or equal to y in case of $x\geq m$. As a result we get admissible set of ciphertexts.

E. Alternative OPE schemes

In our research we consider other ideas, which OPE can be based on. We propose two alternative OPE schemes, research problems of overflow and computational accuracy and try to increase cryptographic strength of schemes.

III. OPE SCHEME BASED ON ARITHMETIC CODING

This scheme builds a representation of integer in a suitable form. Representation preserves the order, so we can talk about order-preserving mapping. Suppose we consider positive integers requiring for their representation not more than n bits. Each number c is mapped to a bit string $b=(a_1,a_2,\ldots a_n)$, where a_1 is MSB. Let us define the order-preserving mapping f. We assume that the string defines certain real number $s\in[0,1)$. The simplest way to define it:

$$s=a_1\frac{1}{2}+a_2\frac{1}{2^2}+\cdots+a_n\frac{1}{2^n}.$$

In other words $s=\frac{c}{2^n}$.

Let us seek a different representation of the number s. To do it let us use ideas associated with arithmetic coding. The equation:

$$G(x)=2^n x-c=0 \qquad (1)$$

has only one solution on the interval $[0,1)$. In case of bisection method of solving the equation (1) we get the number s after n steps. The main idea of arithmetic coding is that segments can be split into parts arbitrarily. In this case approximate solution of equation (1) can be found in fewer steps. That allows us to achieve compression of data while using arithmetic coding.

Let us describe the process of segments splitting. Suppose, $\gamma=\frac{p}{p+q}, \mu=\frac{q}{p+q}$, where p, q are random natural numbers. Obviously, $\gamma+\mu=1$. Let us split segment $[0;1)$ into two parts $\left[0,\frac{p}{p+q}\right),\left[\frac{p}{p+q},1\right)$. After that, if $G\left(\frac{p}{p+q}\right)>0$, we choose $\left[0,\frac{p}{p+q}\right)$ and produce 0. In

case of $G\left(\frac{p}{p+q}\right)<0$, we choose $\left[\frac{p}{p+q},1\right)$ and produce 1. Let us denote the segment we chose by $[a_1,b_1)$.

New segment splits into parts in the ratio $\gamma:\mu$. According to the sign of $G(x)$ in the point, one of the segments is selected. Let us denote it by $[a_2,b_2)$. Proceeding by induction, we compute the interval $[a_n,b_n)$. Its length is $\gamma^r\mu^{n-r}$, where r is number of zeros in b. By construction of $b_n, b_n \in Q$, so it can be expanded in powers of $\frac{1}{2}$ up to m degree, where m is the smallest integer satisfying

$$\frac{1}{2^m}<\gamma^r\mu^{n-r}.$$

That condition may also be written as

$$-m<r\log_2\gamma+(n-r)\log_2\mu$$

or

$$m>-r\log_2\frac{p}{p+q}-(n-r)\log_2\frac{q}{p+q}=r\log_2\frac{p+q}{p}+$$
$$+(n-r)\log_2\frac{p+q}{q}=n\log_2(p+q)-r\log_2 p-$$
$$-(n-r)\log_2 q$$

Therefore, we can calculate m, if we know b,γ,μ. For universality we need m estimation not to depend on r value. E.g, we can require condition $m>n\log_2(p+q)$. Let us approximate b_n with bit string $f(b)=(\beta_1,\beta_2,\ldots,\beta_n)$. Obviously, this transformation preserves order.

General conclusion, which is using adaptive arithmetic coding, is to use different ratio on each step, when approximating the solution of equation (1). It allows making cryptosystem cryptographically stronger.

VI. Key generation

Key is set of ratios, which segments are divided in. Suppose current segment is split in a ratio $p_i:q_i$ on ith step. To be able to decrypt the cipher for n-bit number, the length of the segment obtained after decryption should be less than $\frac{1}{2^n}$. Maximum segment length, which can be obtained as a result of decrypting is $\prod_i\frac{\max(p_i,q_i)}{p_i+q_i}$. Consequently, the key generation algorithm has the following form:

1. Choose random ratio $p_i:q_i$.
2. Check the condition

$$\prod_i\frac{\max(p_i,q_i)}{p_i+q_i}<\frac{1}{2^n}.$$

If condition is satisfied, go to step 3, otherwise – to 1.

3. Complete key generation.

V. Encryption

Suppose we need to encrypt $Num\in\mathbb{N}$. On each algorithm iteration the interval $[a_i,b_i)$ is considered, where $a_0=0, b_0=1$. Let us examine the ith iteration of the algorithm.

Current segment $[a_{i-1},b_{i-1})$ is divided in a ratio $p_i:q_i$. Let point $s\in[a_{i-1},b_{i-1})$ separate it, i.e. $s=a_{i-1}+\frac{(b_{i-1}-a_{i-1})p_i}{p_i+q_i}$. If $s>\frac{Num}{2^n}$, then the output is 0, $a_i=a_{i-1}, b_i=s$. Otherwise the output is 1, $a_i=s, b_i=b_{i-1}$. It can be seen, that $\forall i\ \frac{Num}{2^n}\in[a_i,b_i)$ (by a_i and b_i selection).

After performing k iterations (where k is key size, i.e. the number of ratios) we obtain a bit sequence $\sigma=(\sigma_1,\ldots\sigma_k),\ \sigma_i\in\{0,1\}$, which is ciphertext for Num.

VI. Decryption

Suppose there is a bit sequence $\sigma=(\sigma_1,\ldots\sigma_k),\ \sigma_i\in\{0,1\}$, which is a ciphertext for unknown number Num. Just as in the encryption algorithm, at each iteration is considered the interval $[a_i,b_i), a_0=0, b_0=1$. Consider the ith iteration of the algorithm.

Current segment $[a_{i-1},b_{i-1})$ is divided in a ratio $p_i:q_i$. Let point $s\in[a_{i-1},b_{i-1})$ separates it, i.e. $s=a_{i-1}+\frac{(b_{i-1}-a_{i-1})p_i}{p_i+q_i}$. If $\sigma_i=0$, then $a_i=a_{i-1}, b_i=s$. Otherwise, $a_i=s, b_i=b_{i-1}$.

After performing k iterations (where k is key size) we obtain a segment $[a_k,b_k)$, and $(b_k-a_k)<\frac{1}{2^n}$ because of key selection. As $\frac{Num}{2^n}\in[a_k,b_k)$ number Num is uniquely decoded as follows:

$$Num=\lfloor 2^n a_k\rfloor+1,$$

where $\lfloor x\rfloor$ is the largest integer, which comes before x.

VII. Increasing cryptographic strength of the algorithm

If attacker does not know a secret key, all he knows is that encrypted value belongs to the interval $[a_0,b_0)$. But to make algorithm more secure, this segment can be hidden.

Let us choose an arbitrary strictly increasing function $f(x)$ so that $\lim_{x\to-\infty}f(x)<0$, $f(x)=+\infty$. Suppose $f(a)=0$, $f(b)=2^n$. Let us use function $f(x)$ to encrypt n-bit number $s: 0\le s<2^n$. Using modified arithmetic

encryption algorithm, we encrypt number x from interval $[a,b)$, where $f(x)=s$ (fig.1).

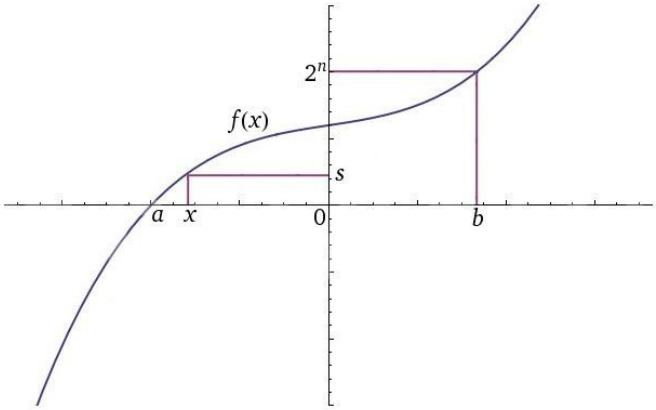


Fig.1 Increasing cryptographic strength of algorithm

In this case secret key is set of ratios p_i, q_i (key for modified arithmetic algorithm we already described earlier) and function $f(x)$, that was used in encryption.

To decode a cipher h we should find a and b such that $f(a)=0$ and $f(b)=2^n$ (they are unique, because $f(x)$ is strictly increasing function), decode x from interval $[a,b)$ and calculate $s=f(x)$. Now if attacker doesn't know a secret key, he doesn't have any information about s.

VIII. Implementation features

The algorithm described above can be implemented using only integers, which allows the implementation to eliminate rounding or computation errors. Nevertheless, the size of the numerator and denominator of fractions a_i, b_i and $x=\frac{Num}{2^n}$ can't be limited in general case, and therefore arbitrary-precision integer arithmetic should be used. As a result algorithm speed is not high enough.

The algorithm can be modified in order to accelerate implementation. On each iteration of the encryption transformation matrix $[a_i,b_i)\to[0,1)$ can be applied to segment $[a_i,b_i)$. Then it is obvious that $a_i\to 0, b_i\to 1$. Point x transforms by the following rule:

$$\text{if } \sigma_i=0, \text{ then } \frac{Num}{2^n}\to\frac{Num}{2^n}\frac{p_i+q_i}{p_i}, \quad (2)$$

$$\text{else } \frac{Num}{2^n}\to\frac{Num}{2^n}\frac{p_i+q_i}{q_i}. \quad (3)$$

It saves time for the calculation of the segment, and long arithmetic is only required for the storage of a fractional number x.

As a result the encoding iteration includes comparing points x and $\frac{p_i}{p_i+q_i}$ and following x recalculation according to the rules (2) and (3). In case of decrypting we need to know final segment $[a_k,b_k)$, so decoding iteration is implemented directly with a_i and b_i recalculation.

In order to remove one of the slowest parts of the implementation - arbitrary-precision integer arithmetic, we can apply rounding to the largest previous integer, which is used in many implementations of standard arithmetic coding[1].

The idea of this approach is that all fractional numbers such as a_i, b_i and x, which belong to the interval $[0,1)$ are multiplied by 2^r, where r is some power of two. Thus, arithmetic operations can be quickly performed. E.g. in case of 32-bit architecture, it is convenient to choose $r=32$. After multiplying the resulting fraction shall be rounded down to the nearest integer. The more is r, the smaller is rounding error, so it is advisable to choose large values for r.

On each iteration of the encoding (or decoding) algorithm initial segment $[a_0,b_0)=[0,2^r)$ decreases, because $[a_i,b_i)\subset[a_{i-1},b_{i-1})$, so rounding error increases and at some point it is impossible to determine if $x=\left\lfloor\frac{Num}{2^n}2^r\right\rfloor$ belongs to segment $[a_{i-1},s)$ or $[s,b_{i-1})$. To avoid it, the length of interval $[a_{i-1},s)$ and $[s,b_{i-1})$ after the process of rounding should be not less than 1. In other words,

$$s-a_{i-1}\ge 1,$$

$$\frac{(b_{i-1}-a_{i-1})p_i}{p_i+q_i}\ge 1,$$

$$b_{i-1}-a_{i-1}\ge\frac{p_i+q_i}{p_i}.$$

Similarly for the interval $[s,b_{i-1})$:

$$b_{i-1}-a_{i-1}\ge\frac{p_i+q_i}{p_i}.$$

Thus, in order to be able to perform the ith iteration, the current length of the segment $[a_{i-1},b_{i-1})$ should not be less than

$$\max\left(\frac{p_i+q_i}{p_i},\frac{p_i+q_i}{q_i}\right)\le p_i+q_i\le\max_i(p_i)+\max_i(q_i) \quad (4)$$

There is a special renormalization operation, which allows to increase the length of the current segment. It can be used in one of three cases:

1. Segment $[a_{i-1},b_{i-1})$, lies in the left half of the interval $[0,2^r)$, i.e. $[a_{i-1},b_{i-1})\subseteq[0,2^{r-1})$. In this case interval $[2^{r-1},2^r)$ is not involved in the encryption process anymore, so we can "bring closer" $[0,2^{r-1})$ segment twice, i.e. use transformation $[0,2^{r-1})\to[0,2^r)$. Then

[1]Moffat, Alistair. ACM Transactions on Information Systems / Alistair Moffat, Radford M. Neal, Ian H. Witten // – 1998. – Vol. 16, No. 3, July 1998.

$$a_{i-1} \to 2a_{i-1},$$
$$b_{i-1} \to 2b_{i-1},$$
$$x \to 2x.$$

2. Interval $[a_{i-1}, b_{i-1})$ lies in the right half of the interval $[0, 2^r)$, i.e. $[a_{i-1}, b_{i-1}) \subseteq [2^{r-1}, 2^r)$. Similarly to the previous case, we can use transformation $[2^{r-1}, 2^r) \to [0, 2^r)$. Then

$$a_{i-1} \to 2(a_{i-1} - 2^{r-1}),$$
$$b_{i-1} \to 2(b_{i-1} - 2^{r-1}),$$
$$x \to 2(x - 2^{r-1}).$$

3. Interval $[a_{i-1}, b_{i-1})$ lies in the central part of the interval $[0, 2^r)$, i.e. $[a_{i-1}, b_{i-1}) \subseteq [2^{r-2}, 3 \cdot 2^{r-2})$. Then let's "bring closer" twice the interval $[2^{r-2}, 3 \cdot 2^{r-2})$, i.e. use transformation $[2^{r-2}, 3 \cdot 2^{r-2}) \to [0, 2^r)$. Then

$$a_{i-1} \to 2^{r-1} - 2(2^{r-1} - a_{i-1}),$$
$$b_{i-1} \to 2^{r-1} - 2(2^{r-1} - b_{i-1}),$$
$$x \to 2^{r-1} - 2(2^{r-1} - x).$$

It is easy to notice that each case of the renormalization increases the length of the segment twice. If none of the renormalization conditions is satisfied, then the length of this interval is strictly greater than 2^{r-2}. According to this we can find the maximum size of p_i and q_i. If they are m-bit numbers, then (by (4)) current segment length after renormalization should be greater than

$$\max_i(p_i) + \max_i(q_i) \le 2^m + 2^m = 2^{m+1}.$$

Then $m_{max} + 1 = r - 2$, and $m_{max} = r - 3$. Larger size makes no sense to choose, because rounding leads to (r-3)-bit p and q numbers.

As a result we turn from calculations with fractional numbers with unlimited numerator and denominator to calculations of the fixed integers. This fact significantly increases the speed of algorithm. The price of this acceleration is a rounding error, which may reduce the strength of cryptographic algorithm.

IX. Statistics

Any order-preserving encryption can be represented as transformation from domain $[0, 2^{b_1})$ to some range $[0, 2^{b_2})$, where b_1, b_2 are the sizes of input and output data respectively, so it is possible to attack a cipher using linear approximation cryptograms on extreme values (fig 2).

In other words, cryptogram for a certain $x \in [0, 2^{b_1})$ is approximated with value $2^{b_2 - b_1} \cdot x$, and plaintext for cryptogram $m \in [0, 2^{b_2})$ – with value $\left\lfloor \frac{m}{2^{b_2 - b_1}} \right\rfloor$. Knowledge of the secret key is not required to calculate the approximate value, only sizes of the input and output data are needed. To defend against such attack we should increase an approximation error.

Therefore, in order to study the cryptographic strength of the algorithm statistics of "segment" lengths was examined. The end points of such "segment" a_i equals successive encrypting function values, i.e. $a_i = [f(i), f(i+1)]$. Obviously, the closer "segment" lengths distribution to uniform distribution, the smaller the error of approximation.

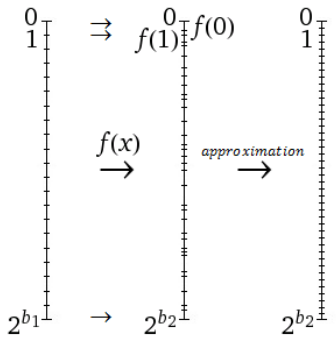


Fig.2 Linear approximation of cryptograms on extreme values

As a result, it was found that the "segment" lengths distribution is close to exponential (fig. 3, the abscissa indicates the segments lengths and the ordinate indicates the number of such segments). As too big error appears, this attack cannot be applied.

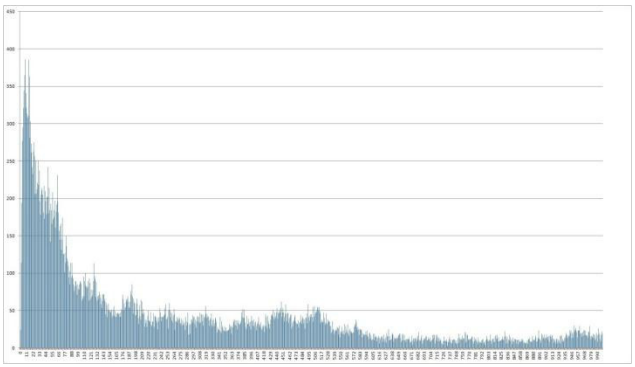


Fig.3 "Segment" lengths distribution of two-byte numbers

X. Matrix based OPE scheme

The suggested scheme allows to avoid overflow accuracy errors as it uses only integers. The scheme does not disclose any information about initial values of encrypted variables, except their order.

Tuple of three numbers (r, k, t) serves as a ciphertext. Suppose we need to encrypt positive integer x. In contradistinction to pseudo-random number generator method the first element of cryptogram is not the sum of numbers of random ascending sequence on the step x, but a number i of a step, such that sum of random numbers exceeds the initial number x on this step. The second element of the cipher is the difference between x and the sum of random sequence on (i-1)th step. It can also be

encrypted in a similar way, using the sum of random elements of a different sequence

$$g_1 < g_2 < \ldots < g_{r-1} \le x < g_r .$$

Obviously, $x - g_{r-1} < g_r - g_{r-1}$. Therefore, to encrypt residual $x - g_{r-1}$ we need to calculate the sum of the sequence elements, which will be limited by $g_r - g_{r-1}$. Instead of pseudo-random sequence elements sum we can use strictly increasing function with special features. The sum of the sum sequence of matrix elements is suggested to be used as strictly increasing function:

$$\sum_{i=0}^{n}\sum_{j=0}^{n} a_i ,$$

where a_i is an elements of matrix A_i of size $n \times n$. We use a certain sequence of matrices

$$A_1, A_2, \ldots, A_r, \ldots$$

First, a certain matrix A_1 is chosen. Each successive matrix can be calculated using some transformation of the previous one. As such transformation we research power function.

$$A_1, A_1^2, \ldots, A_1^r, \ldots$$

In order to prevent the increasing of matrix elements after multiplication operation, we consider matrix elements are from finite field $\mathbb{Z}_m$. Residual $x - g_{r-1}$ is hidden with the help of elements of the matrix, calculated on the rth step. Elements are summed up randomly, and the number k of step, where the sum s_k exceeds $x - g_{r-1}$, is the second element of the tuple. The residual obtained at this stage is the third element of the tuple. Let's look at the scheme in greater detail.

Encryption scheme under consideration is symmetrical, i.e. it uses private key for encryption and decryption of data. In the research the input data of cryptographic algorithm are integer non-negative numbers from 0 to 2^n, where $n = 16, 32, 64$, which corresponds to the size of built-in types unsigned short int, unsigned int, unsigned long int, unsigned long long int of C++ language, which the scheme is implemented on. The result of cryptographic algorithm is a tuple of three numbers (r, k, t), where $r, k, t \in \mathbb{Z}_{\ge 0}$.

XI. KEY GENERATION

The private key in the scheme is non-degenerate matrix $A(\det A \neq 0)$ over the finite field $\mathbb{Z}_m$, $m \ge 2$, with $n \times n$ size, where $n \in N, n \ge 2$ and a certain permutation of σ elements of matrix with such size, called matrix traversal.

A. Non-degenerate matrix generations

Standard way of non-degenerate matrix generation is to generate matrix with random elements and check its non-degeneracy. If the determinant of obtained matrix is equal to zero, generation shall be repeated until this condition is violated.

For small-size matrices this method is quite effective. However with increase of matrix size, multiplication operation takes more and more time, which reduce speed of private key generation.

Direct methods of determinant calculation can be based on permutations sum, or Laplace expansion of smaller degree determinants. However such methods are rather ineffective, because they require time complexity $O(n!)$ for n degree determinant calculation. There are other methods with fewer operations, e.g. Gauss method modification, where matrix is transformed to the form of echelon and determinant is calculated as product of multiplication of diagonal elements. Complexity of this method constitutes $O(n^3)$. If there is available multiplication algorithm of two square matrices of size n in time $M(n)$, where $M(n) \ge n^a$, for certain $a > 2$, then matrix determinant can be calculated in time $O(M(n))$. However if random matrix is degenerate, calculations shall be repeated until we get non-degenerate matrix.

This research suggests using knowingly non-degenerate matrix generation approach, based on the following linear algebra theorem.

Theorem. Square matrix A with non-zero principal minors can be presented in the form of LU lower triangular matrix L, whose main diagonal consists of non-zero elements, times upper triangular matrix U with units on the main diagonal.

Since lower triangular matrix L contains a single diagonal, its determinant equals one. Upper triangular matrix U determinant equals multiplication of elements on main diagonal. Using the property of determinant, we obtain $\det(A) = \det(LU) = \det(L)\det(U) = \det(U)$.

Thus, in order to generate non-degenerate matrix A, it is enough to generate matrices L and U, compliant with the above properties and find their multiplication result. Computational complexity of matrix product by definition is $O(n^3)$, however there are more effective algorithms, for instance, Coppersmith–Winograd algorithm, performing multiplication in $O(n^{2.3727})$.

B. Generation of matrix elements permutation

Algorithm uses permutation of matrix elements, imitating its traversal, a specific order in which to trace the elements of a matrix. Cryptographic robustness cannot be achieved through simple row traversal of matrix elements or row traversal with a shift. The traversal should be generated randomly. Traversing function shall go over all matrix elements (through each element only once), i.e. the function has to be bijective. For instance, we can use affine transformation.

It shall be presented through randomly generated non-degenerate in $\mathbb{Z}_m$ matrix B of size 2×2 (B shouldn't be identity matrix, or else the function implements row traversal with a shift) and vector $c = (c_1, c_2)$, where $c_1, c_2 \in \mathbb{Z}_m$. Matrix non-degeneracy provides inverse mapping, used in decryption. New element coordinates are calculated as follows:

$$\begin{pmatrix} u \\ v \end{pmatrix} = B\begin{pmatrix} i \\ j \end{pmatrix} + \begin{pmatrix} c_1 \\ c_2 \end{pmatrix},$$

where coordinates (i, j) mean initial row traversal $(1,1), (1,2), \ldots (2,1), (2,2), \ldots (n, n)$ and $(u(i, j), v(i, j))$ is traversal, used in algorithm.

Generally, to determine matrix traversal we can use any effective algorithm for random permutation of set of elements generation, e.g. Fisher–Yates shuffle algorithm (Knuth shuffle), with time complexity reduced to $O(n)$, along with its updated versions – Durstenfeld and Sattolo algorithms, using less memory space. When using high quality unbiased random numbers generator, algorithm guarantees equiprobability of permutations.

XII. Encryption

Let's consider data encryption procedure. Suppose we need to encrypt $x \in \mathbb{Z}_{\geq 0}$. At first private key (A, σ) is generated. The first element r of ciphertext is calculated as follows:

$$\sum_{i=1}^{r-1} d(A^i) \leq x < \sum_{i=1}^{r} d(A^i),$$

where

$$d(A) = \sum_{i=0}^{n^2} a_i$$

is the sum of elements a_i of matrix A. Power operation A^i is performed in $\mathbb{Z}_m$.

In order to determine the second element k of the cipher, we need to calculate the sum

$$S = \sum_{i=0}^{k} a_i',$$

where a_i' are elements of matrix σA^r, such that

$$S \leq x - \sum_{i=1}^{r-1} d(A^i).$$

Difference

$$t = x - \sum_{i=1}^{r-1} d(A^i) - S$$

is the third element of the cipher.

XIII. Decryption

Suppose the input of decryption algorithm is a tuple (r, k, t) and the key is (A, σ). First, we shall calculate matrix A^r. Using known permutation σ of matrix elements, we calculate

$$S = \sum_{i=0}^{k} a_i',$$

where a_i' are elements of matrix σA^r. Adding to the sum S the third element of ciphertext t, we obtain a certain number h, which, according to encryption procedure, equals

$$h = t + S = x - \sum_{i=1}^{r-1} d(A^i).$$

Using the first element r of ciphertext we calculate the sum

$$\sum_{i=1}^{r-1} d(A^i)$$

and obtain the number x, that was encrypted

$$x = h + \sum_{i=1}^{r-1} d(A^i).$$

XIV. Constructinon correctness

First of all we have to verify cipher uniqueness. Thus, the encrypted number should be the same number as decrypted. We also shall prove that such encryption is order-preserving.

Theorem 1. If with encrypted N there is cipher Enc(N), then with decryption Dec(Enc(N)) there is number N.

Proof. We need to prove that mapping, determining encryption algorithm, is bijective. In bijective mapping every element of one set corresponds to only one element of another set, along with inverse mapping with the same property.

Bijective mapping properties:

1. A function $f : A \to B$ is bijective if and only if it is invertible, that is, there is a function $g : B \to A$ such that $g \circ f$ is identity function on A and $f \circ g$ is identity function on B.
2. The composition $g \circ f$ of two bijections is again a bijection.

Function $f_1(x, A)$ calculates the parameter r. Since $\det A \neq 0$, then function

$$\sum_{i=1}^{j} d(A^i)$$

is strictly increasing, i.e.

$$\sum_{i=1}^{j} d(A^i) < \sum_{i=1}^{j+1} d(A^i).$$

Obviously, for any x from $\mathbb{Z}_{\geq 0}$ number r is found uniquely. Function $f_2(x, A, \sigma)$ calculates the parameter k. Since

$$x < \sum_{i=1}^{r} d(A^i),$$

hence we get

$$x - \sum_{i=1}^{r} d(A^i) < d(A^r).$$

Therefore, there is the only k, such that

$$S = \sum_{i=0}^{k} a_i', \quad S \leq x - \sum_{i=1}^{r-1} d(A^i),$$

where a_i' are elements of matrix σA^r. It is obvious that permutation of matrix A^r elements does not violate this condition. If functions

$$\sum_{i=1}^{r-1} d(A^i)$$

and S are bijective, then function $f_3(x, A, \sigma)$, calculating t, is also bijective.

Theorem 2. If with encryption of N_1 with the key K we obtained cipher Enc_1, and with encryption N_2 with the same key we obtained cipher Enc_2, and $N_1 > N_2$, then $Enc_1 > Enc_2$.

Proof. Let us say that tuple $\overline{A}_1 = (a_{11}, \ldots, a_{1m+1})$ is greater than vector $\overline{A}_2 = (a_{21}, \ldots, a_{2m+1})$, if $a_{1j} > a_{2j}$, where j is first position number, such that $a_{1j} \neq a_{2j}$, for all j, from m+1 to 0.

We shall consider $x_1, x_2 \in \mathbb{Z}_{\geq 0}$, such that $x_1 < x_2$. According to encryption procedure

$$\sum_{i=1}^{r_1-1} d(A^i) \leq x_1 < \sum_{i=1}^{r_1} d(A^i),$$

$$\sum_{i=1}^{r_2-1} d(A^i) \leq x_2 < \sum_{i=1}^{r_2} d(A^i).$$

Consequently,

$$\sum_{i=1}^{r_1-1} d(A^i) \leq \sum_{i=1}^{r_2-1} d(A^i), \quad \sum_{i=1}^{r_1} d(A^i) \leq \sum_{i=1}^{r_2} d(A^i).$$

These functions are strictly increasing, so $r_1 \leq r_2$. As

$$S_1 \leq x_1 - \sum_{i=1}^{r_1-1} d(A^i),$$

$$S_2 \leq x_2 - \sum_{i=1}^{r_2-1} d(A^i),$$

therefore, $S_1 \leq S_2$. By definition of S function, since all the matrix elements are non-negative, then $k_1 \leq k_2$.

$$t_1 = x_1 - \sum_{i=1}^{r_1-1} d(A^i) - S_1,$$

$$t_2 \leq x_2 - \sum_{i=1}^{r_2-1} d(A^i) - S_2.$$

If $\sum_{i=1}^{r_1-1} d(A^i) = \sum_{i=1}^{r_2-1} d(A^i)$ and $S_1 \leq S_2$, i.e. $r_1 = r_2$ and $k_1 = k_2$, then $t_1 = t_2$.

XV. Cryptographic strength

A. Ciphertext-only attack

Provided that the attacker doesn't have key parameter, i.e. matrix size n and modulus m, then ciphertext-only attack does not seem possible. Matrix size of $n \times n$ modulus m can be presented with m^{n^2} variants.

Suppose we extract a certain ciphertext (r, k, t). Suppose the size of secret matrix is $n \times n$, and operations are performed modulus m. As the matrix is non-degenerate and non-identity, minimum sum of its elements equals $(n+1)$ and maximum $n^2(m-1)$. Thus, initial number lies within the range from $(r-1)(n+1)$ to $rn^2(m-1)$. The sum S lies within the range from 0 to $k(m-1)$ and may be presented with m^k variants.

B. Chosen-plaintext attack

All order-preserving encryptions are vulnerable to such kind of attack. Let us encrypt a sequence of numbers $x_1, x_2, \ldots$. We shall consider ciphertexts of form $(r_i, 0, 0)$. With these values of encryption function, the sum is

$$\sum_{i=1}^{l} d(A^j) = x_i,$$

Hence for subsequent plaintexts $i+1, i+2, \ldots$. We can find residuals

$$\sum_{i=1}^{l} d(A^j) - x_i$$

encrypted with the last two elements of the tuple.

Next we shall consider ciphertexts of form $(r_i, k_t, 0)$. Since the last tuple element equals zero, corresponding residual equals

$$\sum_{j=1}^{k_t} a_j,$$

where a_j are elements of matrix A^{r_1} in accordance with secret traversal. Thus, examining successively $k_1 = 1, k_2 = 2, \ldots$, we can determine the elements of matrix A^r. During security enhancement, matrix elements are summed up randomly, with $n^2!$ possible variants. Besides, if $r > 1$, matrix root is an expensive operation.

Another way to enhance security is applying strictly increasing function value f(x) to initial number x before encryption procedure. For instance, f(x) can be Ax+B, where $A, B \in \mathbb{Z}$, $A > 1$, $0 < B < A$, and instead of x there is initial number. The procedure of encryption substitution is as follows. Suppose, $Num \in \mathbb{Z}_{\geq 0}$ a number, that needs to be encrypted. Let us randomly choose number A, which is the part of the key. Allowed value range for B is limited by A, therefore, the bigger number A, the more possible variants there are for B. Number B is also randomly chosen, and adding number B enables two neighbor numbers to grade into numbers, whose difference is a random number, i.e $Num_2 - Num_1 = A(x+1) + B_2 - (Ax + B_1) = A + B_2 - B_1$.

Thus, in order to find the number, following number Num_2, we need to sort out at worst $A + B_3 - B_2 = 3A - 2$ numbers, which enhances the construct security.

Next we substitute x for Num, calculate Ax+B and encrypt the deduced number. After decryption, we need to perform integer division of decrypted number by A in order to get initial number. This scheme doesn't require storing of coefficient B within the key and, therefore gives opportunity to use different B values for any numbers, including equal numbers.

XV. Statistics

The plots depicting dependence of r, k, t parameters on size of initial unencrypted values are provided below. We used secret matrix of size 10 × 10 with elements from $\mathbb{Z}_{10}$ and parameter $A = 2^8$ for encrypting 1000 subsequent 8-bit numbers.

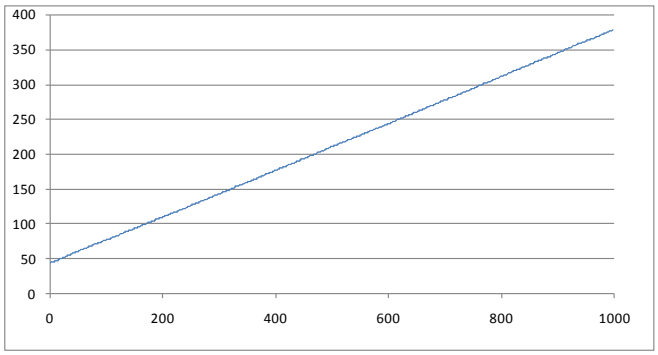


Fig.4 Dependence of r parameter on plaintext size

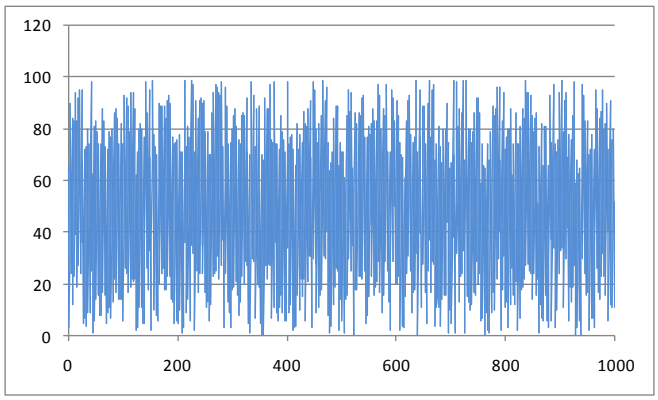


Fig.5 Dependence of k parameter on plaintext size

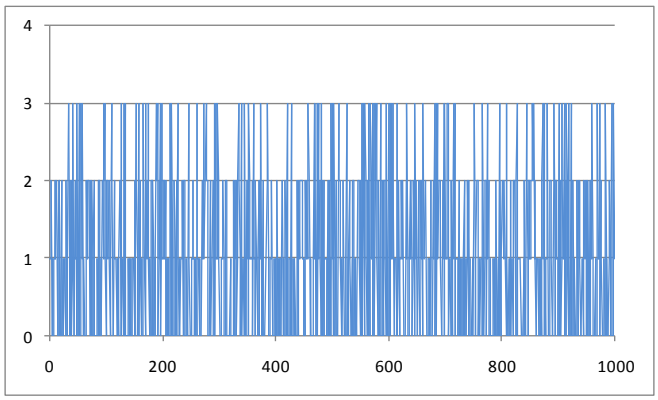


Fig.6 Dependence of t parameter on plaintext size

XVI. Further work

Currently we are examining other ways of attacking given OPE schemes and increasing cryptographic strength of them. We are also looking for optimal parameters of schemes, which provide an acceptable balance of speed and security.

There are several possible approaches to accelerate implementation such as using GPU for calculations, using specified matrix implementation, using faster standard algorithms required in schemes, etc.

After appropriate results are achieved, these schemes will be embedded as third-party libraries in security database service implemented in our laboratory.

References

[1] G. Bebek. Anti-tamper database research: Inference control techniques, Technical Report EECS 433 Final Report, Case Western Reserve University, 2002.

[2] Gultekin Ozsoyoglu, David A. Singer, and Sun S. Chung. Anti-Tamper Databases: Querying Encrypted Databases. In 17th Annual IFIP WG 11.3 Working Conference on Database and Applications Security, 2003, http://dx.doi.org/10.1109/ICDEW.2006.30

[3] R. Agrawal, J. Kiernan, R. Stikant, and Y. Xu, Order-preserving encryption for numeric data, ACM SIGMOD International Conference on Management of Data, pp. 563-574, 2004.

[4] G. Amanatidis, A. Boldyreva, and A. O'Neill, Provably-secure schemes for basic query support in outsourced databases, Working Conference on Data and Applications Security, pp. 14-30, 2007.

[5] A. Boldyreva, N. Chenette, Y. Lee, and A. O'Neill, Order-preserving symmetric encryption, Eurocrypt, pp. 224-241, 2009.

[6] A. Boldyreva, N. Chenette, and A. O'Neill, Order-Preserving Encryption Revisited: Improved Security Analysis and Alternative Solutions, Crypto11, 2011.

[7] Schneier, B., Wiley J.: Applied Cryptography Second Edition 1996 ISBN 0-471-11709-9..

[8] D. Boneh and B. Waters, Conjunctive, subset, and range queries on encrypted data, TCC, 535-554, 2007.

[9] H. Hacigumus, B.R. Iyer, C. Li, and S. Mehrotra, Executing SQL over encrypted data in the database-service-provider model, ACM SIGMOD Conference on Management of Data, 2002.

[10] M. Halloush and M. Sharif, Global heuristic search on encrypted data (GHSED), International Journal of Computer Science Issues (IJCSI), 1:13-17, 2009.

[11] Raluca A. Popa, Frank H. Li, Nickolai Zeldovich, An Ideal-Security Protocol for Order-Preserving Encoding. IEEE Symposium on Security and Privacy 2013, http://dx.doi.org/10.1109/CISS.2012.6310814

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 901–910
DOI: 10.15439/2014F77
ACSIS, Vol. 2

A Comparison between Business Process Management and Information Security Management

Gaute Wangen
Norwegian Information Security Laboratory
Gjovik University College
Teknologiveien 22, 2802 Gjovik, Norway
Email: gaute.wangen2@hig.no

Einar Arthur Snekkenes
Norwegian Information Security Laboratory
Gjovik University College
Teknologiveien 22, 2802 Gjovik, Norway
Email: einar.snekkenes@hig.no

***Abstract*—Information Security Standards such as NIST SP 800-39 and ISO/IEC 27005:2011 are turning their scope towards business process security. And rightly so, as introducing an information security control into a business-processing environment is likely to affect business process flow, while redesigning a business process will most certainly have security implications. Hence, in this paper, we investigate the similarities and differences between Business Process Management (BPM) and Information Security Management (ISM), and explore the obstacles and opportunities for integrating the two concepts. We compare three levels of abstraction common for both approaches; top-level implementation strategies, organizational risk views & associated tasks, and domains. With some minor differences, the comparisons shows that there is a strong similarity in the implementation strategies, organizational views and tasks of both methods. The domain comparison shows that ISM maps to the BPM domains; however, some of the BPM domains have only limited support in ISM.**

Keywords: Information Security, Information Security Risk Management, Business Process Management, BPM Methodology Framework, ISO/IEC 27001, ISO/IEC 27002, ISO/IEC 27005, NIST SP 800-39

I. Introduction

INFORMATION technology and systems play a crucial role by supporting the organization in achieving its goals and objectives. The main goal of information security (IS) is to secure the business against threats and ensure success in daily operations, and aid the businesses in reaching the desired level of reliability and productivity through ensuring integrity, availability and confidentiality [1]. We define the main profit of IS risk management (ISRM) as maximizing long term profit in the prescence of faults, conflicting incentives and active adversaries.

Business Process management (BPM) is a discipline that combines knowledge from information technology and management sciences and centers on business processes [2]. It is used to represent business processes (BP) for analysis and improvement purposes [3], [4]. The main goals of BPM is to align the organization's business processes to the organization's mission, goals and objectives and improve efficiency to create a competitive advantage [3], [5].

Some of the existing information security frameworks mention risk management (RM) of business processes in some form, e.g. ISO/IEC 27005:2011 defines BPs as a primary asset [6], and NIST SP 800-39 suggests RM of Mission/Business Process as tier 2 in the multi tier organization-wide risk management model [7]. While the purpose of both IS management (ISM) and BPM is similar, to map and improve organizational performance in their own way, they remain two different disciplines that require two different sets of skill.

In this paper, we investigate the similarities and differences between BPM and ISM, and explore the obstacles and opportunities for integrating the concepts of ISM and BPM. The BPM methodology framework [8] by BPTrends as described by Harmon [5] and Mahal [3] represents the main sources used to describe BPM, and we use the ISO/IEC 27000-series [6], [9], [10] and NIST SP 800-39 [7] to describe ISM.

A. Problem Description

While it can be said that the scope of ISM is turning towards BPs security, BPM and ISM remain two different disciplines and are most of the time regarded as separate activities [11]. However, the disciplines mutually affect each other's objectives, e.g. re-engineering a BP will often have security implications, and introducing an information security control is likely to affect the BP flow. In addition, the impact of a materialized security risk will usually affect the business. A different set of skills is required to risk manage a BP than an IT-system; one requires knowledge of BPM methods, and the other technical insight in information security. In addition, there exists several types of BPs, ranging in abstraction level, from value chain at the very top of the organization, to work instruction & procedures [3], [5], see Fig. 1. People employed at different levels of the organization, perceive and worry about different risks [12], and focus on a variety of different goals in their work efforts [5]. The difference in abstraction makes it likely that one ISRM approach designed for a low level BP is not likely to be applicable for risk managing the higher abstractions, such as value chain or core processes. Hence, there is a need to make sure that IS and BPM activities are aligned. Very little has been published in terms of investigations regarding to what extent IS and BPM guidelines and methods are well aligned, overlapping or in conflict. The aim of this paper is to contribute towards the filling this gap.

The remainder of this paper is structured as follows; In Sect. II, we present related work. In Sections III & IV we

978-83-60810-58-3/$25.00

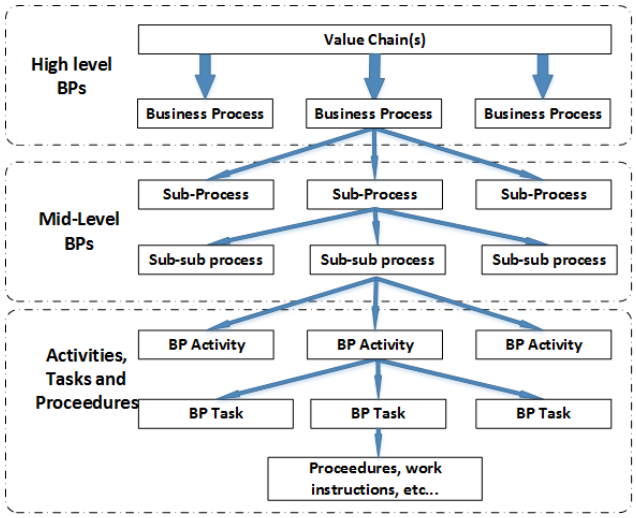


Fig. 1. Example of a Business Process Hierarchy.

introduce relevant IS and BPM concepts used in this article. Sect. V introduces the research method. Sections VI, VII & VIII presents comparisons of ISM and BPM and discussions of findings. The three areas of comparisons are Lifecycles, Organizational Views & corresponding tasks, and Domains. Conclusion and Future Work are given in Sect. IX.

II. Related Work

Much of the published research within combination of BPM and ISM focus on risk analysis of BPs; Milanovic et. al. [13] presents a framework for modeling BP availability. The framework takes into account services, the underlying ICT-infrastructure and people, and has a special focus on dependencies between these layers. Jallow et.al. [14] present a framework for risk analyzing BPs, using modeling activities and Monte Carlo analysis for calculating risks and forecasts. Asnar and Massacci [15] takes the GRC management approach to information security, and presents a method for analyzing and designing security controls in an organizational setting using BPs. Zoet et.al. [16] introduces the different kinds of risk that affect a BP and establishes the relationship between operational risk, compliance risk, internal controls and business processes. Zoet et.al. also present an integrated framework for dealing with RM and compliance from a BP perspective. Taubenberger and Jurens [17] suggest to improve security processes by using BP models to move away from probabilities.

There also exists approaches for risk managing BPs; In 2000, Kokolakis et.al. [18] presented a paper discussing the use of BPM for IS. The authors argue that the asset-based approach of ISRM treats IS as an add-on feature aiming to minimize the overhead cost. The authors suggests that the combination of BPM and IS-SAD (information security analysis and design) techniques can be used for security re-engineering of a BP, and integration of IS. The authors presents an overview of existing BPM approaches and requirements they should support to be used in ISRM.

Jakoubi and Tjoa [11] introduce a reference model for considering information within the BPM and RM domains. The authors argue for a stronger interweaving between RM and BPM, and present an approach for reengineering business processes as risk-aware. Herrmann and Herrmann [19] introduces the MoSS BP (Modeling Security Semantics of Business Processes) frame, based on object-oriented process models. The authors introduce several security properties and correlations between security requirements and BP elements, together with the following general approach to risk managing business processes, the three first steps focus on identificaiton of: (i) Business Processes and their actors. (ii) And valuation of assets and their security levels. (iii) Security requirements - and responding vulnerabilities and threats. While the two last steps address risk analysis and treatment: (iv) Assessment of risk. (v) Proposal, design and implementation of countermeasures.

AURUM [20] supports the NIST SP 800-30 standard [21], and is a framework for addressing IT risks which utilizes business processes for RM. AURUM prioritizes BPs based on importance, and derives the important assets from the BP. The method then continues to determine asset importance and conducts risk analysis based on Bayesian threat networks.

Ozkan and Karabacak [22] suggests that process modeling can be used to ease the use of risk analysis methods and move the IS focus from hardware and software over to IT processes. The authors suggests using process modeling to model the activities of the information processing and to determine the scope of the risk analysis. The CERT Resilience Management Model v 1.0 [23] (CERT RMM) is an approach for handling the challenge of operational resilience in day to day operations. The notion is that organizations deliver services that are supported by BPs' which are further supported by assets.

III. IT Governance, Information Security Risk & Management

Gregory [24] state that *"The purpose of IT governance is to align the IT-organization with the needs of the business"*. IT governance involves a series of activities to achieve this goal such as creating IT-policy, internal prioritizing between e.g. mission, objectives and goals, program and project management [24]. It also includes the responsibility for managing risks appropriately, and verifying that resources are used responsibly [7].

A. Information Security Management (ISM)

Generally, the main goal of information security is to secure the business against threats and ensure success in daily operations by ensuring confidentiality, integrity, availability (CIA) and non-repudiation [1]. Information can be present in many forms within the organization, it may be stored on a physical medium, be in the form of paper, or it can be an employee's knowledge and experience. Common for all these is that they are all valuable assets to an organization and

their security needs assurance. One of the main components of ISM is to establish a security program, often referred to as an information security management system (ISMS). The ISMS is a collection of security related documents often with the company wide security policy as the main document. The purpose of the ISMS is to ensure CIA through management of the organization; by choosing and implementing the appropriate security measures and controls. These measures can be chosen from e.g. the ISO/IEC 27002 [10], which is a standard consisting of security measures and how to implement them. The ISMS can be implemented following a Plan-Do-Check-Act (PDCA) cycle of continuous improvement [1], [6], see Fig. 2.

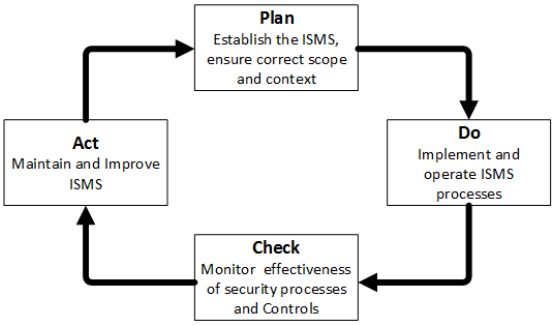


Fig. 2. Plan Do Check Act-phases of ISMS implementation as described in ISO/IEC 27000:2009 [1].

The security documentation of the ISMS is represented by a top-level security policy, generally founded in the organization's mission, vision, goals, values and objectives. Further represented by topic/issue-specific policies, standards, procedures and routines.

B. Information Security Risk Management (ISRM)

There exists several definitions of risk, ISO/IEC 31000:2009 [25] standard explains risk as *the effect of uncertainty on objectives*, and *Risk management* as a set of activities and methods applied in an organization to manage and control the many risks that can affect achievement of business goals. Hence, the main goal of ISRM is to maximize the long term profit, and optimally manage risks presented by potential failures, conflicting incentives and active adversaries.

A risk assessment is the *overall process of risk analysis and risk evaluation* [1], and risk analysis (RA) is the *systematic use of information to identify sources to estimate the risk* [1]. Risk evaluation is the *"process of comparing the estimated risk against given risk criteria to determine the significance of the risk"* [1].

ISO/IEC 27005:2011 [6] is a standard specialized for ISRM and defines the formal process of managing risks as an iterative process of reviewing and monitoring risks, including: context establishment, risk assessment, communication and treatment to obtain risk acceptance [6]. Risks for information systems are generally analyzed by using a probabilistic risk analysis (PRA) [6], [21], where impact to the organization (e.g. loss if a risk occurred) and the probability of the risk occurring is calculated. Probability calculation in ISRM has previously recieved critisism for relying too much on subjective estimates, and being too much like guesswork [24], [26], [27]. Risk evaluation uses the results from the analysis, and if the risk is found unacceptable, risk treatments are implemented, which consists of choosing a strategy and measures for controlling undesirable events.

C. Context Establishment for ISRM

The term "Context Establishment" is from the ISO/IEC Risk Management standard 27005 [6], and defines both the external and the internal parameters that must be considered when managing risks. The internal context for ISRM will usually be a product of different factors, such as IT systems, stakeholders, governance, contractual relationships, culture, capabilities, business objectives, and others. Examples of relevant external factors for establishing context are external stakeholders, external environment, laws and regulations, and other factors that can affect the organizations objectives.

Many established ISRM methods center around assets, the *NIST Specification for Asset identification* [28] uses three main classes of information system related assets; (i) Persons, (ii) Organization, and (iii) Information Technology. In addition, it provides nine sub-classes of assets of Information technology. In contrast to this, ISO/IEC 27005:2011 uses two primary asset classes; (i)Business processes & activities" and (ii)Information, with supporting assets: (i) Hardware, (ii) Software, (iii) Network, (iv) Personell, (v) site, and (vi) organization's structure.

A control can exist as automatic or manual, an automatic control performs its function with little or no human interaction, and a manual control requires a human to operate it, and generally fall within three major categories [24]: (i) Physical - represents controls that are found in the physical world, such as fences, doors with locks, and laptop wires. (ii) Technical - represents controls that are implemented in the form of information systems, they are usually in a logical form, such as a firewall, antimalware, and computer access control. (iii) Administrative - represents controls in form of e.g. policies and procedures that forbid certain activities, such as the IS policy.

The 14 Control Clauses and security domains from ISO/IEC 27002:2011 [10] and ISO/IEC 27001:2013 [9] are:

1) Information Security Policy - Top level documented security objectives for the whole organization, determined by management.
2) Organization of Information Security - IS Roles and Responsibilities, and IS management in general.
3) Human Resources Security - IS requirements and controls for recruitment of staff, terms of employment, security awareness training and process for termination.
4) Asset Management - The management and application of hardware and software assets, and classifying and handling of information.

5) Access Control - Effective password, privilege and user management on operating systems, applications and within networks.
6) Cryptography - Controls for securing CIA of information using encryption.
7) Physical and Environmental Security - Securing the human and system environment, including entry controls, power and cabling security.
8) Operations Security - Ensure CIA of operations and facilities.
9) Communications Security - Key security aspects of managing systems securely, such as backups, antivirus, media and laptop security
10) System Acquisition, Development and Maintenance - Secure development of software and maintenance of systems to maintain ongoing security
11) Supplier Relationships - Protect the organization from security breaches caused by third parties.
12) Information Security Incident Management - The reporting, recording, management and review of security incidents.
13) Information security Aspects of Business Continuity Management - Determine requirements, plan and training for response in the event of disasters.
14) Compliance - Ensuring compliance with legal requirements, including IPR, computer misuse and privacy legislation.

IV. Business Process Modelling and Management

A business process (BP) is a set of activities within an organization whose objective is to produce a desired result [29]. A process is, in short, "How work gets done" [3], and work is the *"exertion of effort directed to produce or accomplish something"* [4]. The purpose of modeling a BP is to describe the logical order and dependence, such that the practitioners can achieve a comprehensive understanding of the process [29]. A process generally has some sort input and transforms this into an output, e.g. a manufacturing process will take raw material as input, process this material, and output a product. We borrow the explanation from Mahal [3]:"a process is triggered by an event, governed by some rules using relevant knowledge, and executed through people using enabling technology and supporting infrastructure, such as facilities". A common abbreviation used to describe the components of a BP is IGOE - Inputs, guides, outputs and enablers [3], [5].

Besides from documenting processes, BPM can be used to facilitate large scale software developments to support BPs, BP analysis and improvement re-engineering [29]. The top-level representation of the BPM approach seen in Fig. 3.

A. The BPM Lifecycle

The BPM lifecycle represent the key activities in BPM. There is no uniform view of the number of BPM-LC phases [30]. Ko [31] state that there are many views of what steps the BPM life cycle actually consists of, and presents van der Aalst et.al.'s (2003) [32] view due to succinctness and relevance. Van der Aalst (2013) [2] has also published a newer review of the key activities in BPM after [31] was published. Wetzstein et.al. [30] present a general version of the BPM-LC. An analysis of the different lifecycle steps from [2], [30], [32], [33] show that they have the following steps in common:

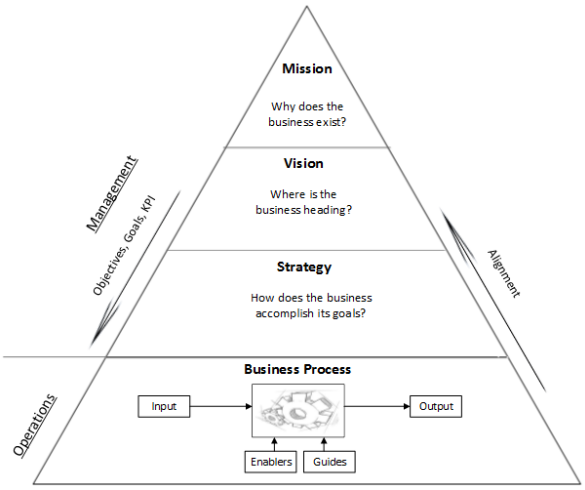


Fig. 3. Connection between Mission, Vision, Strategy and Business Processes. *Based on Mahal [3]*

1) Modeling and Design - Map/re-design or create a process model for analysis and/or enactment.
2) System Configuration & Implementation - Configure the system and implement the process model for enactment.
3) Enact/Execution - Deploy and execute the BP model using set configuration control and support concrete cases.
4) Monitor/Analyze - Analyze a process model studying the BP and/or event logs.
5) Manage/Diagnosis - Adjust/improve process, reallocate resources, manage large collections of BP models.

B. BPTrends Associates' BPM Methodology

The BPM Methodology Framework [8] is a best practices framework that provides a view of BPM sorted into three levels with associated steps. The framework recognizes the variety of goals at the different levels of the organization. The framework sorts the different levels into enterprise, process and implementation levels. The *Enterprise* level centers on corporate strategy, and focus on understanding and modeling BP architecture, definig performance measures, governance systems, aligning enterprise capabilities and prioritizing efforts. The main ongoing task consist of managing enterprise processes.

The *Process* level runs process improvement projects, where modeling, redesign and improvement of existing processes is in focus, taking processes from AS-IS to TO-BE. The main day-to-day tasks are BP execution and management.

The *Implementation* level focuses on designing human, software and information systems to implement BPs. It consists of

various IT and HR methodologies that are used for maintaining resources and continuous improvement.

C. BP Domains

Fig. 4 illustrates the BP domains, and shows how the different aspects of business support the BP, which ultimately determines enterprise performance. The general purpose of a BP is to transform an input to a desired output. The enterprise delivers value to its stakeholders and customers, and enterprise performance can be described using a set of measurable goals and objectives. KPIs provide the mechanisms for measuring performance. Information, knowledge and insight is what fuels the BP. The BP execution transforms the information into knowledge which is applied to create solutions. The "Guides" manages and controls the input/output transformation [3]. Put in the information security language; Guides are generally about governance and controls. The "Enablers" are the reusable resources of an organization that support the BP in transformation of input to output [3]. We leave inputs and outputs out of scope in this comparison. An explanation of the BP domains in the hexagon is as follows [3]: *Guides* provide

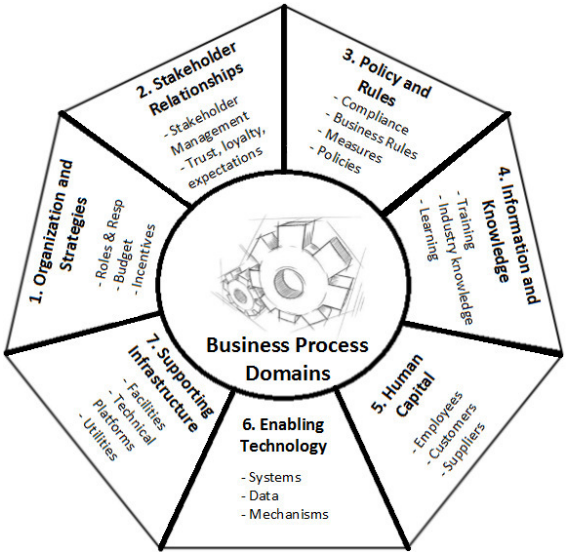


Fig. 4. Illustration of Guides and Enablers that contribute to the BP. *Based on [3], [34]*

governance, stakeholder expectations, direction, funding, rules and compliance restraints to the business process.

1) Organization and Strategies - Constitutes the organization's governance and its support structure. This domain covers consistent management, cohesive policies, processes, roles and responsibilities. It also includes organizational alignment and strategy development to achieve vision and deliver results.
2) Stakeholder Relationships - This domain constitutes both the external and internal stakeholders of the organization. It covers stakeholder management of expectations, trust and loyalty. The stakeholders are people who have vested in the success of the organization and can benefit from its performance.
3) Policy & Rules - Constitutes the business policies and rules of the organization, and are established to ensure compliance and mitigate risks through appropriate controls. The policies provide a decision-making framework at all levels of the organization.
4) Information and Knowledge - Encompasses training, learning and industry knowledge.*Defined as a guide in [3], [5].*
   Enabler are the reusable resources of an organization that support the BP in transformation of input to output. Enablers provide execution capabilities for the BP.
5) Human Capital - Constitutes of the people who enables the process, namely employees, customers, and suppliers. For the employee it is about their competence, which encompasses of a combination of knowledge, skills and behavior. Capable people are essential to optimally executing a process.
6) Enabling Technology - Constitutes of the technology that enables the BP. Includes information technologies such as business applications, data stores, and mechanisms such as production lines, robots, and engineering equipment.
7) Supporting Infrastructure - Constitutes of production facilities, technical platforms, communications, utilities and energy, and other infrastructure. Can also be considered as the capital asset of the organization.

V. Method

The primary research method adopted in this work is analytical. This article uses theoretical comparisons and mapping of BPM and ISM, for each BP activity we look for a corresponding IS activitiy. Similarly, for each IS activity, we look for a corresponding BP activity. This process will identify the intersection of BP and IS as well as what activities that are missing if BP and IS "compliance" is desired.

Following Ko et.al. [33] we start at the very top of the abstraction levels, comparing the generic lifecycles of BPM and ISM. Staying at a high level of abstraction, we compare organization/risk views and corresponding tasks. Lastly, we do a domain comparison of the BPM and ISM.

VI. A Comparison of ISM and BPM Lifecycles

The purpose of this section is to look for similarities and possibilities of integration between the top-level implementation strategies of the ISMS and BPM. We compare the high level steps of the plan-do-check-act (PDCA) lifecycle of the ISMS [9] and BPM lifecycle (BPM-LC) and look for common ground. Both cycles represent high-level views of the general activities of each approach. As there is no uniform view on the BPM-LC, we use the steps summarized in this article. We make the assumption that the ISMS lifecycle is compliant with the original PDCA-cycle, and compare the BPM-LC with the PDCA cycle as described by Moen and Norman [35].

TABLE I
A COMPARISON OF THE GENERIC PDCA STEPS AND THE BPM LIFECYCLE

PDCA steps/ BPM Lifecycle	**Plan**	**Do**	**Check**	**Act**
1. Modeling	X			
2. Implement/ Sys Config		X		
3. Enact/ Execution		X		
4. Analyze/ Monitor			X	
5. Manage/ Diagnosis				X

Table I shows that the generic BPM-lifecycle is loosely related to a PDCA notion of continuous improvement. A further comparison of the ISMS and BPM lifecycle approaches shows:

1) *Plan - Modelling:* The Plan-phase in ISMS is applied to establish context and scope the ISMS, together with planning for ISRM. In BPM, the steps in the modelling-phase maps existing BPs and plan/re-design BPs for enactment and analysis. Similar for both approaches is that they both establish the context and scope in this phase, the BPM uses BPs while IS uses e.g. an asset-based approach to establish organizational context. ISO/IEC 27005:2011 [6] names BPs as one of two primary assets, which may open for a combined approach of BPM context establishment.
2) *Do - "System Configuration" & "Implementation and Enact/Execution":* The steps in the Do-phase of the ISMS-lifecycle consists of implementing the processes associated with the ISMS. Usually in form of implementing risk treatment plans as a result of the ISRM program.
   The system configuration and implementation-phase in BPM implements designs by configuring process aware information systems and the underlying infrastructure. While the Enact/Execution phase executes and enacts the BP model. Both these BPM-phases correspond to the Do-phase in the PDCA cycle. Similar for both the ISMS and BPM lifecycles is that they both *implement* plans.
3) *Check - Analyze/Monitor:* This ISMS-phase monitors and reviews the effectiveness of implemented security process and residual risks. While the BPM-phase monitors and analyzes BPs for optimization. Both the IS and BPM lifecycles utilizes this phase for *monitoring and analysis* of the implemented processes.
4) *Act - Manage/Diagnosis:* The ISMS act-phase is mainly used to improve existing security processes based on analysis. The Manage and Diagnosis phase is utilized to adjust and improve BPs based on results from the previous lifecycle phase. This phase is also used to reallocate resources between BPs and manage large collections of BPs. Common for both lifecycles is implementing improvements based on analysis results from the previous phase.

We see from this comparison that the approaches are closely related; they are both founded on the PDCA principle, and the main tasks of each step is also similar.

VII. A COMPARISON OF ORGANIZATIONAL VIEWS

People employed at different levels of the organization both perceive and worry about different risks [12], which is also similar for the different concerns in the BPM hierarchy [5]. There is therefore a difference in what kind of information is needed to conduct tasks for both BPM and ISM at different levels of the organization. The purpose of this section is therefore to compare and map the organizational views and associated tasks presented in BPM and ISRM literature.

The BPM Methodology Framework represents a view of BPM sorted into levels including enterprise, process and implementation level, with recommended BPM steps per level (see [3], [5], [8]). NIST SP 800-39 [7] presents three different tiers for ISRM views, the comparison between the organizational views can be seen in table II.

TABLE II
A COMPARISON OF ORGANIZATIONAL VIEWS FROM THE NIST SP 800-39 [7] AND BPM METHODOLOGY FRAMEWORK [3], [5], [8]

Abstraction level	**Category**	**Multitier Org -Wide RM**	**BPM Methodology Framework**
Level 1	**Perspective**	Organizational	Enterprise
	Management	Top management	Organizational Management
	Main Tasks	Strategic risk management	Corporate Strategy in BPM, Supply chain
Level 2	**Perspective**	Mission/ Business Processes	Processes
	Management	Middle management	Process Management
	Main Tasks	RM of M/BP	Process Improvement
Level 3	**Perspective**	Information Systems	Implementation Level
	Management	Operations	Activity Management
	Main Tasks	Tactical Risk	Implementation of Information systems

The top-level comparison of the organizational views reveal a strong similarity. This is not surprising as one of NIST SP 800-39's main focus areas is securing BPs. Looking closer at the comparison we see a strong similarity in perspectives, tasks and responsibilities at each level:

- *Level 1* - We consider top management and organizational management to represent the same point of view. Both have a top-level management focus and are concerned

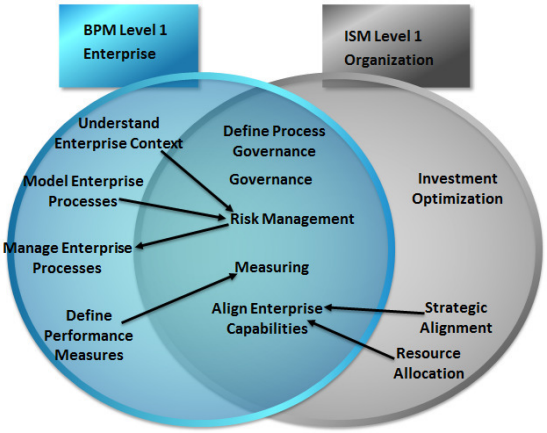


Fig. 5. Illustration of common BPM & ISM Level 1 tasks. Arrows indicate that a task is part of an activity, and that conducting the individual task will not complete the activity.

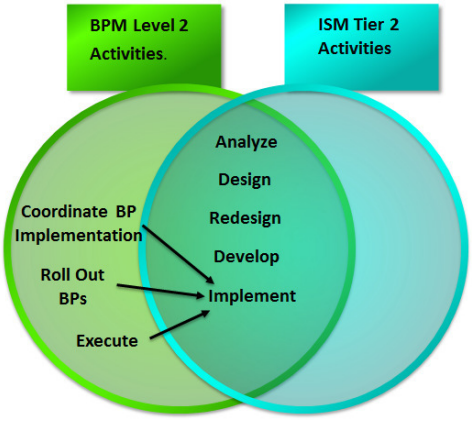


Fig. 6. Illustration of common BPM & ISM Level 2 tasks. Arrows indicate that a task is part of an activity, and that conducting the individual task will not complete the activity.

with governance and strategy tasks. We use the BPM tasks as described by [3], [8] to compare the subtasks from ISRM. Since there is no standardized steps per level from NIST SP 800-39, we analyzed and summarized the following steps for level 1 [7]: (i) Governance - assign roles and responsibilities to provide strategic direction, mission and objective achievement, risk management and resource usage, (ii) Strategic Alignment - of mission and business functions, (iii) Execution of Risk Management - frame, assess, respond to, and monitor risk (iv) Resource Allocation - of RM resources, (v) Measuring - monitoring and reporting RM metrics to ensure aligment, and (vi) Investment optimization - based on RM in support of organizational objectives.

The results from the comparison between ISRM and BPM level 1 sub-tasks can be seen in Fig. 5. The comparison show that the NIST RM function cover both *understanding the enterprise context* and *modelling enterprise processes* under Risk Framing, both activities necessary to conduct ISRM. However, the RM function only contributes to *Managing enterprise processes* which also includes activities such as establishing a BP services charter [3]. The same can be said for *Strategic aligment* of risk decisions, which is a part of completing *Aligning enterprise capabilities*, but does not complete the task. Our comparison show that there is no support for *Investment optimization* based on risk management at this level in BPM. Conducting resource allocation of RM resources will not complete any BPM tasks, but is a part of the *aligning enterprise capabilities* activity.

Comparing the other way, we see that there is no single ISRM Level 1 subtask to understand enterprise context and model enterprise context, but both are necessary steps in *executiion of RM* task. While *defining performance measures* is a part of the ISRM activity *measuring*, we cannot say that completing the BPM activity also completes the ISRM task. However, managing enterprise processes also measures processes and allocates resources.

- *Level 2* - Middle management and Process management are descriptions of the same responsibilities and points of view, only differentiated by organizational structure (e.g. matrix based for process management, or traditional department-based organization for middle management) [5]. Both have a BP perspective, and are concerned with modeling, prioritizing and re-designing processes. Further comparison of level 2 subtasks is seen in Fig. 6, where we see that the Level 2 BPM activities resemble the BPM lifecycle. As there are no standard steps in NIST SP 800-39, we have summarized the following level 2 steps from [7] for developing Risk-aware BPs: (i) Design - Existing BP (AS-IS), (ii) Develop - secure BP (TO-BE), (iii) Implement - secure BP. The standard also suggests to develop Secure Enterprise Architecture (EA) as a Level 2 task, which comprises maximizing effectiveness of BPs and information resources. We regard this task as present in all the BP-ISRM steps, and therefore do not count it as a standalone task.

  Our understanding of the NIST SP 800-39 tier two steps is that implementing a secure BP includes the BPM tasks "Coordination" (preparing for implementation), "Rolling out" and "Executing". Which means that all the BPM activities are supported in the ISRM approach. Comparing the other way shows that the "Analyze" and "Redesign" activities are covered by the ISRM steps, and that three remaining tasks together complete the ISRM "Implement" activity.
- *Level 3* - The information systems and implementation level perspective represents the operations and activity management point of view. The processes are found at the lower levels in the BPM hierarchy (see section 1), and represents where "the rubber meets the road" [3]. We consider this to represent the same management and perspective. Although both BPM and ISRM share the

operations view, they have slightly different concerns; IS is focused on securing information systems from tactical risks and managing controls, while BP is concerned with designing systems to implement with BPs.
As BPM employs several methodologies at this level, and the BPTrends associates' BPM Methodology framework does not extend to software and HR development [8], we have no standard tasks to compare to the ISRM. Mahal [3] mentions that one commonly used BPM method at this level is the software development lifecycle (SDLC). Risk managing the SDLC is also the main approach in NIST SP 800-39. Althought concrete HR-strategies are not present in the NIST standard, it does discuss organizational culture and it does also discuss the topic of trust, which we can not see mentioned in the BPM literature.

VIII. A COMPARISON OF ISM AND BPM DOMAINS

The main objective of this section is to compare the ISM and BPM domains to investigate if all control objectives can be integrated using BPM, and that all relevant aspects of BPM are covered in the control objectives. IS encompasses many fields related to information technology and systems, the ISO/IEC-standards in the 27000-series are industry standards and we use them as representatives of what must be covered to achieve IS (Notably ISO/IEC 27001 & 27002 [9], [10]). Therefore, to compare BPM and ISM approaches we use the 14 security domains and controls from ISO/IEC 27002 [10]. We mutually compare the IS domains to the domains of BPM defined by Burlton [34] and refined by Mahal [3] and Harmon [5].

A. Summary of Comparison, ISM and BPM

This section contains a summary of the integration results of IS into BPM. Table III shows a high level comparison of how the control clauses are supported by the BPM-domains.

The comparison of the ISM and BPM domains shows that we can integrate the security clauses and controls into the BPM domains of enablers and guides, and model them as BPs. An example is the implementation of the controls from the Information security incident management-security categories, illustrated in Fig. 7, which shows how the guides and enablers support the process.

One significant finding was that the domains of BPM does not directly consider internal or external attackers. This can in some cases be considered as a weakness of BPM as it concerns itself availability and integrity of BPs. RM is suggested as a supporting practice in development of the guides "Policy & Rules" [3]. The attacker might be considered as a part of general RM, but RM is such a wide discipline that it is likely to mean different things to different people [27].

BPM also presents a bit different view of assets; as the context, represented by BPs, is established before identifying the assets. In traditional ISRM, the situation is the other way around; first the asset that needs protection is identified, and then the context is modeled around the asset. Besides from

TABLE III
SUMMARY OF BPM-ISM AND ISM-BPM COMPARISON.
LEGEND: - "X" MARKS HOW THE ISM DOMAINS ARE COVERED AND CAN BE IMPLEMENTED IN THE BPM DOMAINS.
- "0" MARKS WHICH ISM DOMAINS SUPPORT BPM DOMAINS AND WHERE.

Domains BPM / ISM Domains	1.Organization and Strategy	2.Stakeholder Relationships	3.Policy and Rules	4.Information and Knowlegde	5.Human Capital	6.Enabling Technology	7.Support Infrastructure
1.Information Security Policy	X 0		X 0	X 0	X		
2.Organization and IS	X 0	0	X 0	0	X	X	X
3.Human Resources Security	X		X 0	X 0	X 0	X	
4.Asset Managment	X		X 0	X 0	X	X 0	X 0
5.Access Control			X 0	X 0	X 0	X 0	X
6.Cryptography			X 0	0	X	X 0	X
7.Physical and Environment Security			X 0	0	X	X 0	X 0
8.Operations security	X		X 0	X 0	X	X	X
9.Communications sec			X 0	X 0	X	X	X 0
10.System acquis, developm and mainte			X 0	X 0	X	X 0	X
11.Supplier relations		X (0)	X 0	X 0	X 0	X	X
12.IS incident man	X		X 0	X 0	X	X 0	X
13.IS aspect of BCM	X		X 0	X 0	X	X 0	X
14.Compliance			X 0	0	X	X	

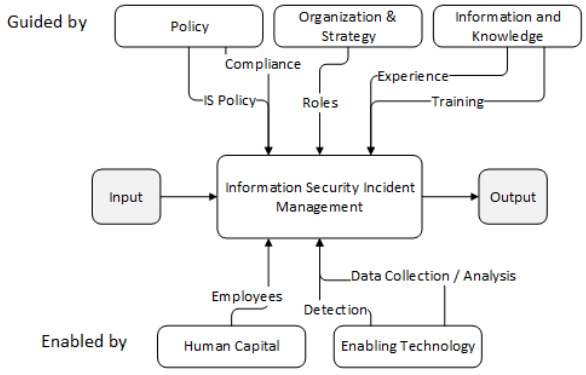


Fig. 7. The illustration shows how the IS Incident Management control can be modelled within the BP domain.

knowledge, intangible assets are not reflected in the BPM domains.

Another result that can be seen from the comparison is that the enabler "Human Capital", which generally represents employees, are needed to implement and operate every ISM control domain. However, the comparison show that out of fourteen control domains, only four are related to the security of human capital.

B. Summary of Comparison, BPM and ISRM

This section contains a summary of the integration results of BPM into ISM. Our comparison shows that the controls in ISO/IEC 27002:2013 are properly scoped to address four of the seven BPM domains. The enabler-domains were all addressed, but there were issues when addressing three of the Guide-domains:

1) Organization and Strategies: ISO/IEC 27001, section 5.1 a) emphasizes IS policy's compatibility with the organizations strategic direction, however, it is not mentioned in one of ISO/IEC 27002's 114 controls that the IS policy should be aligned with business. We can make the assumption of alignment from clause control objective 5.1, which is to provide management direction and support for IS in accordance with business requirements and compliance. The control itself state that the policy should be defined and approved by management. This points to a difference in perspective between the two disciplines, where BPM hammers organizational alignment of BPs as one of its main mantras.

2) Stakeholder Relationships: Nurturing both internal and external stakeholder relationships is an essential component of BPM; stakeholder identification, steering expectation, ensuring trust and loyalty are essential to BPM success [3], [5], [36]. Section "6.1 Internal organization" [10] covers some stakeholder groups (without using that term), as authorities and "special interest groups" are both types of stakeholders. The suggested controls put emphasis on maintaining contact with these stakeholders. However, these external groups are per BPM definition not important stakeholders, ISO/IEC 27001:2013 address the stakeholder needs in section 4.2 *Understanding the needs and expectations of interested parties*, but we can not see this reflected in the control objectives. The ISMS-program risk failing if key stakeholders lose interest, several instances of failure due to not having sufficiently powerful allies is highlighted in [22]. Although not completely neglected by IS, there is a clear gap between how much emphasis BPM and ISRM put on stakeholder management.

3) Information and Knowledge: It is a given that information is covered by all of the security domains. In BPM, information is utilized as knowledge by employees to fuel BPs [3], and knowledge is generally possessed by employees. The "Return of Assets"- security control (8.1.4) briefly mentions knowledge; *In cases where an employee, contractor or third party user has knowledge that is important to ongoing operations, that information should be documented and transferred to the organization.* This reflects a preventive control at the *end* of an employment. Capturing knowledge presents difficulties, as the interviewer must know exactly what questions to ask and the subject must be cooperative and willing to communicate the information in a comprehensive way.

This brings up the question if an ISRM process can identify and protect critical knowledge. Knowledge is viewed as an intangible asset [37], but e.g. is not included in the asset overviews in [28] or [6]. However, loss of availability due to lack of knowledge is a plausible IS risk (e.g. during incident handling), combined with the importance of knowledge in BPM, makes it an important business area to secure. Depending on the skill of the analyst, knowledge runs the possibility of being overlooked by ISO/IEC 27005:2011 and asset-based approaches.

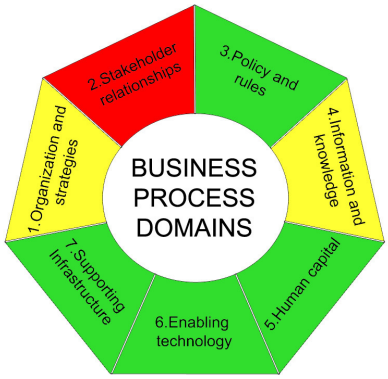


Fig. 8. Heatmap indicating how well ISM covers the BPM domains, green signals no issues, red signals significant issues.

IX. Conclusion

We have shown in this article that both the top-level BPM and ISM approaches are based on a Deming-cycle (PDCA) of continuous improvement, and that the main tasks of each step are similar.

We have shown that there is a strong similarity between the BPM Methodology framework and the ISRM standard NIST SP 800-39, as both approaches uses similar organizational views, only applying different names. We have also shown that the tasks and goals of each level are similar, with some key differences: the tier/level 1 ISRM approach does not include an activity for managing enterprise processes, and BPM does not include risk based investment optimization and trust-issues.

When comparing BPM and ISM domains we found that the ISM tasks can be supported by BPM, but that BPM does not include the concept of internal or external attackers. Further we found that ISO/IEC 27001/2 standards emphasized, but not controlled that the IS policy was aligned with business requirements. We also found a large gap between how much emphasis ISM and BPM put on stakeholders. Where BPM have fully adopted the principles of stakeholder management and recognized its importance, there is no real approach adopted in ISM to address stakeholders. We also found that the need for securing knowledge possibly is underestimated in ISM.

A. Future Work

As our findings are theoretical, we suggest further validation of the results from this article. This article has also shown that there is some common ground between BPM and ISM, and this warrants further investigation to determine if a joint approach is feasible. This work has revealed the potential for further research concerning stakeholder management in information security.

ACKNOWLEDGEMENTS

The authors of this paper thanks the anonymus reviewers for their valuable comments and suggestions. The PHD-student is sponsored by COINS Research School for IS.

REFERENCES

[1] *Information technology, Security techniques, ISMS, Overview and vocabulary*, International Organization for Standardization Norm, ISO/IEC 27000:2009. [Online]. Available: http://dx.doi.org/10.3403/30236519

[2] W. M. van der Aalst, "Business process management: A comprehensive survey," *ISRN Software Engineering*, vol. 2013, 2013. [Online]. Available: http://dx.doi.org/10.1155/2013/507984

[3] A. Mahal, *How Work Gets Done: Business Process Management, Basics and Beyond*. Technics Publications, LLC, 2010.

[4] R. Damelio, *The basics of process mapping*. Taylor & Francis US, 2011.

[5] P. Harmon *et al.*, *Business process change: A guide for business managers and BPM and Six Sigma professionals*. Morgan Kaufmann, 2010. [Online]. Available: http://dx.doi.org/10.1016/b978-012374152-3/50043-4

[6] *Information technology, Security techniques, Information Security Risk Management*, International Organization for Standardization Std., ISO/IEC 27005:2011.

[7] G. Locke and P. Gallagher, "800-39 nist sp, managing information security risks - organization, mission, and information systems view," National Institute of Standards and Technology, Tech. Rep., 2008.

[8] "The bpm methodology framework," http:\\www.BPTrends.com, visited April 2014.

[9] *Information technology - Secuirty techniques - Information security management systems - Requirements*, International Organization for Standardization Norm, ISO/IEC 27001:2013. [Online]. Available: http://dx.doi.org/10.3403/30192065

[10] *Information Technology, Security Techniques, Code of Practice for Information Security Management*, International Organization for Standardization Std., ISO/IEC 27002:2013. [Online]. Available: http://dx.doi.org/10.3403/30186138

[11] S. Jakoubi and S. Tjoa, "A reference model for risk-aware business process management," in *Risks and Security of Internet and Systems (CRiSIS), 2009 Fourth International Conference on*. IEEE, 2009, pp. 82–89. [Online]. Available: http://dx.doi.org/10.1109/crisis.2009.5411973

[12] A. G. Kotulic and J. G. Clark, "Why there aren't more information security research studies," *Information & Management*, vol. 41, no. 5, pp. 597–607, 2004. [Online]. Available: http://dx.doi.org/10.1016/j.im.2003.08.001

[13] N. Milanovic, B. Milic, and M. Malek, "Modeling business process availability," in *Services-Part I, 2008. IEEE Congress on*. IEEE, 2008, pp. 315–321. [Online]. Available: http://dx.doi.org/10.1109/services-1.2008.9

[14] A. Jallow, B. Majeed, K. Vergidis, A. Tiwari, and R. Roy, "Operational risk analysis in business processes," *BT Technology Journal*, vol. 25, no. 1, pp. 168–177, 2007. [Online]. Available: http://dx.doi.org/10.1007/s10550-007-0018-4

[15] Y. Asnar and F. Massacci, "A method for security governance, risk, and compliance (grc): a goal-process approach," in *Foundations of security analysis and design VI*. Springer, 2011, pp. 152–184.

[16] M. Zoet, R. Welke, J. Versendaal, and P. Ravesteyn, "Aligning risk management and compliance considerations with business process development," in *E-Commerce and Web Technologies*. Springer, 2009, pp. 157–168. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-03964-5_16

[17] S. Taubenberger and J. Jürjens, "It security risk analysis based on business process models enhanced with security requirements," in *Modeling Security Workshop, Toulouse, France*, 2008.

[18] S. Kokolakis, A. Demopoulos, and E. A. Kiountouzis, "The use of business process modelling in information systems security analysis and design," *Information Management & Computer Security*, vol. 8, no. 3, pp. 107–116, 2000. [Online]. Available: http://dx.doi.org/10.1108/09685220010339192

[19] P. Herrmann and G. Herrmann, "Security requirement analysis of business processes," *Electronic Commerce Research*, vol. 6, no. 3-4, pp. 305–335, 2006. [Online]. Available: http://dx.doi.org/10.1007/s10660-006-8677-7

[20] A. Ekelhart, S. Fenz, and T. Neubauer, "Aurum: A framework for information security risk management," in *System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on*, 2009, pp. 1–10.

[21] G. Stoneburner, A. Goguen, and A. Feringa, *NIST 800-30, Risk Management Guide for Information Technology Systems, Special publication*, National Institue of Standards and Technology (NIST) Std., 2002.

[22] S. Ozkan and B. Karabacak, "Collaborative risk method for information security management practices: A case context within turkey," *International Journal of Information Management*, vol. 30, no. 6, pp. 567–572, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.ijinfomgt.2010.08.007

[23] R. A. Caralli, J. H. Allen, and D. W. White, *CERT Resilience Management Model (CERT-RMM): A Maturity Model for Managing Operational Resilience*. Addison-Wesley Professional, 2010.

[24] P. H. Gregory, *All in one - CISA - Certified Information Systems Auditor - Exam Guide*. McGraw-Hill Companies, 2012.

[25] *Risk Management - Principles and Guidelines*, International Organization for Standardization Std., ISO/IEC 31000:2009. [Online]. Available: http://dx.doi.org/10.3403/30246105

[26] V. Bier, "Challenges to the acceptance of probabilistic risk analysis," *Risk Analysis*, vol. 19, no. 4, pp. 703–710, 1999. [Online]. Available: http://dx.doi.org/10.1023/A%3A1007093805693

[27] G. Wangen and E. Snekkenes, "A taxonomy of challenges in information security risk management," in *Proceeding of Norwegian Information Security Conference / Norsk informasjonssikkerhetskonferanse - NISK 2013 - Stavanger*, vol. 2013. Akademika forlag, 2013.

[28] J. Wunder, A. Halbardier, and D. Waltermire, *Specification for Asset Identification 1.1*. NIST - US Department of Commerce, National Institute of Standards and Technology, 2011.

[29] R. S. Aguilar-Saven, "Business process modelling: Review and framework," *International Journal of production economics*, vol. 90, no. 2, pp. 129–149, 2004.

[30] B. Wetzstein, Z. Ma, A. Filipowska, M. Kaczmarek, S. Bhiri, S. Losada, J.-M. Lopez-Cob, and L. Cicurel, "Semantic business process management: A lifecycle based requirements analysis." in *SBPM*, 2007.

[31] R. K. Ko, "A computer scientist's introductory guide to business process management (bpm)," *Crossroads*, vol. 15, no. 4, p. 4, 2009. [Online]. Available: http://doi.acm.org/10.1145/1558897.1558901

[32] W. M. Van Der Aalst, A. H. Ter Hofstede, and M. Weske, "Business process management: A survey," in *Business process management*. Springer, 2003, pp. 1–12. [Online]. Available: http://dx.doi.org/10.1007/3-540-44895-0_1

[33] R. K. Ko, S. S. Lee, and E. W. Lee, "Business process management (bpm) standards: a survey," *Business Process Management Journal*, vol. 15, no. 5, pp. 744–791, 2009. [Online]. Available: http://dx.doi.org/10.1108/14637150910987937

[34] R. Burlton, *Business process management: profiting from process*. Pearson Education, 2001.

[35] R. Moen and C. Norman, *Evolution of the PDCA Cycle*. Associates in Process Improvement, 2011.

[36] A. Josey, *TOGAF Version 9: A Pocket Guide*. Van Haren Pub, 2009.

[37] D. J. Teece, "Capturing value from knowledge assets: The new economy, markets for know-how, and intagible assets." *California management review*, vol. 40, no. 3, 1998. [Online]. Available: http://dx.doi.org/10.2307/41165943

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 911–918
DOI: 10.15439/2014F122
ACSIS, Vol. 2

Security Evaluation of Bistable Ring PUFs on FPGAs using Differential and Linear Analysis

Dai Yamamoto*†, Masahiko Takenaka
* Fujitsu Laboratories Ltd.
Kanagawa, Japan
{yamamoto.dai, ma}@jp.fujitsu.com

Kazuo Sakiyama
† The University of Electro-Communications
Tokyo, Japan
sakiyama@uec.ac.jp

Naoya Torii
Fujitsu Laboratories Ltd.
Kanagawa, Japan
torii.naoya@jp.fujitsu.com

***Abstract*—Physically Unclonable Function (PUF) is expected to be an innovation for anti-counterfeiting devices for secure ID generation, authentication, etc. In this paper, we propose novel methods of evaluating the difficulty of predicting PUF responses (i.e. PUF outputs), inspired by well-known differential and linear cryptanalysis. According to the proposed methods, we perform a first third-party evaluation for Bistable Ring PUF (BR-PUF), proposed in 2011. The BR-PUFs have been claimed that they have a resistance against the response predictions. Through our experiments using FPGAs, we demonstrate, however, that BR-PUFs have two types of correlations between challenges and responses, which may cause the easy prediction of PUF responses. First, the same responses are frequently generated for two challenges (i.e. PUF inputs) with small Hamming distance. A number of randomly-generated challenges and their variants with Hamming distance of one generate the same responses with the probability of 0.88, much larger than 0.5 in ideal PUFs. Second, particular bits of challenges in BR-PUFs have a great impact on the responses. The value of responses becomes '1' with the high probability of 0.71 ($>$ 0.5) when just particular 5 bits of 64-bit random challenges are forced to be zero or one. In conclusion, the proposed evaluation methods reveal that BR-PUFs on FPGAs have some correlations of challenge-response pairs, which helps an attacker to predict the responses.**

I. Introduction

RECENTLY, the concept of Internet of things (IoT) has been widely spread. Various things such as vehicles, home appliances, medical devices and sensing devices are connected to the Internet. It is expected that this provides us a lot of new services and products in the field of industry, education, healthcare, agriculture, etc. First of all, the secure IoT requires us to realize an authentication system for things. This is because counterfeiting the things causes serious security problems for the services and products based on the IoT concept. Generally, hardware-based approaches are often used for the authentication of things. For example, cryptographic hardware using integrated circuits (ICs) stores a secret key in its internal memory. A secure cryptographic protocol using the cryptographic hardware enables us to authenticate the things, making the secret key invisible from the outside. This approach prevents a leakage of the key outside, and makes it impossible to counterfeit things. However, recent research has found that the secret key could be revealed by de-packaging the IC and analyzing the IC mask design [2]. Therefore, further techniques are necessary to protect the cryptographic hardware storing the secret keys.

Recently, Physically Unclonable Functions (PUFs) have been focused as a solution to the secure authentication for things [3]. PUFs are realized in individual IC chips, and have a completely identical circuit structure. In spite of the identical circuit, PUFs generate the unique output values (responses) to the same input value (challenge) for each individual IC. This uniqueness is provided by process variations in memory characteristics or wire/gate delay occurring in the manufacturing process of each IC chip [4] [5]. Even if an attacker de-packages and analyzes ICs of PUFs, she cannot analyze the process variations due to the identical PUF structure. As a result, she cannot reveal the challenge-response pairs of PUFs. Therefore, PUFs can be utilized for a secure authentication system for things.

There are two categories in the PUFs on ICs: *memory-based* PUFs utilizing the memory characteristics and *delay-based* PUFs utilizing wire/gate delay variations [6]. One of the most feasible and secure memory-based PUFs is latch-based PUFs (LPUFs) [7] [8]. The LPUF generates an N-bit response based on N outputs from N RS latches. The RS latch is composed of cross-coupled logic gates, and is similar to a memory cell. Each bit of the response is generated from each latch output in a stable state after a metastable state. The metastable state is affected by the memory characteristics, thus the latch outputs (i.e. response bits) are also unique for each individual IC. One of the most famous delay-based PUFs is Ring Oscillator PUFs (RO-PUFs) [9]. The RO-PUF has M number of ROs, one of which is composed of odd number of cascaded inverters as a ring. The RO-PUF derives 1-bit responses from the difference of oscillator frequencies between two arbitrary ROs. Consequently, 1-bit response becomes zero or one, depending on which RO has a larger frequency. The number of responses is ${}_M\mathrm{C}_2$, which corresponds to the number of combinations of M ROs taken 2 at a time. The oscillator frequencies are affected by the wire/gate delays, which makes the responses unique for each individual IC.

Bistable Ring PUFs (BR-PUFs), having both properties of memory-based and delay-based PUFs, were proposed and self-evaluated by Chen et al. (hereinafter called "developers") [10] [11]. There are two major differences between BR-PUFs and

The preliminary version of this paper was presented in a Japanese domestic symposium without peer review [1].

978-83-60810-58-3/$25.00

RO-PUFs: (1) the structure of a ring, (2) response generation. (1) A BR-PUF is composed of cascaded inverters as a ring (hereinafter called "primitive BR-PUF"). A primitive BR-PUF is similar to an RO-PUF in terms of the ring of cascaded inverters. The difference is that the number of the inverters is not odd but even. Hence the primitive BR-PUF does not keep oscillation, but make the transition from metastable to stable state like memory-based PUFs. The primitive BR-PUF derives a 1-bit response from which stable state the ring is, e.g. '10101010' or '01010101' in a ring of 8 inverters. (2) A primitive BR-PUF generates just one 1-bit response because it consists of one ring, while an RO-PUF includes multiple rings. To generate multiple 1-bit responses, the BR-PUF has a *basic component* instead of the inverters in the ring. The basic component consists of two logic gates, either of which is selected by 1 bit of challenge [10]. The BR-PUF with 64 basic components, for example, has 64-bit challenge to select the logic gates. The BR-PUF is organized by 2^{64} different types of rings, the logic gates of which are differently selected depending on the values of challenges. Therefore, the BR-PUF can generate multiple challenge-response pairs without having multiple rings like the RO-PUF.

In this paper, we evaluate security aspects of BR-PUFs implemented on FPGAs. The reason why we focus on this PUF is that BR-PUFs have the following advantages with both memory-based and delay-based PUFs, as claimed by the developers:

- BR-PUFs are similar to delay-based PUFs in that the number of challenge-response pairs is exponential to the bit length of challenges, which makes difficult the predictions of responses.
- BR-PUFs also have the resistance against a machine learning attack and a modeling attack like memory-based PUFs.

BR-PUFs are evaluated by developers themselves. These self-evaluation results are very useful for users to understand the effectiveness of the proposed PUFs. However, the evaluation results may be different depending on PUF implementations since PUFs are based on physical characteristics in ICs. Hence it is quite important to evaluate and analyze newly proposed PUFs by third-party researchers as attackers. These BR-PUFs with excellent characteristics have not been evaluated by other researchers yet.

A. Our Contributions

In order to evaluate the security of PUFs, we focus on the difficulty of predicting responses. This difficulty is one of requirements for PUFs. Consequently, responses for unknown challenges should be unpredictable and non-biased even when some challenge-response pairs are known. In this paper, we propose novel two methods of evaluating this difficulty:

Our Evaluation Method (i):

> What is the probability that PUFs generate the same responses for two challenges with small Hamming distance?

Our Evaluation Method (ii):

> What is the probability that PUFs generate the same responses for multiple challenges whose particular bits are forced to be zero or one?

These methods are inspired by well-known cryptanalysis methods: differential cryptanalysis [12] and linear cryptanalysis [13]. The Evaluation Method (i) focuses on how differences in the challenge (plaintext) lead to differences in the response (ciphertext). The Evaluation Method (ii) is based on the idea that the response (ciphertext) is linearly approximated by particular bits of the challenge (plaintext). If the probabilities in these methods are close to 0.5, the evaluated PUFs are highly secure because they can generate non-biased responses independently of the values of challenges.

In this paper, we evaluate the security of BR-PUFs on Xilinx FPGAs (Spartan-6) according to our two evaluation methods. We analyze a number of challenge-response pairs obtained from BR-PUFs consisting of 64 inverters implemented on FPGAs. As a result, our case study supports that BR-PUFs on FPGAs have undesirable performance in the differential and linear evaluations; the probability is far from ideal 0.5. The differential evalaution shows that two types of rings for the challenges with small Hamming distance have many common logic gates (i.e. similar circuit characteristics), which are likely to generate the same response. The linear evaluation implies that BR-PUFs have some special inverter gates which have a great impact on the responses.

This paper is the first time that BR-PUFs on FPGAs have some security issues of response predictions. More importantly, our two evaluation methods can be used as universal methods for evaluating the security of other types of PUFs.

B. Organization of the Paper

The rest of the paper is organized as follows. Section II gives an outline of the BR-PUF. Section III proposes two evaluation methods of the difficulty of predicting responses. Section IV evaluates the BR-PUFs implemented on an FPGA platform according to the evaluation methods. Finally, in Section V we summarize our work and comment on future directions.

II. Bistable Ring PUF

The BR-PUF was proposed by Chen et.al. in 2011 [10]. Figure 1 shows the basic mechanism of the BR-PUF. The BR-PUF consists of the even (e.g. eight in Fig. 1) number of inverters (INVs), which are connected as a ring. After voltage is supplied, the ring has two possible stable states, '10101010' ('A'-state) or '01010101' ('5'-state), enumerating inverter's outputs beginning from INV_1. The ring generates 1-bit response according to which state the ring falls into. The BR-PUF is similar to the delay-based RO-PUF in terms of having inverter rings. It also has the same characteristic with the memory-based Latch-PUF, having two possible states.

Figure 2 shows the circuit structure of the BR-PUF, presented in [10]. The inverter in Fig. 1 is implemented by a *BR-S*, which is a basic component of the BR-PUF. The l-th BR-S, i.e. BR-S_l ($1 \leq l \leq 64$), is composed of two NOR gates,

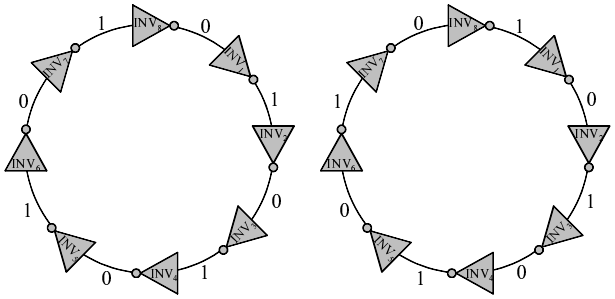


Fig. 1. Two possible stable states on a primitive BR-PUF with 8 inverters.

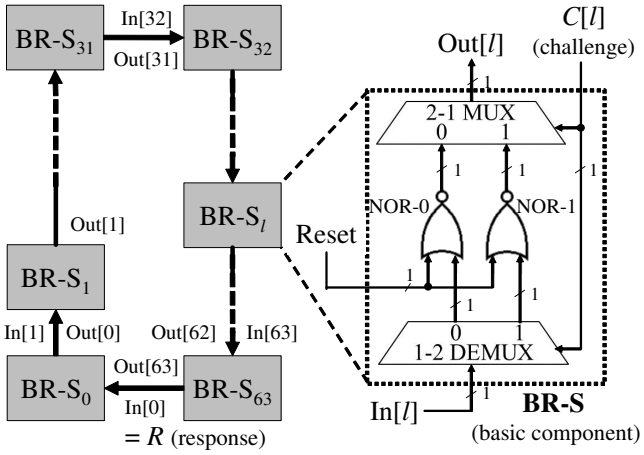


Fig. 2. Circuit structure of Bistable Ring PUF.

a 2-to-1 MUX and a 1-to-2 DEMUX. A 1-bit challenge $C[l]$ is input to the BR-S$_l$ to select either of the NOR gates. The BR-PUF with 64 BR-Ss has 64-bit challenges, which means 2^{64} different types of rings can be organized. Each NOR gate has different characteristics, i.e. drive capability or gate/wire delay. Hence the value of challenges has a great impact on the decision of stable states, either A-state or 5-state, as claimed by the developers. A 1-bit response is extracted from an arbitrary signal between two BR-Ss, e.g. the output from BR-S$_{63}$, i.e. Out[63] (=In[0]) in Fig. 2. The BR-S$_l$ works as an inverter when reset signal equals to 0. In contrast, the input and output of BR-S$_l$, In[l] and Out[l], can be forced to zero when the reset signal is 1 (i.e. neither A-state nor 5-state). This enables us to generate responses at any time after power up. In conclusion, BR-PUF has multi-bit challenges and generates a number of challenge-response pairs at any time.

III. Proposed Evaluation Methods

It is well known that responses of some delay-based PUFs are predictable through a machine learning attack [14]. The developers of BR-PUFs claim that BR-PUFs have a resistance against such an attack [10]. This resistance is based on the complex and non-linear behavior of BR-PUFs, different from other delay-based PUFs. Hence they claim that an attacker cannot predict responses of BR-PUFs. For the verification of this resistance, we consider that the correlation among challenge-response pairs should be evaluated experimentally. This evaluation, unfortunately, has not been performed by other researchers yet.

In this paper, we propose novel two methods of evaluating PUFs in terms of the resistance against response predictions. In the proposed method (i), we evaluate whether or not challenges with small Hamming distance result in highly correlated responses. In the proposed method (ii), we evaluate whether or not we obtain the same responses with high probability if certain bits of challenges are forced to zero or one. In the following, we explain these methods, assuming the case of evaluating BR-PUFs.

A. Proposed Method (i): Differential PUF Analysis

A group of challenges with small Hamming distance may cause the problem that most of NOR gates are selected commonly, so the characteristics impacting on the responses are also similar one another. In detail, let R_j's be the responses obtained from 64-bit challenges C_j's. Here, let $\tilde{R}_j^{(1,i)}$ be the response obtained from $\tilde{C}_j^{(1,i)} (1 \leq i \leq 64)$ whose Hamming distance is one from C_j (i.e. the only i-th bit from least significant bit (LSB) is different). In ideal PUFs, $\tilde{R}_j^{(1,i)}$ and R_j have little correlation. If a correlation exists, $\tilde{R}_j^{(1,i)}$ has a possibility to be easily predicted by an attacker when challenge-response pairs (C_j, R_j) are known. This means that the implemented BR-PUFs have a serious security issue.

The proposed method (i) is inspired by the well-known cryptanalysis method: differential cryptanalysis [12]. The differential cryptanalysis evaluates the avalanche effect: the effects of the changes of plaintext bits on ciphertext bits. In the proposed method (i), we evaluate how differences in the challenge (plaintext) lead to differences in the response.

B. Proposed Method (ii): Linear PUF Analysis

We consider that some logic gates and wires may be quite different from many other ones. This is because of the process variations in the circuit characteristics such as drive capability or gate/wire delay. If such gates and wires exist in a ring of the BR-PUF, the stable state falls into either state with high probability. As a result, the number of independent challenge-response pairs is very small, which is a security problem for PUFs.

This method is inspired by linear cryptanalysis [13]. In this cryptanalysis, an attacker tries to find linear equations with plaintext bits and ciphertext bits which have a high bias. This provides us with the inspiration of the general method of evaluating PUFs in terms of response predictions.

From the view point of designers of PUFs, they must design a secure PUF whose responses cannot be predicted by an attacker. Of course, machine learning attacks can evaluate a tolerance against the response predictions. However, a concrete method of designing such a secure PUF has not been established yet. Therefore, in this paper, the proposed

evaluation methods provide fundamental principles to design secure PUFs, in terms of correlations between challenges and responses. Next section experimentally evaluates the BR-PUFs implemented on FPGAs according to our proposed methods.

IV. Experimental Evaluation

A. Experimental Setup

Figure 3 shows our experimental system, which consists of two boards: a custom-made board with a Xilinx Spartan-6 FPGA (XC6SLX16-2CSG324C) and a commercially-available Spartan-3E starter kit board with a Xilinx Spartan-3E FPGA (XC3S500E-4FG320C). We implemented the BR-PUF circuit with 64 BR-Ss on the Spartan-6 FPGA, and the peripheral circuits such as the block RAM and RS232C module on the Spartan-3E FPGA. An Spartan-6 FPGA chip was put on a socket of the custom-made board, being therefore easily replaceable by another chip. We evaluated 4 BR-PUFs implemented on 4 Spartan-6 FPGA chips: $\mathrm{FPGA}_x (1 \leq x \leq 4)$.

Our response acquisition process was as follows. When the RS232C module in the Spartan-3 FPGA received a start command from a user PC, the module sent a start signal to a CTRL module. The CTRL module got a 64-bit linear feedback shift register (LFSR) to generate 2,048 random challenges $C_j (1 \leq j \leq 2,048)$. According to [15], the tap sequence of the LFSR was set to [64, 63, 61, 60], and the initial value was set to '0x123456789ABCDEF0'. The 64-bit challenge was divided into four 16-bit values, which were sent and stored to the flip-flops (FFs) on Spartan-6 FPGA. The reset signal to the BR-PUF was changed from 1 to 0, then the response acquisition was started. Not only 1-bit output but also all of 64-bit output from BR-Ss was stored into the 64-bit flip-flop. This enables us to confirm whether or not the response is stable; if the 64-bit value has at least two consecutive 1's/0's, the response is regarded as unstable state, vice versa. In our experiment, the 64-bit value was stored after sufficient time (i.e. approximately 6 ms) from the reset signal changing to 0 in order to make the response as stable as possible. The 64-bit value was sent to a block RAM on the Spartan-3E bit-sequentially, and was transmitted to the user PC through an RS232C port.

Both design and implementation of the BR-PUF are very important because they have a large impact on the eventual response behavior of the PUF itself. Hence we take great care of the symmetric layout of the BR-PUF as follows. Figure 4 shows our custom layout of a BR-PUF with 64 BR-Ss on a Spartan-6 FPGA. The 64 BR-Ss were implemented on the ring-shaped neighboring CLBs (configurable logic blocks), expecting that the wire lengths between all BR-Ss are identical. This symmetric layout is expected to make a uniform ring and a bias of responses as small as possible.

Before we perform an experimental evaluation, we verify the implemented BR-PUFs according to the responses R_j's for the 2,048 random challenges C_j's. Average Hamming distance between two arbitrary 64-bit challenges among the 2,048 challenges is 32.00. This is extremely close to theoretical value $(= 64/2)$, so our using challenges are enough random. By

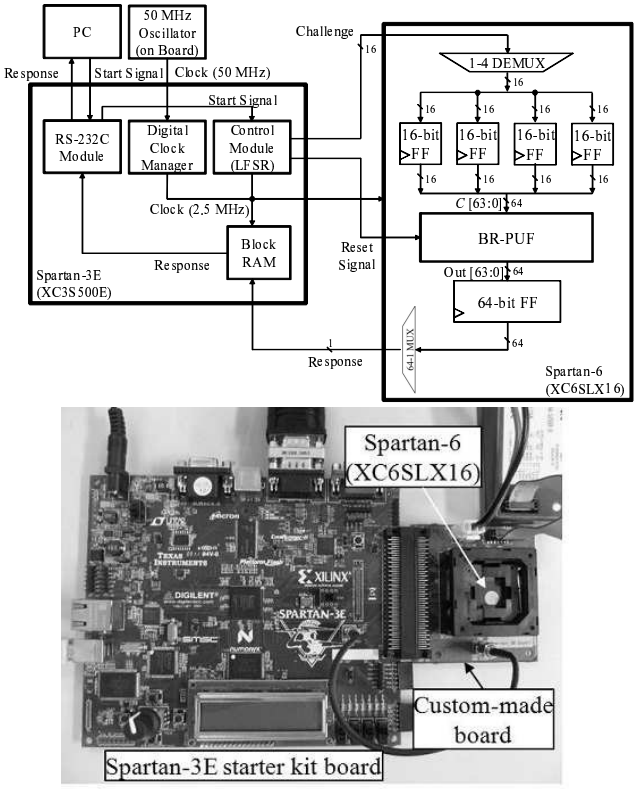


Fig. 3. Experimental system.

using these challenges, we evaluate average hamming distance between two arbitrary responses among the 2,048 responses (i.e. ${}_{2048}\mathrm{C}_2$ combinations). The results are 0.50, 0.49, 0.49 and 0.46 in four BR-PUFs, respectively. These are very close to the ideal value (= 0.5), so our implemented BR-PUFs are verified to generate almost non-biased responses for random challenges.

B. Experimental Results - using Proposed Method (i)

This section evaluates the correlation among the responses obtained from challenges with small Hamming distance. We generate a certain number of challenges $\tilde{C}_j^{(k,i)}$ satisfying the following condition: the Hamming distance between C_j and $\tilde{C}_j^{(k,i)}$ being equal to k, i.e. HD(C_j, $\tilde{C}_j^{(k,i)}$) $= k$. For example, in the case for $k = 1$, we generate 64 challenges $\tilde{C}_j^{(1,i)} (1 \leq i \leq 64)$, where i-th LSB is just different. In our experiment, we evaluate the responses in $k = 1, 2, 4, 8$ and 16. In the case where $k > 1$, however, the number of challenges $\tilde{C}_j^{(k,i)}$ is ${}_{64}\mathrm{C}_k$, which becomes quite large for the value of large k. Due to time constraints, we generate the following two types of challenges $\tilde{C}_j^{(k,i)}$:

Type A

Neighboring k bits are different between C_j and $\tilde{C}_j^{(k,i)}$ as shown in Fig. 5(I).

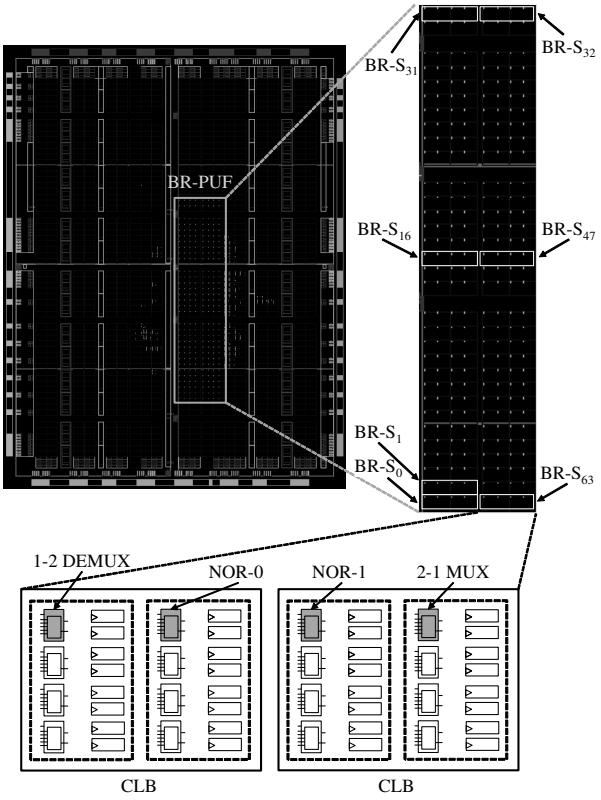
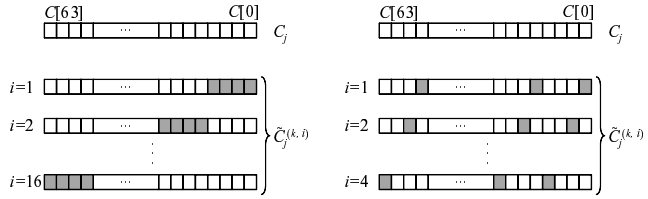


Fig. 4. Implementation of our BR-PUF with 64 BR-Ss on a Spartan-6 FPGA.

Type B
: Intervals of $64/k$ bits are different as shown in Fig. 5(Ⅱ).

Table I shows the number of $\tilde{C}_j^{(k,i)}$ in each k according to the aforementioned types. We generate 184 ($= 124 + 60$) challenges $\tilde{C}_j^{(k,i)}$ for each of 2,048 C_j's. Hence we obtain the total of 378,880 ($= 2,048 \times 185$) challenge-response pairs from each BR-PUF.

Figure 6 shows the ratios of the challenges $\tilde{C}_j^{(k,i)}$ which generate the same responses as each C_j. These results are the means of 4 implemented BR-PUFs. From the result of Type A in $k = 1$, 88.0% of challenges $\tilde{C}_j^{(1,i)}$ lead to the same responses as C_j. This ratio should be around 50% in secure PUFs. The larger the value of k is, the lower the ratios of such challenges are. However, even in the Type A of $\tilde{C}_j^{(16,i)}$ where HD(C_j, $\tilde{C}_j^{(16,i)}$)=16, the probability is approximately 0.665, which is larger than ideal 0.5. Additionally, there is almost no difference between both types in Fig. 6. This indicates that the similarity of responses depends not on the locations of the different bits, but just on the Hamming distance of the challenges. Consequently, if a challenge-response pair is known to an attacker, she has a high possibility to predict the responses for challenges with small Hamming distances by using the known challenge.



(I) Type A ($k = 4$) (Ⅱ) Type B ($k = 16$)

Fig. 5. Two types of challenges $\tilde{C}_j^{(k,i)}$ (Colored bits are different between C_j and $\tilde{C}_j^{(k,i)}$).

TABLE I
NUMBER OF CHALLENGES FOR k IN BOTH TYPES.

k	Type A	Type B
1	64	N/A
2	32	32
4	16	16
8	8	8
16	4	4
Sum	124	60

Different from other PUFs, the generation time of responses, i.e. the duration period for stable states, is quite different depending on values of challenges in BR-PUFs [10]. The generation time has a strong impact on the *reliability* and *uniqueness* of the responses, systematically defined as PUF performance metrics in [16]. Especially, the responses obtained in a short transient time have little uniqueness[1]: a small difference among BR-PUFs because circuit layout influences the responses strongly. Hence we should select and use the only responses with long transient time, as presented in [10]. In the above-mentioned evaluation we focus on all of challenge-response pairs without consideration of the transient time. We anticipate that highly-unique responses with the long transient time have a lower similarity, even if the challenges have a small Hamming distance. To confirm this we obtain the 64-bit outputs of BR-Ss, i.e. responses for 2,048 C_j's, in a short time of approximately 70μs after the reset signal to the BR-PUF is zero. 1,658 (approximately 80.96%) out of 2,048 C_j's lead to stable responses with alternate bits. Here, we focus only on the remaining of 390 C_j's and perform the same evaluation as above mentioned, whose results are shown in Fig. 7. The correlation between the value of responses and the Hamming distance of challenges becomes small, as we expected. However, the correlation still exists: 68.1% of challenges $\tilde{C}_j^{(1,i)}$ lead to the same responses as C_j's. This indicates that the responses of BR-PUFs are predictable even if we use the selection of challenge-response pairs, presented by developers. In conclusion, this dependency of the responses on the Hamming distance of challenges might facilitate an attacker to succeed in her modeling attack, and predict most of unknown responses.

[1]According to the BR-PUFs on ASICs self-evaluated by the developers through SPICE simulations in [11], the PUF performances such as reliability and uniqueness are not affected by the generation time of responses.

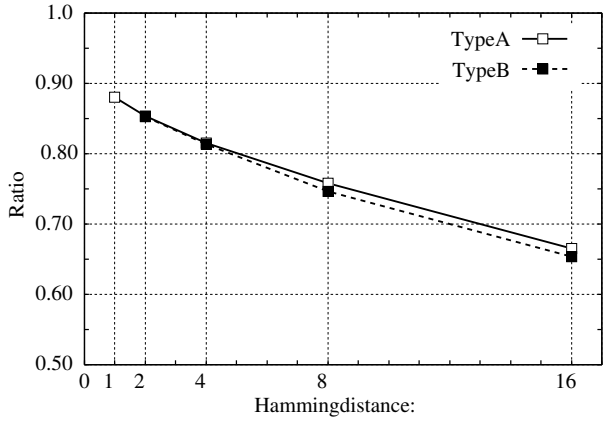


Fig. 6. Ratios of challenges $\tilde{C}_j^{(k,i)}$ generating the same responses as C_j for k in both types.

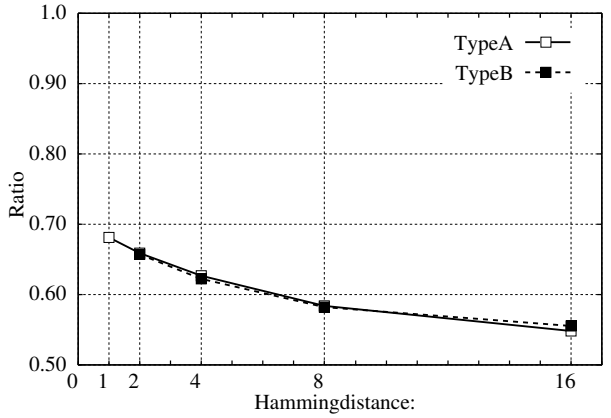


Fig. 7. Ratios of challenges $\tilde{C}_j^{(k,i)}$ generating the same responses as C_j whose transient time is longer than 70 μs.

C. Experimental Results - using Proposed Method (ii)

This section evaluates whether or not BR-PUFs have a BR-S with an *influential* NOR gate, which has a decisive impact on the value of responses. We anticipate that the influential NOR gate has quite different circuit characteristics from other NOR gates. The location of the influential NOR gate is defined by the following two parameters: *enforced bit* ($\leq$ 64) and *enforced value* (0/1). The enforced bit means the location of the BR-S including the influential NOR gate. The enforced value represents either NOR gate in the BR-S. For example, if the enforced bit is 33 and the enforced value is 1, the influential NOR gate is the NOR-1 gate in BR-S_{33}.

As a preliminary experiment to confirm the existence of influential NOR gates, we analyze the 2,048 challenge-response pairs (C_j, R_j) same as Section IV-B. 64-bit challenges of BR-PUFs correspond to the way of selecting NOR gates in BR-Ss. We extract part of C_j's from 2,048 ones whose certain m ($1 \leq m \leq 5$) bits are the same one another, i.e. common NOR gates are selected. Our software program searches all patterns of selecting m NOR gates (${}_{64}\mathrm{C}_m \cdot 2^m$ combinations). Due to time constraints, we set m to less than 6. Table II shows the number of responses (=‘1’s) for the part of C_j's. We explain how to read the table with the specific example of $m = 3$, as follows. Out of 2,048 there are 236 C_j's whose 58th, 13rd and 6th LSBs are 1, 0 and 1, respectively. The number of responses whose values are ‘1’s is 205, which is 86.9% of 236 R_j's. Hence these three NOR gates are predicted to be influential NOR gates, i.e. (enforced bit, enforced value)= $(58, 1), (13, 0), (6, 1)$. Table II also shows the 6 patterns of influential NOR gates for each m. From Table II, we see that more than 65% of responses become 1 in the BR-PUF with just one influential NOR gate (i.e. $m = 1$). The number of the influential NOR gates is around 10 in the 64 BR-Ss. The larger the number of influential NOR gates ($= m$) is, the larger the percentage of responses (=‘1’s) is, i.e. the larger impact on the responses. Especially, all responses become 1 when $m = 5$. In conclusion, according to the analysis of 2,048 C_j's, we demonstrate that our BR-PUF on FPGA has influential NOR gates with a decisive impact on the values of responses.

Above-mentioned results are obtained from a BR-PUF on FPGA_1. We also confirm that the other three BR-PUFs on FPGA_2, FPGA_3 and FPGA_4 have influential NOR gates. BR-PUFs on FPGA_1, FPGA_2 and FPGA_3 generate responses biased to one, while the BR-PUF on FPGA_4 outputs responses biased to zero. The locations of influential NOR gates are different from each FPGA. These are caused by the characteristics of BR-Ss.

As a further experiment, we evaluate the responses for much larger number of challenges than 2,048. First, additional 2^{15} C_j's ($1 \leq j \leq 2^{15}$) are obtained by using the LFSR on the Spartan-3E FPGA. Next, we generate $\hat{C}_j$'s whose enforced bits are changed to the enforced values according to Table II. This means that influential NOR gates are definitely included in the rings of our BR-PUF, and the other NOR gates are selected randomly. Figure 8 shows the ratio of responses equal to 1 for $\hat{C}_j$'s. The line graph represents the average result of six patterns of influential NOR gates as shown in Table II. The upper and lower bounds for error-bars mean the maximum and minimum results of the six patterns, respectively. From Fig. 8, we see that the responses are biased to one when our BR-PUF includes influential NOR gates. The probability of responses being one is 71.4% and 54.5% when the number of influential NOR gates is set to 5 and 1, respectively. The reason why the degree of the bias is smaller than in Table II is more likely that responses are affected by other influential NOR gates not shown in Table II. In conclusion, an attacker who knows some challenge-response pairs could reveal the properties (i.e. influential NOR gates) of her target BR-PUF like Table II. After that, she has a high possibility to predict unknown challenge-response pairs. To minimize the impact of the influential NOR gates, special layout and implementation custom-designed for each BR-PUF are required, however, increase the manufacturing costs dramatically.

TABLE II
INFLUENTIAL NOR GATES AND THEIR IMPACT ON A BIAS OF RESPONSES.

m	**Influential NOR gate(s)** Enforced bit (i-th LSB) : Enforced value (0/1)	# of responses (= 1) / # of responses for challenges with left-column's NORs
1	53:0	701 / 1046 (67.0%)
	25:0	716 / 1044 (68.6%)
	19:0	700 / 1041 (67.2%)
	18:1	678 / 1008 (67.3%)
	06:1	682 / 1011 (67.5%)
	01:0	709 / 1037 (68.4%)
2	53:0, 25:0	411 / 539 (76.3%)
	52:1, 01:0	384 / 505 (76.0%)
	37:0, 06:1	384 / 502 (76.5%)
	25:0, 18:1	400 / 514 (77.8%)
	15:0, 09:0	402 / 528 (76.1%)
	09:0, 06:1	384 / 504 (76.2%)
3	58:1, 13:0, 06:1	205 / 236 (86.9%)
	54:1, 25:0, 18:1	204 / 239 (85.4%)
	53:0, 17:0, 11:0	215 / 252 (85.3%)
	43:0, 37:0, 06:1	219 / 257 (85.2%)
	25:0, 20:1, 19:0	234 / 275 (85.1%)
	25:0, 18:1, 01:0	231 / 271 (85.2%)
4	63:0, 59:0, 37:0, 06:1	124 / 132 (93.9%)
	58:1, 52:1, 13:0, 06:1	112 / 120 (93.3%)
	54:1, 25:0, 18:1, 01:0	114 / 121 (94.2%)
	53:0, 28:1, 11:0, 00:1	122 / 131 (93.1%)
	43:0, 40:1, 32:1, 01:0	123 / 132 (93.2%)
	27:0, 25:0, 18:1, 06:1	123 / 132 (93.2%)
5	53:0, 51:0, 45:0, 18:1, 07:0	79 / 79 (100%)
	58:1, 41:0, 32:1, 19:0, 06:1	76 / 76 (100%)
	59:0, 43:0, 32:1, 13:0, 01:0	61 / 61 (100%)
	63:0, 59:0, 45:0, 18:1, 17:0	61 / 61 (100%)
	52:1, 51:0, 35:1, 20:1, 01:0	45 / 45 (100%)
	48:1, 32:1, 26:1, 10:1, 02:1	41 / 41 (100%)

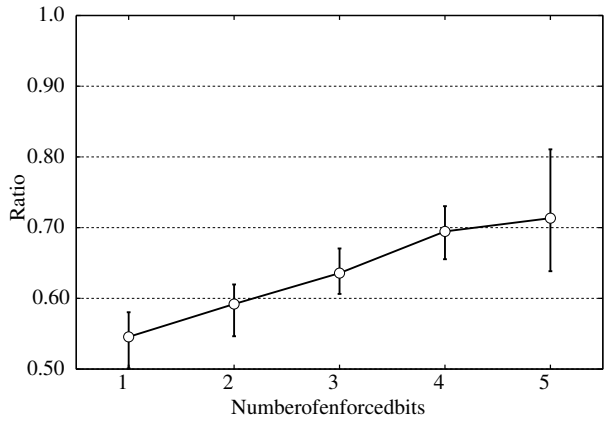


Fig. 8. Ratio of responses (= 1) for $\hat{C}_j$'s whose m-bit enforced bits are changed to enforced values according to Table II.

D. Discussion

1) **Are the wire lengths between all BR-Ss identical?***:* There is no evidence that all of the wire lengths are completely identical. The reason is that it is difficult to exactly control the wire length because logic gates on FPGAs are fixed on grid-pattern layouts. As shown in Fig. 4, we implement BR-PUFs on FPGAs as carefully as possible. In spite of the careful implementations, our BR-PUFs have security issues: correlations of challenge-response pairs. This indicates that it is difficult for many designers to implement BR-PUFs on FPGAs securely. Therefore, this paper gives useful information for designers of BR-PUFs.

2) **Is the implementation of BR-PUFs in this paper the same as that in original paper?***:* We derive 64-bit outputs from all of the 64 BR-Ss, instead of just one in original. We consider that the original implementation is not the best option. This is because deriving only one output may lead to unbalance of capacitive loads on the output of each BR-S, which causes influential gates. We derive outputs from all of BR-Ss in order to prevent this unbalance.

3) **Is it appropriate to derive general results from a single implementation on a specific FPGA family (Xilinx Spartan 6)?***:* Xilinx Spartan 6 FPGAs are relatively newly-released, and have almost the same structure as other types of FPGAs such as Xilinx Virtex 6. We expect, therefore, that similar results are confirmed on the Xilinx FPGA family. In contrast, we need further evaluations on other FPGA families (e.g. developed by Altera) or ASICs because their structures are completely different from that of the Xilinx FPGA family.

4) **Why BR-PUFs are evaluated?***:* It is ture that BR-PUFs are not one of the most famous PUFs. However, we consider BR-PUFs to be excellent and promising PUFs bacause BR-PUFs have advantages of both memory-based and delay-based PUFs. That is why we focus on BR-PUFs in this paper. Our main contribution is to propose the evaluation methods for PUFs: differential and linear PUF analyses. The experimental evaluation of BR-PUFs is a case study of PUF evaluations based on the proposed methods. The proposed methods can be used to evaluate not only BR-PUFs but also other types of PUFs, e.g. arbiter-based PUFs [17].

V. CONCLUSION

In this paper, we proposed the evaluation methods for PUFs: differential and linear PUF analyses. Based on these methods, we experimentally analyzed responses obtained from BR-PUFs using 64 BR-Ss, composed of two NOR gates, implemented on Xilinx Spartan-6 FPGAs. We evaluated the probability of a prediction of the responses R_j for challenge C_j ($1 \leq j \leq 2,048$). According to differential analysis for BR-PUFs, we demonstrated that approximately 88.0% and 66.5% of responses become 1 for challenges with Hamming distance of 1 and 16, respectively. These results are much larger than about 50% in ideal BR-PUFs. Hence an attacker has a high possibility to predict the responses for challenges with small Hamming distances from her known challenge-response pairs. According to linear PUF analysis, we demonstrated that BR-PUFs have some *influential NOR gates*, which cause a strong bias of responses. The probability of responses being one is 71.4% and 54.5% when the number of influential NOR gates is 5 and 1, respectively. An attacker has a high possibility to predict unknown challenge-response pairs by specifying the location of influential NOR gates. Our experimental results are the first time that BR-PUFs present undesirable PUF behavior due to the response prediction, and

compromise the whole security of a system based on BR-PUFs. More importantly, our two evaluation methods can be used as universal framework for evaluating the security of other PUFs (e.g. Arbiter PUFs).

Other implementations of BR-PUFs would probably not behave likewise Spartan-6 FPGAs. For example, the bias of responses has a possibility to improve if BR-PUFs are implemented on other types of FPGAs or ASICs. Future work should include a discussion of security evaluation of BR-PUFs on various platforms.

REFERENCES

[1] D. Yamamoto, M. Takenaka, and N. Torii, "Performance and Security Evaluation of BR-PUF on FPGAs (in Japanese)," in *The 30th Symposium on Cryptography and Information Security (SCIS 2013)*, 2013.

[2] R. Torrance and D. James, "The State-of-the-Art in IC Reverse Engineering," in *Cryptographic Hardware and Embedded Systems (CHES 2009)*, 2009. doi: 10.1007/978-3-642-04138-9_26 pp. 363–381. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-04138-9_26

[3] R. S. Pappu, "Physical One-Way Functions," Ph.D. dissertation, Massachusetts Institute of Technology, 2001.

[4] B. Gassend, D. Clarke, D. Lim, M. van Dijk, and S. Devadas, "Identification and Authentication of Integrated Circuits," *Concurrency - Practice and Experience*, vol. 16, no. 11, pp. 1077–1098, 2004. doi: 10.1002/cpe.805. [Online]. Available: http://dx.doi.org/10.1002/cpe.805

[5] B. Gassend, D. E. Clarke, M. van Dijk, and S. Devadas, "Silicon physical random functions," in *ACM Conference on Computer and Communications Security*, 2002. doi: 10.1145/586110.586132 pp. 148–160. [Online]. Available: http://doi.acm.org/10.1145/586110.586132

[6] R. Maes and I. Verbauwhede, "Physically Unclonable Functions: A Study on the State of the Art and Future Research Directions," in *Towards Hardware Intrinsic Security: Foundation and Practice*. Springer, 2010, pp. 3–37. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-14452-3_1

[7] Y. Su, J. Holleman, and B. Otis, "A 1.6pJ/bit 96% Stable Chip-ID Generating Circuit using Process Variations," in *IEEE International Solid-State Circuits Conference (ISSCC 2007)*, 2007. doi: 10.1109/ISSCC.2007.373466 pp. 406–611. [Online]. Available: http://dx.doi.org/10.1109/ISSCC.2007.373466

[8] Y. Su, J. Holleman, and B. P. Otis, "A Digital 1.6pJ/bit Chip Identification Circuit Using Process Variations," *Solid-State Circuits, IEEE Journal of*, vol. 43, no. 1, pp. 69–77, 2008. doi: 10.1109/JSSC.2007.910961. [Online]. Available: http://dx.doi.org/10.1109/JSSC.2007.910961

[9] G. E. Suh and S. Devadas, "Physical Unclonable Functions for Device Authentication and Secret Key Generation," in *Design Automation Conference (DAC 2007)*, 2007. doi: 10.1109/DAC.2007.375043 pp. 9–14. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/DAC.2007.375043

[10] Q. Chen, G. Csaba, P. Lugli, U. Schlichtmann, and U. Rührmair, "The Bistable Ring PUF: A New Architecture for Strong Physical Unclonable Functions," in *Hardware-Oriented Security and Trust (HOST 2011)*, 2011. doi: 10.1109/HST.2011.5955011. [Online]. Available: http://dx.doi.org/10.1109/HST.2011.5955011

[11] ——, "Characterization of the Bistable Ring PUF," in *Design, Automation & Test in Europe Conference & Exhibition (DATE 2012)*, 2012. doi: 10.1109/DATE.2012.6176596. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/DATE.2012.6176596

[12] E. Biham and A. Shamir, "Differential Cryptanalysis of DES-like Cryptosystems," *J. Cryptology*, vol. 4, no. 1, pp. 3–72, 1991. doi: 10.1007/BF00630563. [Online]. Available: http://dx.doi.org/10.1007/BF00630563

[13] M. Matsui, "Linear Cryptoanalysis Method for DES Cipher," in *EUROCRYPT*, 1993. doi: 10.1007/3-540-48285-7_33 pp. 386–397. [Online]. Available: http://dx.doi.org/10.1007/3-540-48285-7_33

[14] U. Rührmair, F. Sehnke, J. Sölter, G. Dror, S. Devadas, and J. Schmidhuber, "Modeling Attacks on Physical Unclonable Functions," in *ACM Conference on Computer and Communications Security*, 2010. doi: 10.1145/1866307.1866335 pp. 237–249. [Online]. Available: http://dx.doi.org/10.1145/1866307.1866335

[15] R. Ward and T. Molteno, "Table of Linear Feedback Shift Registers," University of Otago, New Zealand, Tech. Rep., 2007. [Online]. Available: http://www.eej.ulst.ac.uk/~ian/modules/EEE515/files/old_files/lfsr/lfsr_table.pdf

[16] A. Maiti, V. Gunreddy, and P. Schaumont, "A Systematic Method to Evaluate and Compare the Performance of Physical Unclonable Functions," in *Embedded Systems Design with FPGAs*. Springer New York, 2013, pp. 245–267. [Online]. Available: http://dx.doi.org/10.1007/978-1-4614-1362-2_11

[17] J. W. Lee, D. Lim, B. Gassend, G. E. Suh, M. van Dijk, and S.Devadas, "A Technique to Build a Secret Key in Integrated Circuits for Identification and Authentication Applications," in *IEEE VLSI Circuits Symposium 2004*, 2004. doi: 10.1109/VLSIC.2004.1346548 pp. 176–179. [Online]. Available: http://dx.doi.org/10.1109/VLSIC.2004.1346548

Frontiers in Network Applications, Network Systems and Web Services

SYMPOSIUM SoFAST-WS focuses on modern challenges and solutions in network systems, applications and service computing. The Symposium builds upon the success of Frontiers in Network Applications and Network Systems (FINANS'2012) and 4th International Symposium on Web Services (WSS' 2012) held in 2012 in Wroclaw, Poland. These two events are now integrated into one event to fully exploit the synergy of topics and cooperation of research groups.

The topics discussed during the symposium include different aspects of network systems, applications and service computing. The primary objective of the symposium is to bring together researchers and practitioners analyzing, developing and administering network systems, with particular emphasis on Internet systems. Authors are invited to submit their papers in English, presenting the results of original research or innovative practical applications in the field.

Topics

Topics include (but are not limited to):

- Architecture, scalability and security of Open API solutions,
- Technical and social aspects of Open API and open data,
- Service delivery platforms - architecture and applications,
- Telecommunication operators API exposition in Telco 2.0 model,
- The applications of intelligent techniques in network systems,
- Mobile applications,
- Network-based computing systems,
- Network and mobile GIS platforms and applications,
- Computer forensic,
- Network security,
- Anomaly and intrusion detection,
- Traffic classification algorithms and techniques,
- Network traffic engineering,
- High-speed network traffic processing,
- Heterogeneous cellular networks,
- Wireless communications,
- Security issues in Cloud Computing,
- Network aspects of Cloud Computing,
- Control of networks,
- Standards for Web services,
- Semantic Web services,
- Context-aware Web services,
- Composition approaches for Web services,
- Security of Web services,
- Software agents for Web services composition,
- Supporting SWS Deployment,
- Architectures for SWS Deployment,
- Applications of SWS to E-business and E-government,
- Supporting Enterprise Application Integration with SWS,
- SWS Conversational Protocols and Choreography,
- Ontologies and Languages for Service Description,
- Ontologies and Languages for Process Modeling,
- Foundations of Reasoning about Services and/or Processes,
- Composition of Semantic Web Services,
- Innovative network applications, systems and services

Event Chairs

Furtak, Janusz, Military University of Technology, Poland

Grzenda, Maciej, Orange Labs Poland and Warsaw University of Technology, Poland

Legierski, Jarosław, Orange Labs Poland, Poland

Luckner, Marcin, Warsaw University of Technology, Poland

Szmit, Maciej, Orange Labs Poland, Poland

Program Committee

Afonso, Joao, Foundation for National Scientific Computing, Portugal

Baghdadi, Youcef, Sultan Qaboos University, Oman

Benslimane, Sidi Mohammed, University of Sidi Bel-Abbès, Algeria

Chainbi, Walid, ENISO, Tunisia

Chojnacki, Andrzej, Military University of Technology, Poland

Ciocoiu, Catalin, Orange Labs Products & Services, France

Cocucci, Osvaldo, Orange Labs Products & Services, France

Dabrowski, Andrzej, Warsaw University of Technology, Poland

Davies, John, Glyndwr University, United Kingdom

Fernández, Alberto, Universidad Rey Juan Carlos, Spain

Frankowski, Jacek, Orange Labs, Poland

Fuchs, Lothar, Institute for technical and scientific hydrology, Germany

Furtak, Janusz, Military University of Technology, Poland

Gaaloul, Walid, Institut Mines Télécom, France

García-Domínguez, Antonio, University of Cádiz, Spain

García-Osorio, César, University of Burgos, Spain

Gibert, Philippe, Orange Labs Products and Services, France

Grabowski, Sebastian, Research and Development Centre Orange Labs Poland, Poland

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 921–924
DOI: 10.15439/2014F228
ACSIS, Vol. 2

Automated Discovery of Worldwide Content Servers Infrastructure - the SNIFFER Project

Andrzej Bak and Piotr Gajowniczek
Institute of Telecommunications
Warsaw University of Technology
Nowowiejska 15/19, 00-665 Warsaw, Poland
Email: bak@tele.pw.edu.pl

Marcin Pilarski [1,2] and Marcin Borkowski [1]
[1] Faculty of Mathematics and Information Science
Warsaw University of Technology
pl. Politechniki 1, 00-661 Warsaw, Poland
[2] Orange Labs, Telekomunikacja Polska S.A.
Obrzezna 7, 02-679 Warsaw, Poland

***Abstract*—Service architecture of the Internet becomes more and more complex as it expands as a medium for large-scale distribution of diverse content. Dynamic growth of various content distribution systems, deployed by influential Internet companies, content distributors, aggregators and owners, has substantial impact on distribution of the network traffic and the scalability of various Internet services. The SNIFFER project, presented in this paper, aims to create a service for observing and tracking the long-term growth of various Internet Storage Networks (grids, clouds, Content Delivery Networks, Information-Centric Networks), using the OpenLab and PlanetLab environment. It can be useful to track and map the spreading of such Storage Networks on a global scale, providing more insight into the evolution of Internet towards a content-centric, distributed delivery model.**

I. Introduction

IN RECENT years we have observed an enormous increase in popularity of many Internet services, e.g., Facebook, DailyMotion, YouTube etc. It was possible due to an exponential growth of the number of broadband users and substantial increase in the availability of access bandwidth. During the last five years the Internet backbone traffic has been increasing at a compound aggregate rate of approximately 40% to 50% per year and for the countries of the European Union (EU) the cumulated monthly traffic ranges from 7,500 to 12,000 PB.

The increase of bandwidth usage is closely related to the growth of video traffic in the Internet, spurred by the undeniable trend towards active searching for the preferred content and watching it at the most convenient time. The success of catch-up services (iPlayer, Hulu), online movie rentals over the Internet (Netflix) and watching YouTube movies or podcasts on the TV only confirms this observation.

In order to serve the constantly increasing demand, Internet content service providers deploy content servers in multiple locations all over the world. To obtain high level of scalability and facilitate optimal distribution of popular content to geographically diverse population of end users, such content is usually distributed over multiple physical servers, for example by using the CDN (Content Distribution Networks) technology that utilizes storage located in the network. Such infrastructure, belonging to influential Internet companies, content owners, aggregators, distributors or CDN operators, consists of tenths of thousands of servers deployed throughout the world. Nowadays, it makes up a critical part of the Internet and has substantial impact on distribution of the network traffic and scalability of various Internet services beyond the first and middle mile.

Despite that, very little is known about the topologies, geographical spread, expansion and growth of systems that serve the most popular Internet content worldwide. The main objective of the SNIFFER experiment described in this paper is therefore to create a replicable base for long-running service using OpenLab and PlanetLab environment in order to better observe and track the long-term growth of Storage Networks distributing popular Internet content. The knowledge about location of the content servers and the possibility to monitor long term changes in the infrastructure deployed by popular content distributors, aggregators and owners, would allow better understanding of the nature, complexity and evolution trends of the Internet. It can be also used to improve planning of the Internet underlying transmission resources, which is important as the popular services are progressively more demanding, mainly because of the proliferation of multimedia rich content.

Similar attempts to Internet content server discovery were already undertaken, but lacked versatility (were limited to particular Internet services, such as YouTube [5] [6] [7] or CDNs [8]), sustainability and long-term observation capabilities. In the SNIFFER project we aim to achieve the above goals by developing the following elements that will constitute the final service running on the base of the PlanetLab infrastructure:

- The intercept mechanism, collecting web URLs for pattern discovery and matching to popular Internet services.
- The content server discovery mechanism, providing translation of the discovered web hostnames into IP addresses, clustering, and geo-location of discovered servers.
- The visualization service for easy access to discovered results.

The project uses common Internet protocols, PlanetLab infrastructure and capabilities of Orange Polska as the largest ISP in Central Europe to obtain a large sample of web-related customer activities. The general architecture of the SNIFFER system is presented in Fig. 1.

978-83-60810-58-3/$25.00

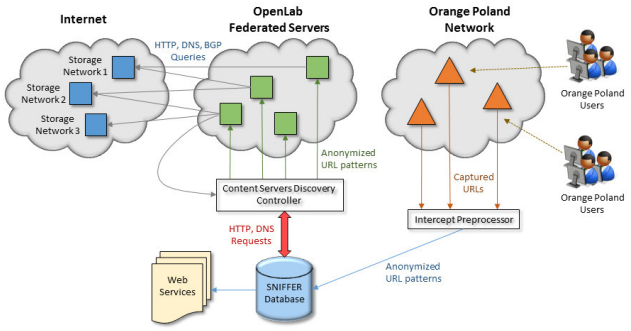


Fig. 1. General architecture of the SNIFFER system

The following sections provide an overview of the main modules of the SNIFFER system.

II. Intercept and Pattern Discovery

This module consists of two functional parts: the capture tool and the pattern discovery tool. The aim of the capture tool is to intercept user traffic and collect URLs of the visited web pages. For this task we use a server equipped with specialized DAG traffic intercept card and an adequate storage. The server is connected to the network with public IP address and 1 Gbps connection. The server also runs the TStat [12] software, functioning as a passive sniffer. It allows capturing the specified traffic at network and transport level. In case of the SNIFFER project, TStat has been prepared for logging TCP communication flows, particularly HTTP requests, as the HTTP GET method contains the URL address of the requested content and hostname of the content server.

The discovered unique URLs are stored on the project server for pattern discovery and further processing. The main function of this part is to analyze the collected URLs and generate URL patterns for selected services in a format that can be easily expanded to generate host names for web server discovery. This approach allows generating hostnames that were not actually intercepted, but for which it is probable that they will resolve to an IP address because of the similarity to some of the intercepted ones. Service selection exploits the fact that the domain names used by the internet service providers are usually equivalent to naming of the services provided to the end users.

III. Content Servers Discovery

To facilitate the conversion of discovered web host names to IP addresses, SNIFFER uses the DNS along with selected PlanetLab servers. The number of distinct IP addresses is used as an indirect measure of the quantity of content servers. A single IP address may however represent a number of physical machines that are indistinguishable for this tool without additional knowledge.

The discovery system takes hostname patterns obtained from previous task, expands them to create a larger set of hostnames, and searches for their IP addresses. The DNS servers are used to resolve the IP addresses, but for each hostname the returned IP address may depend on where the query was issued, as each local DNS can map the hostname to a different server. The goal is to discover as many IP addresses, to which a given URL is resolved in different network areas, as possible.

To obtain wide geographical distribution of queries, and a representative set of server addresses, we use the PlanetLab infrastructure. PlanetLab nodes are located in over 90 sites all over the world, therefore a huge set of local DNS servers can be queried. The system searches for both A and CNAME records. As each CNAME record is followed by complementary A record, at least one IP address is gathered for each hostname.

Popular services quite often have many A records (many IP addresses) assigned to one hostname. For example, a query for domain name *youtube.com* returns 16 A records pointing to 16 different IP addresses. Those 16 records do not exhaust the global list, as the same query executed from different host or after some time may return a different list of 16 IP addresses. Therefore, in case where multiple A records exists, all IP addresses are collected by the querying PlanetLab node.

The SNIFFER experiment does not require that all available PlanetLab nodes are used, as the data acquired by the algorithm are differentiated by geographical location, and so the responses from relatively close nodes are often similar and do not contribute much to the results. Therefore, from all available PlanetLab nodes about two from each top level domain were selected (95 total). As most of those top level domains suggest the country that the node is located in, the selection was driven towards obtaining a uniform distribution of nodes around the globe (to the extent limited by the fact that PlanetLab does not have nodes in every country). PlanetLab nodes availability varies daily, nodes go off-line for various reasons, and therefore at the algorithm initialization the list of on-line nodes is created. Usually around 80% of nodes is ready for use at the same time.

IV. Content Servers Clustering and Geo-tagging

The content server discovery tool collects thousands of IP addresses. Many of the servers behind these addresses are located in the same data centers. To get more insight into geographic distribution of the discovered servers, we employed a clustering algorithm that groups the servers together according to their approximate physical location at city level resolution. The IP address is converted to geographic coordinates using IP geo-location services. However, this approach is not sufficient to distinguish server clusters because of limited geo-location accuracy. Therefore, the algorithm also uses IP trace-route information collected from various locations around the world.

Each IP address from the set of IP addresses of the servers discovered for the particular service denotes a host. Actually, it can be a range of hosts behind NAT or a number of IP addresses located on the same machine. In case of NAT, the group can be treated as one powerful host without the loss of precision for the clustering algorithm. The second case leads to ineffective wasting of public IP addresses so this approach is most probably not used in content distribution systems.

A. Phase 1 - Collection of Gateways

For each IP address the algorithm checks the route through the Internet. The route to the host can be different if checked from different locations around the globe. The algorithm is not collecting the whole route but only the last routing device next to the host itself, called a gateway. If the last device is not discoverable, the second device closest to the target is collected, and so on. The gateway with network distance to the target IP address equal to N will be hereafter denoted by *gwN*.

The addresses of gateways leading to the same server can be different when the path is checked from different locations. The reason for this is that data centers rarely use a single edge router and may also utilize more than one ISP connection for efficiency and reliability. The algorithm collects *gwN*s for given IP addresses from more than 90 PlanetLab nodes and stores them on a dedicated server.

B. Phase 2 - Aggregation

In the second phase of the algorithm only the gateways with network distance one (*gw1*) are considered. Hosts are aggregated by the same gateways, creating clusters. In addition, the number h_N, denoting how often the host was accessible through the particular gateway, is stored.

C. Phase 3 - WHOIS Tagging

The IP address of each gateway is looked up in worldwide domain names register (known as the WHOIS database) to determine the single owner IP range (CIDR) it is in. This is necessary to group similar (belonging to the same organization) gateways later on. The names of clusters formed later are derived not from gateways but from CIDR's. Additionally, those CIDRs/ranges represent the network providers for the data centers. One issue in this process is that even if the WHOIS database is publicly available, the format of the answer is not standardized. It may return the CIDR notation (eg.201.218.32/19), but also the range (e.g., 195.182.218.0-195.182.219.255). Some WHOIS queries also fail, leading to dropping the data related to such query (however, the loss is marginal).

D. Phase 4 - Cluster Candidates

For each unique host IP, the data from previous phase is aggregated into the triple $\{$*cluster name*, *gateway list*, $h_N\}$. The cluster name is formed from all unique CIDRs from the set of triples with the same host IP. To make those names easily comparable, CIDR ranges were lexicographically sorted. The gateway list includes all gateways associated with the host IP and the h_N now represents a cumulative value for all of them.

E. Phase 5 - Cluster Geo-tagging

In this phase the location of the host part of all triples is acquired using geo-tagging tool, and the result is appended to the name of the cluster candidate, as the algorithm assumes that the physical location of the host determines the cluster position on the map. In this way some cluster candidates that have the same name will now have distinct names as they hold hosts at different locations. After geo-tagging, cluster candidates become final clusters.

F. Phase 6 - Aggregation of results

The results of cluster geo-tagging are aggregated by cluster names. The clustering process may omit some IP addresses from the input data due to trace-routing limitations. If a trace cannot find the gateway at distance one (*gw1*) it searches for more remote gateways (*gw2*, *gw3* etc.) but in phase 2 of the algorithm those gateways are filtered out. If a host is not reachable from any of the used PlanetLab nodes via *gw1* it is excluded from the clustering. To address this issue the list of left out IP numbers is processed by the algorithm once again with filtering in phase 2 changed to *gw2*. The resulting clusters are less reliable than the ones obtained in the first run, thus they are stored separately. After second run there may still be some IP addresses left, but at this stage there is no need for the third run of the algorithm with *gw3* filtering, as the number of them is usually minimal.

G. Remarks on Clustering and Geo-tagging Implementation

On each PlanetLab node the routes to tested hosts (IP addresses) are checked with the excellent *Paris traceroute* tool. The important advantage that this tool holds over the classic *traceroute* is the immunity to routers' load balancing. The whole process of checking all IP addresses for given service is rather time consuming so the cache for the queries is used and stored on SNIFFER server. A list of hosts sent to the PlanetLab node is filtered so that only the IP addresses not yet traced (from this node) are tested. The remaining traces are removed from the cache.

Clusters formed in this way should represent close estimation of real life data centers. However, as the algorithm is based on trace-routing data, it detects layer 3 network connections but cannot detect layer 2 links. Consider an example where many hosts are connected to the Internet through two gateways but the internal subnetwork (VLAN) is spanned over 3 switches, where one of them is located in a different data center and connected via a VLAN tunneling protocol (there are various technical methods to extend a single VLAN in such a way). This case can lead the algorithm to aggregation of data centers with the same gateways and different physical locations into one center.

At this point the clustering geo-tagging steps in. However, the geo-tagging accuracy varies a lot between various methods and IP databases. Not all owners of IP addresses want to reveal the exact location of the hosts, therefore the geo-tagging services and tools are imprecise by nature and evolve in time as the IP networks change. Currently, SNIFFER uses only a free of charge MaxMind GeoCity Lite database. It offers city level location service but in practice for a lot of IP addresses the tagging results are not accurate enough. Many tags can be resolved only down to a country or even a continent level. This deficiency is affecting the precision of the clustering algorithm.

V. SNIFFER VISUALIZATION SERVICE

SNIFFER web interface is available at *https://sniffer.mini.pw.edu.pl/*. The website was designed to present the most important results generated by the SNIFFER experiment in a graphical form, as the "snapshots" of worldwide content server infrastructure, taken at various points in time.

SNIFFER web pages use world maps rendered by the Google Maps V.3 engine and API (customized for the specific requirements of the project using JavaScript code). The GUI server is running on open source tools, such as Apache WWW server, MySQL database, Drupal Content Management System and other applications and libraries. It pulls the preprocessed experiment data from the data server in a daily routine, importing cluster lists, patterns, metadata related to specific run of the experiment, and special datasets prepared for comparing the results from various time points. The data is then rendered using the mechanism of Drupal views.

Data presentations accessible from SNIFFER website are created dynamically from database content which makes them very flexible. At present it is possible to visualize location of content servers of Akamai and YouTube discovered in a selected experiment, or in a form of differential maps showing changes in the discovered infrastructure between two different runs. In addition, a user can access various details and statistics of the experiment data, such as IP addresses, CNAMEs, and patterns found during the discovery process.

An example experiment executed in April 2014 took about 40h. The 315 URL patterns identified by intercepting end-user web requests were further used by the Content Servers Discovery module to search for Akamai servers using DNS queries from 76 geographically dispersed PlanetLab nodes. About 10,000 IP server addresses were discovered in result of this process, grouped into 585 clusters and geo-located to produce the map shown in Fig 2.

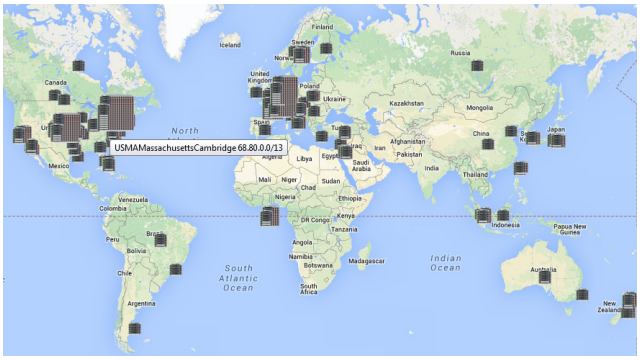


Fig. 2. Discovered locations of the Akamai servers

VI. CONCLUSION

In the paper we have described the SNIFFER project aiming to create a long-running measurement platform to monitor the location and evolution of the content distribution servers in the Internet.

The key difficulties encountered during the ongoing development of the platform were mostly related to precise clustering and geo-location of the discovered data centers. The clustering algorithm that uses *traceroute* and *whois* tools appeared more complex in practice than it was foreseen because of the difficulties in obtaining the proper gateway addresses and their actual locations. Some of the problems can be attributed to the deficiencies of the free IP geo-location database used in building the system.

During testing of the SNIFFER system in the development phase some variability in IP addresses found in consequent experiments was observed. This phenomena can be seen on differential maps and may arise in result of load balancing performed by service providers in conjunction with the scale on which they operate. Despite the fact that 95 PlanetLab servers deployed around the world were used to resolve and trace thousands of host names in each experiment, the architecture of investigated systems is so vast that each time some IP addresses fall outside the search. The providers purposeful approach to hide the actual architecture of their systems cannot be also excluded.

The experimental results from the SNIFFER project will be periodically published on the project web page *https://sniffer.mini.pw.edu.pl*. The discovery service in its current form is running from just March 2014, so the results should be still treated as preliminary. We hope however, that after SNIFFER platform is refined and its measurements database grows up, it will be useful in providing insight into evolution and growth of various Storage Networks related to popular Internet services or effective distribution of content.

REFERENCES

[1] V. Gehlen, A. Finamore, M. Mellia, M. Munafò, *Uncovering the Big Players of the Web*, Proc. TMA'12, 2012, pp. 15-28, doi: 10.1007/978-3-642-28534-9_2

[2] L. Grimaudo, M. Mellia, E. Baralis, *Hierarchical Learning for Fine Grained Internet Traffic Classification*, Proc. IWCMS'12, 2012, pp. 463-468, doi: 10.1109/IWCMC.2012.6314248

[3] A. Finamore, V. Gehlen, M. Mellia, M. Munafò, S. Nicolini, *The Need for an Intelligent Measurement Plane: The Example of Time-Variant CDN Policies*, Proc. NETWORKS'12, 2012

[4] I. Bermudez, M. Mellia, M. Munafò, R. Keralapura, A. Nucci, *DNS to the Rescue: Discerning Content and Services in a Tangled Web*, Proc. ACM IMC'12, 2012, pp. 413-426, doi: 10.1145/2398776.2398819

[5] R. Torres, A. Finamore, J.R. Kim, M. Mellia, M. Munafò, S. Rao, *Dissecting Video Server Selection Strategies in the YouTube CDN*, Proc. IEEE ICDCS'11, 2011, pp. 248-257, doi: 10.1109/ICDCS.2011.43

[6] V.K. Adhikari, S. Jain, Y. Chen, Z.-L. Zhang, *Reverse Engineering the YouTube Video Delivery Cloud*, Proc. IEEE Hot Topics in Media Delivery Workshop, 2011

[7] V.K. Adhikari, S.. Jain, Y.. Chen, Z.-L.. Zhang, *Where Do You 'Tube'? Uncovering YouTube Server Selection Strategy*, Proc. IEEE ICCCN'11, 2011, pp.1-6, doi: 10.1109/ICCCN.2011.6006028

[8] C. Huang, A. Wang, J. Li, K.W. Ross, *Measuring and Evaluating Large-Scale CDNs*, Proc. IMC'08, 2008

[9] T. Leighton, *Improving Performance on the Internet*, Commun. ACM, Feb. 2009, Vol 52, No 2, pp. 44-51, doi: 10.1145/1461928.1461944

[10] B. Wong, A. Slivkins, E. Gun Sirer, *Meridian, a Lightweight Network Location Service without Virtual Coordinates*, Proc. SIGCOMM'05, 2005, pp. 85-96

[11] B. Wong, I. Stoyanov, E. Gun Sirer, *Octant: A Comprehensive Framework for the Geolocalization of Internet Hosts*, Proc. 4th USENIX Conf. on Networked Systems Design and Implementation (NSDI), 2007

[12] *TCP STatistic and Analysis Tool*, http://tstat.tlc.polito.it/index.shtml

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 925–932
DOI: 10.15439/2014F192
ACSIS, Vol. 2

Graph Based Messaging APIs – concept and implementation

Michał Cieszko(1,2)
(1) Orange Labs
ul. Obrzeżna 7
02-691 Warsaw, Poland
(2)Warsaw University of Technology Faculty of Electronics and Information Technology
ul. Nowowiejska 15/19
00-665 Warsaw, Poland
Email:
cieszko.michal@gmail.com

Jarosław Legierski
Orange Labs
CBR
ul. Obrzeżna 7,
02-691 Warsaw, Poland
Email: jaroslaw.legierski@orange.com

***Abstract*—This paper presents an idea of new communication Application Programming Interfaces (APIs) based on graph oriented data stored in database. The use case described in paper allows the sending of traditional telecommunication messages (SMS or USSD) to dedicated people identified as best recipients based on their skills and location. A decision algorithm implemented in API, besides the organizational data takes into consideration geo-location of user' mobile phones.**

I. Introduction

NOWADAYS, with the Internet based on Web 2.0 we can observe a large expansion of social frameworks and services. Social portals such as an Facebook [1], Google+ [2], Twitter [3] and lately Instagram [4] have become very popular, especially among young people. The natural representation of connections between people linked by social networks or network of professionals (e.g. Linked In [5] or Viadeo [6]) is represented by data structures in a graphs form. This data structure in a very good way represents: people (e.g. as a graphs vertices), connection between them (graphs edges) and other parameters such as interests and hobbies (represented by attributes associated with its nodes or edges stored in graph). During the last 20 years in the Internet we can observe a large expansion of data generated by people and for people. This trend called "Big Data" in large part results generation of data sets which have the structure of a graph and can be stored as graphs attributes.

The paper is organized as follows: Section I presents a short introduction. Section II contains an overview of some related work. Section III describes the idea of a graph based messaging communication services. Section IV elaborates on our proof-of-concept prototype and system architecture. Section V presents future work, plans and intentions. The last section VI concludes this paper.

II. Existing Solutions

A. APIs exposed graph oriented data sets

Graph data sets and based on them services are used by many portals and application in the Internet. In this paper we concentrate on graph based data exposed in Application Programming Interfaces (API).

The most interesting API, based on a graph is offered by Facebook: (The Graph API [7]). This API allows: reading publishing, updating and deleting of records correlated with Facebook users activities: accounts, posts, groups, events, location and places.

Another information set (defined as a free, knowledge graph with millions of people, places, and things.) offers Google as an Freebase - Google's Knowledge Graph [8]. The Freebase APIs support a function responsible for searching and manipulating of this open database repository. Google also expose an additional data description language (Search Cookbook [9]), which allows including additional semantic information in text queries.

Another similar API based on Twitter data – GraphEdge API [10],[11] which allows for Twitter data manipulation has lately evolved to GraphEdge Pro service [12]. This solution helps PR marketing agencies track, report, and analyze the activity of their customers Twitter accounts.

The next service GraphMuse API [13] – is a solution dedicated for developers interested in the exploration of Facebook data. This easy RESTfull API returns: friends, groups, family or collective interests data sets in the JSON format.

CicerOOs Semantic Graph API [14] allows semantic searches of Points of Interest (POI) related to geographical objects (data correlated with tourism information in Italy). Apart from simple search, this API

978-83-60810-58-3/$25.00

can return the suggested possible topics of interest (e.g. restaurants, castles, galleries, parks) taken from the semantic graph.

A completely different, from above mentioned area of application is provided in the *RunKeeper's* Health Graph API [15] – which offers developers access to all of health oriented information such as: nutrition information, workouts, sleep and body measurements data, blood glucose levels etc.

The *Viadeo* [16], service which operates professional social networks provides the Graph API that allows developers to integrate professional social context with websites, applications and services. A social portal using this API offers developers access to public information on some objects: members, connections, jobs, articles, news, comments, etc.

Table 1. Exposed in the Internet graph oriented APIs

API	Functionality	Data Source
Facebook The Graph API	Data search and manipulation	Own data
Freebase Google API	Data search and manipulation	Own data
GraphEdge API	Data mining second level API	Twitter API
GraphMuse API	Data sets access simplification	Facebook API
CicerOOs Semantic Graph API	Simple and semantic data search	Own data
RunKeeper's Health Graph API	Data search and manipulation	Own data
Viadeo Graph API	Data search and manipulation	Own data

B. *Messaging applications and services based on graph data*

In the Internet some applications and services can be identified, which offer messaging functions based on information from graph oriented data sets. The *Meetbymaps.com* portal [17] offers the Instant Messaging (chat) communication service for Facebook users based on location and events (realtime geolocated chat) using the Facebook Graph API. Another portal, based on Facebook graph *WishMindr* [18] is the freely accessible virtual wish list. Using this application the end user can prepare wish lists from any site and add preferred gifts using the *WishMindr* search. *WishMindr* also offers an email communication channel used for reminders about wish lists for e.g. friends and family.

The next solution *oGoWoo* www.ogowoo.com [19] is a free advertising system. The *oGoWoo* service uses the Facebook API to post promotional and advertising messages on users Facebook walls on their behalf. As benefits, the application users get gifts and discounts. Another service *Decoda* [20] allows you to share favorite lyrics between Facebook users (e.g. marking favorite lines from a song and posting them to Facebook as a status or posting lyrics to Facebook wall). The service can also post information to a friend's Facebook wall thanks to the Facebook Graph API. Table 2 provides examples of communications applications using Facebook Graph API.

Table 2. Example communication driven applications based on Facebook Graph APIs

Application	Communication channel
Meetbymaps	IM (chat)
WishMindr	e-mail
oGoWoo	Facebook walls
Decoda	Facebook walls

C. *Telco Messaging based on API*

SMS, MMS and USSD based messaging is known in telecommunication for many years. Short Messaging Services and Multimedia Messaging Service can be originated (sent) and terminated (received) by network terminals (mainly phones). Unstructured Supplementary Service Data (USSD) can by initiated by phones (in the form of USSD codes e.g.*100#, *665*1#). Communication Service Providers (CSP) in the last decade of XX century, have started opening their networks to external developers by exposition of the Telco API in the Internet. This concept extends traditional device based messaging model to an API originated or terminated communication function. In the network architecture it appears as an additional exposed API element – Service Delivery Platform (SDP). SDP using API in Web Services form can provide external developers with a large set of communication functions from CSP networks such as send/receive: SMS, MMS or USSD, mobile terminal location, click to call, payment etc. for new innovative applications [21],[22],[23],[24]. Basic Telco APIs set is standardized by GSMA as OneAPI specification [25] based on RESTful architecture style Web Service [26]. Another specification ParlayX [27] uses Service Oriented Architecture and SOAP protocol [28].

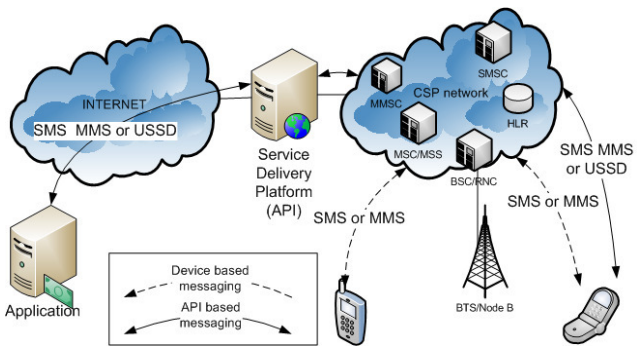


Fig 1.Telco messaging: based on devices and APIs

III. THE IDEA OF GRAPH BASED MESSAGING SERVICES

This chapter presents the idea of Graph Based Messaging (GBM) services exposed in APIs form. There are many solutions used in a single system that allow the communication of groups of people depending on their status, location or additional features e.g. the solutions based on the Unified Communications paradigm [29],[30], [31] or applications described in chapter II. The application based on the Unified Communication are dedicated to enterprise sector and mostly allow end users communicate using: voice calls, voice and web conferences, or chats. In UC, a system's subscribers are addressed using names, e-mails or phone numbers and applications are trusted and private from an enterprises point of view. Companies in the B2B sector have a lot of information which is graph oriented: organizational structures, information about calls and actions from private telecommunication or IT systems. This information can be very useful for the development of new applications and services but should be extracted from many ICT systems, structured and exposed in the form of dedicated Application Programming Interfaces. The second problem is related to the confidentiality of the exposed data. A dedicated API should also isolate sensitive private information which greatly simplifies the development of an external application from a legal point of view. The solution described in chapter 2 solutions use a graph API exposed in the public domain (in the Internet) and exposes a user's identifier (e.g e-mails or stored in Facebook data) based on permissions granted by users.

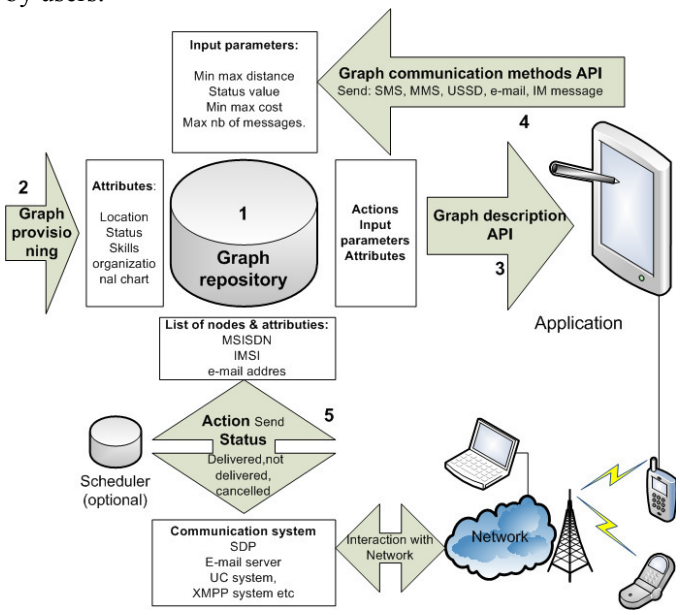


Fig 2. Idea of a Graph Based Messaging APIs

In the solution presented above the main element is **Data Repository** (1) – this is a database which contains the graph oriented data sets. Next, a very important is the **Graph provisioning** (2) – an interface for graph data sets maintenance (add/remove nodes, attributes), this element can be implemented in the form of: a web interface, an API or a batch file.

Graph description API (3) – is a high level API for external application defined as: a graph structure, graph attributes, and possible actions. By using this API an external application can automatically get graph meta data description.

Graph communication methods API (4) – a set of APIs dedicated to invoking some actions (e.g. send messages, display graph data sets etc.)

Actions (communication functions) (5) – executed by a graph messaging service, interacts with communication systems such as e.g.:

- Service Delivery Platform – sending SMS, MMS, USSD, Terminal Location,
- SMPT server – sending and receiving e-mails,
- UC system – read/set presence status , monitor office phone calls, invoke voice and web conferences etc.
- XMPP server – dedicated for sending Instant Messaging information (chats).

The idea of an API based on graph oriented communication services presented in this paper allows developers to think about communication more widely. The presented concept gives external developers a simple programming interface and provides following advantages:

- sensitive user data (e-mails, MSISDN etc.) is not exposed,
- developers do not need to implement graph specific algorithms on the application side (e.g. clustering, shortest path)
- developers are isolated from telecommunication specific protocols

Using the above presented concept allows developers to create applications in a simpler way. The majority of useful and needed (e.g. mobile) applications were invented by people from normal society. Graph oriented data exposed by social portals, communication providers or enterprises has a lot of potential. Developers, companies and people from societies are looking for a new ideas and applications. Graph based messaging API is the answer to computerized society needs.

IV. TEST DATA AND MEASUREMENT

As a test data set, in the presented in this paper proof of concept of a Graph Based Messaging APIs, the

organizational structure of the department of Open Communication Systems and Open Data Department in Orange Poland R&D Center was used. This structure is presented in Fig 3 and contains:

- graph vertices- represented by people (names)
- graph edges – based on ongoing projects and other activities virtual teams
- graph attributes - associated with vertices: employees: e-mails, phone numbers, terminal location and people skills. Skills are defined in following form:

DEV - developer
IN – IN network specialist
API – API expert
SDP – Service Delivery Platform expert
BD – Big Data expert
LOC – mobile terminal location specialist
DM – Data Mininig expert
STAT – Statistics
STD - Student

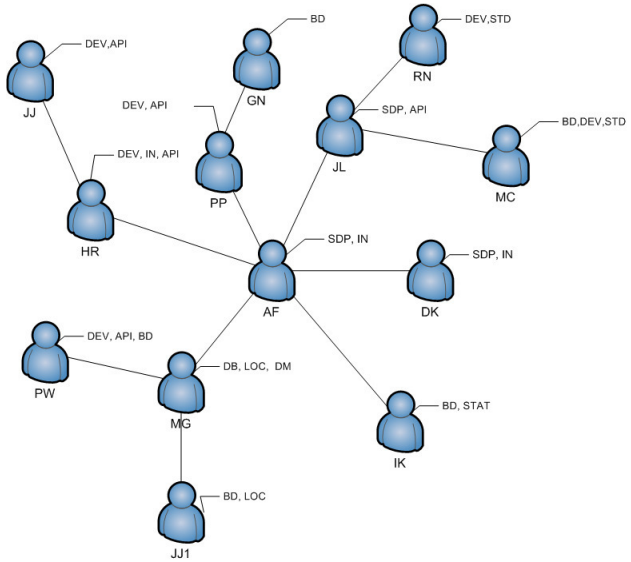


Fig 3. Test data set

Mathematically, a graph G is represented as G(V, E, A) where V are vertices (people), E a set of edges connecting some vertex in V (virtual teams), A set of attributes A={a1,a2, a3 , …an} is defined as list of skills, contacts (phone numbers, e-mails) and mobile terminals geolocation coordinates.

V. SYSTEM ARCHITECTURE

In this section a Graph Based Messaging API system architecture realizing messaging functions is described. The developed system, based on a graph database as a source of information, offers end users messaging functions such as SMS, USSD or e-mail in an Application Programming Interfaces form. The project integrates Gmail, Google Maps and Orange API services and creates a coherent communication system which it's high level architecture is presented in Fig.4.

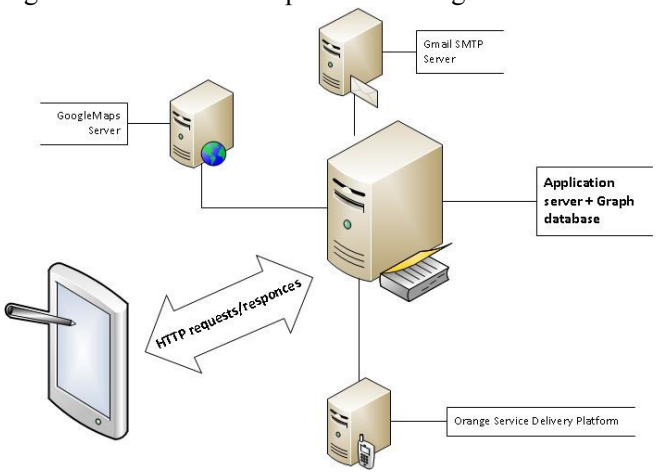


Fig 4. A high level architecture of the graph based messaging API

A. *Programing environment*

The Graph Based Messaging API system was developed in JRuby [33] – a Java implementation of the Ruby programming language and deployed on Linux Ubuntu server. The application server is exposed in the Internet using a public IP address. JRuby provides a set of classes and Standard Libraries for the Ruby language. The JRuby environment also offers a possibility of downloading many additional libraries (External Gems) created by independent developers and external companies. An example of an external gems, used during system implementation, are libraries that connect system with an open-source Neo4j graph database developed by Neo Technology. An embedded version of Neo4j database in High-availability Cluster mode was used in the graph based messaging API system as data repository.

B. *Database*

The system presented in this paper system was developed based on the Neo4j database in community open source version dedicated to the creation and testing of application prototypes. The Open Communication Systems and Open Data Department structure presented in Fig.3 was used as a test data set. A traditional, plain relational data model will not be efficient with these kinds of data sets because of a large number of relations between people which is characteristic for social networks. In comparison with relational databases, graph databases such as Neo4j [34] are faster and generate less processor load. Neo4j in the embedded mode offers an opportunity to store, on the local server, data consisting of users, places and competences entities with basic information like personal data or dynamic parameters defining unique terminal location or its actual status. Each of the users (vertices, nodes) is in relation with others by

assignment to a common workplace or some competencies shown in figure 5. All relationships are stored as graph attributes in both of the related entities, defined as incoming or outgoing. A graph oriented structure allows access to all nodes connected by a relationship with and additional set of parameters. From a data model point of view, operations like adding new relationships or attributes are easy to perform and do not demand a reorganization of data structure.

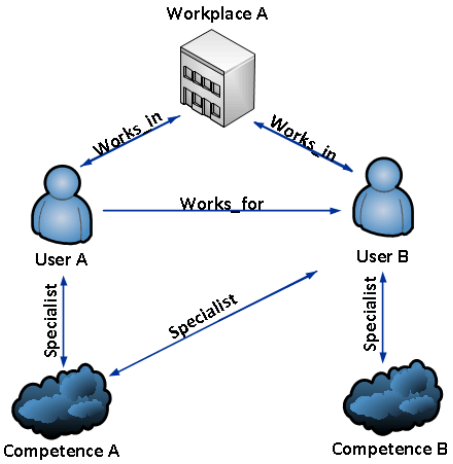


Fig. 5. Model of relationships

C. *Graph data provisioning - Web Application*

A dedicated web application was developed for end-user graph data provisioning and managing. This application allows for the creation of personalized accounts and edition and deleting of user data. Using web forms [Fig. 6] it is possible to set or modify data necessary to create or enlarge an user account. An interesting function is the definition of the location of workplaces/meeting places, which allows the implementation of messaging services based on user location. What is more, the system and proposed data model is extensible and its functionality can evolve depending on user needs.

Fig. 6. Web form interface

D. *Data Flow*

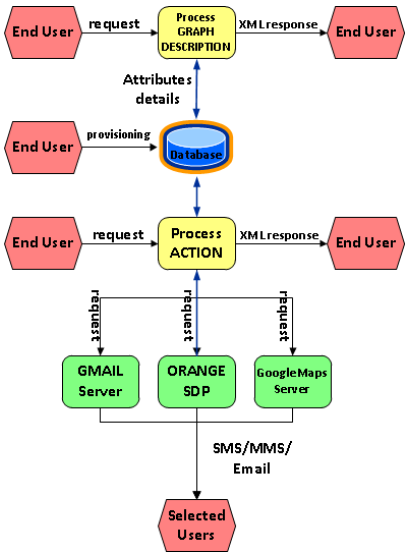


Fig. 7. Data flow in system

The model presented in fig. 7 shows the system data flow used to realize all implemented API messaging methods.

E. *Graph description API*

The Graph description API module exposes in RESTlike Web Service form, information about all the metadata (nodes attributes and communication functions) exposed by the system. The end GBM Application user (mainly the application used GBM API) can request, by using an URL, all actual graph attributes and methods that are possible to perform. The Graph description API return data in the presented below XML format.

```
<GraphDescriptionAPI>
<functions>
        <function>
                <functionname>SendSMS-
        competence</functionname>
                <link>
                http://80.48.104.236:8093/graph_api/SMS?
        competence=%&msg=
                </link>
                <description>
                Sending SMS messages to users with
        specified competencies
                </description>
        </function>
        . . .
        </function>
</functions>
<attributes>
        <competence>
                <value>DEV</value>
                <description>Developer</description>
        </competence>
        . . .
        <location>
                <value>BARBARY</value>
                <description>ul.Św.Barbary2</description>
        </location>
        . . .
</attributes>
</GraphDescriptionAPI>
```

In the presented above structure, the following functions and attributes described in Table 2 are defined as XML elements.

Table 2. Graph description API XML elements

XML element	Description
GraphDescription API	XML root element
functions	List of communication functions exposed by Graph
Function	communication function
functionname	Name of communication function e.g send SSMS or USSD
Link	URL for the server exposed communication function
description	Function description
attributes	List of graph attributes which can be used as parameters values in Graph communication functions
competence	Attribute name in this case – competence (skill)
Value	Attribute value (e.g. BD – Big Data)
description	Attribute value description (e.g. – Big Data)

The existing and exposed in graph database functions and attributes of the API presented in this chapter are similar to the concept of the Web Services Description Language (WSDL) for SOA or WADL (Web Application Description Language) proposed for RESTful Web services, but unlike them functions and attributes are dynamically changed with the change of contents of the database (for example, after adding new competences - employees skills) - they appear in the Graph description API as new attributes.

F. *The Graph communication API methods*

The Graph based messaging offer three communication functions: Send SMS, Send USSD and send e-mail as an easy RESTlike methods, which can be used in external applications for invoking communication between people which data contains an implemented graph database.

Sending SMS or USSD messages can be executed by calling an URL with the following syntax:
http://80.48.104.236:8093/graph_api/SMS?competence=value&msg=value
where:
competence - is the list of people skills (described in chapter 4)
message – is a text message sent to a group of people e.g:
http://80.48.104.236:8093/graph_api/SMS?competence=BD,SDP,LOC,STAT&location=OBRZEZNA&msg=Hello

An application server parses parameters and checks graph database content. A communication action (sending SMS or USSD messages) is processed by SDP (via Orange API) to the set of selected users. After a successful sending process, the API return a XML response with summary information about the number of informed people shown below:

```
<objects type=”array”>
   <object>
      <status>Success</status>
      <method>SMS</method>
      <ammout>5</ ammout>
      <message>Hello</ message >
   </object>
</objects>
```

The email sending process is similar to the SMS and USSD messages. The send email action has its own JRuby controller method which is accessible by the URL bellow.

http://80.48.104.236:8093/graph_api/Email?competence=value&location=value&message=value
where:
competence - is a list people skills (described in chapter 4)
message – is a text message sent to a group of people
location – a list of predefined location where An application will search for people who are located in a particular place
e.g:
http://80.48.104.236:8093/graph_api/Email?competence=ALL&location=OBRZEZNA&message

Calling this method allows the user application to send with the SMTP protocol an email messages to the list of selected people based on employees skills.

Successfully delivered messages are summarized in a XML response shown bellow.

```
<objects type=”array”>
   <object>
      <status>Success</status>
      <method>Email</method>
      <ammout>3</ ammout>
      <message>Hello</ message >
   </object>
</objects>
```

An additional special email API (DANGER e-mail type)

http://80.48.104.236:8093/graph_api/Email?type=DANGER

informs about a danger situation of the person which called the API function (all information about the person who invoked the API is collected from the username by using the HTTP Basic authentication method). In this special case, the graph application sends to the all the people in the nearest location a special e-mail containing a GoogleMap picture with a marker pointing to the location of the person who send requested as shown in Fig.8.

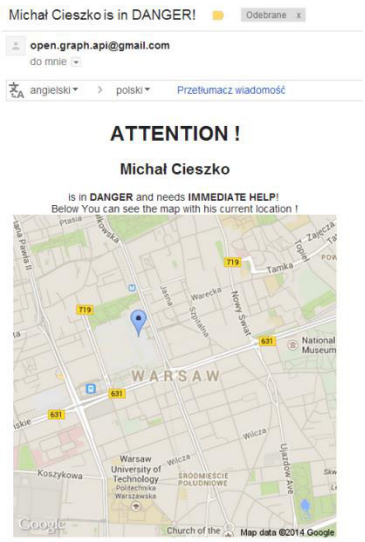


Fig. 8. An example DANGER type e-mail message

The presented above special type of GBM location API can be very useful e.g. in case a health problem with the person sending the request when immediate help is mandatory. In this case the API might be invoked by a mobile application using build in hardware like accelerometer, thermometer etc.

G. *End User applications*

The Graph Based Messaging concept includes functionality dedicated to wide range of applications and services. From the security point of view, API (using Base authentication method) invocations are processed only for correctly authenticated API users. The API was developed in a RESTlike architecture style. For an API exposed by the Graph Based Messaging system, the authors have developed a few example applications. The first one is a Windows application which offers the sending of SMS, USSD or e-mail messages to the users having defined skills or located in dedicated places.

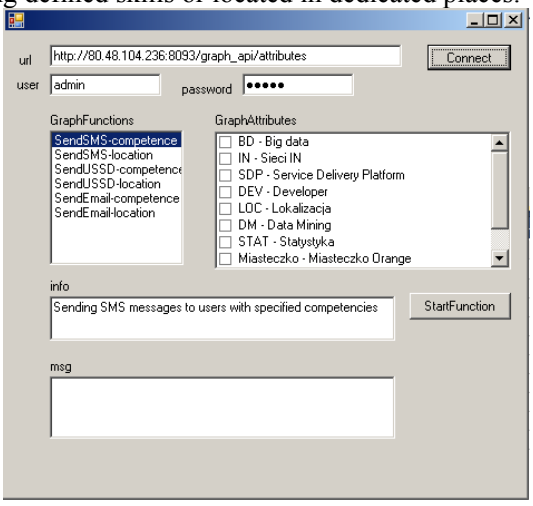


Fig. 9. Windows Client Application

After start, the application end user can choose any of the available *GraphFunctions* and then select some of the allowed *GraphAttributes*. The *GraphFunction* description appears inside the *info* section. Action (sending of messages) can be invoked using the *StartFunction* button which makes a HTTP request and invoke the selected Graph communication API method. A response received from the API is presented in the *info* field as is shown in Fig.11.

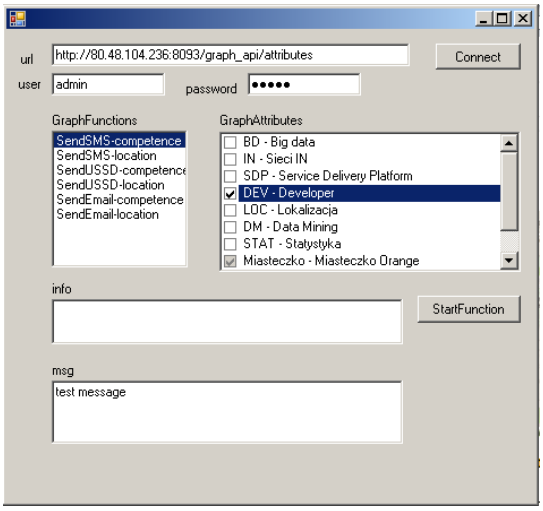


Fig. 10 Windows Client Application – an example function call

The second end user interface is dedicated web application. This subsystem is based on a web graphical user interface and offers additional features which makes it easier to invoke some actions (e.g. sending SMS/USSD/Email), check the system status or visualize a mobile terminal's (and themselves phone user) locating accuracy [Fig.11].

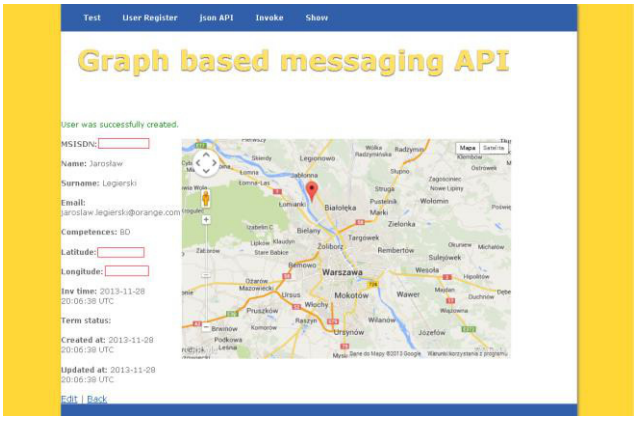


Fig. 11. End user Web GUI

Values of competence or predefined location parameters are store in database and represent the set of users which are related to requested skills and locations.

VI. FUTURE WORK

In the next system version the authors want to integrate

the Graph Based Messaging system with Facebook Graph API as a source of information about users, friends and other attributes, which can be very useful e.g. in e-mail messaging in case of emergency (DANGER e-mail API described in chapter IV F). The next functionality is a provisioning method, which can import and create a graph structure based on data from a text file provided by business clients. In future system versions there are also planned to be implement personal features (as stored in graph additional attributes e.g. users calendars), to extend the area of application usage.

VII. Summary

The Graph Based Messaging system is a prototype which presents the potential area of usage the traditional telecommunication channels connected with social networks technologies such as graph databases. The described communication oriented functionality in the API form allow to solving of many problems:

The system isolates user data (such as MSISDN number, e-mail) from the API requester and themselves thus addressing legal problems (sensitive data are not exposed). All data is processed on the server side, client applications are light and not generate client processor load from a computational point of view.

Client applications are not complicated from a mathematical point of view, graph oriented algorithms (clustering, shortest path) are implemented on the server side and the end result of them is available via the API.

The large advantage of the presented solution is the implementation of Neo4J – dedicated graph database. Relational databases generate many constraints. Programmers have found it difficult to append new objects and relationships to existing solutions while keeping system stability at the same time. The Graph based messaging system is a solution providing good performance while keeping all user data in full secret. What is more, the end user (API client) does not need to know the dataflow inside the system and details about implemented algorithms. The Author's aim was to provide the concept of the solution based on graph data structure with a wide range of functions and easy to implement in different business usage scenarios.

The prototype of the Graph Based Messaging System has been developed under the Orange Labs Open Middleware 2.0 Community program [35].

References

[1] Portal Facebook https://www.facebook.com/ [28.11.2013]
[2] Portal Google + https://plus.google.com/ [28.11.2013]
[3] Portal Twitter https://twitter.com/ [28.11.2013]
[4] Portal Instagam http://instagram.com/ [28.11.2013]
[5] Portal LinkedIn www.linkedin.com [28.11.2013]
[6] Portal Viadeo http://pl.viadeo.com/en/ [28.11.2013]
[7] The Graph API , Facebook https://developers.facebook.com/docs/graph-api/ [27.11.2013]
[8] Freebase API https://developers.google.com/freebase/ [28.11.2013]
[9] Search Cookbook https://developers.google.com/freebase/v1/search-cookbook [28.11.2013]
[10] GraphEdge API http://www.programmableweb.com/api/graphedge
[11] Waldron Faulkner GraphEdge Blog http://waldronfaulkner.wordpress.com/ [28.11.2013]
[12] GraphEdge professional http://agency.graphedge.com/index.php [28.11.2013]
[13] GraphMuse API http://graphmuse.com/api#introduction [28.11.2013]
[14] CicerOOs Semantic Graph API http://www.ciceroos.it/api-documentation [28.11.2013]
[15] Health Graph API http://developer.runkeeper.com/healthgraph [28.11.2013]
[16] Viadeo Graph API http://dev.viadeo.com/documentation/ [28.11.2013]
[17] Portal Meetbymaps.com http://www.meetbymaps.com/en [17.01.2014]
[18] Portal WishMindr http://www.wishmindr.com [17.01.2014]
[19] Portal oGoWoo www.ogowoo.com [17.01.2014]
[20] Portal Decoda www.decoda.com [17.01.2014]
[21] Podziewski, A.; Litwiniuk, K.; Legierski, J., "Emergency button — A Telco 2.0 application in the e-health environment," Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on , vol., no., pp.663,677, 9-12 Sept. 2012
[22] Wawrzyniak, P.; Korbel, P.; Borowska-Terka, A., "Student information delivery platform using telecommunications open middleware APIs," Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on , vol., no., pp.871,874, 8-11 Sept. 2013
[23] Korbel, P.; Wawrzyniak, P.; Grabowski, S.; Krasinska, D., "LocFusion API - Programming interface for accurate multi-source mobile terminal positioning," Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on , vol., no., pp.819,823, 8-11 Sept. 2013
[24] Korbel, P.; Skulimowski, P.; Wasilewski, P.; Wawrzyniak, P., "Mobile applications aiding the visually impaired in travelling with public transport," Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on , vol., no., pp.825,828, 8-11 Sept. 2013
[25] GSMA OneAPI http://www.gsma.com/oneapi/ [17.01.2014]
[26] L. Richardson, Sam Ruby, David Heinemeier Hansson, RESTful Web Services, O'Reilly, 2007
[27] Open Service Access (OSA); Parlay X web services, 3 GPP, http://www.3gpp.org/ftp/Specs/html-info/29199-01.htm, [17.01.2014]
[28] SOAP specification W3C, http://www.w3.org/TR/soap [17.01.2013]
[29] "LifeWorks Die Kommunikation der Zukunft beginnt", Information and Communication Networks A50001-N1-S20-1 bySiemensAG2004
[30] "Siemens ICN Announces LifeWorks Concept for Integrating Personal, Enterprise and Carrier Technologies", http://www.prnewswire.com/news-releases-test/siemens-icn-announces-lifeworks-concept-for-integrating-personal-enterprise-and-carrier-technologies-71170057.html
[31] D.Bogusz, J.Legierski, A. Podziewski, K. Litwiniuk, „Telco 2.0 for UC – an example of integration telecommunications service provider's SDP with enterprise UC system", Wrocław 2012.
[32] A.C, Weaver, B. B. Morrison, Social Networking, Computer, Volume: 41 , Issue: 2, 2008
[33] JRuby portal http://jruby.org/ [24.03.2014]
[34] Ian Robinson, Jim Webber, Emil Eifrem, Graph Databases, O'Reilly Media, 2013
[35] Open Middleware 2.0 Community Program portal www.openmiddleware.pl [30.03.2014]

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 933–940
DOI: 10.15439/2014F241
ACSIS, Vol. 2

Requirements for IMS services and applications over interoperable broadband Public Protection & Disaster Relief Networks and Commercial Communication Networks

Henryk Gierszal
Adam Mickiewicz University, Umultowska 85, 61-614 Poznań, Poland
Email: gierszal@amu.edu.pl

Anna Stachowicz,
Filip Majerowski,
Bartłomiej Kowalczyk,
Michał Goryński
ITTI, Rubież 46, 61-612 Poznań, Poland
Email: {AStachowicz, FMajerowski, BKowalczyk, MGorynski}@itti.com.pl

Vassilis Kassouras
KeMeA,
4 P. Kanellopoulou,
Athens 101 77, Greece
Email:
kassouras@gmail.com

Spase Drakul
THYIA Technologies Sarl, Chemin de Roilbot 19, Pregny-Chambésy, Switzerland
Email:
drakul@bluewin.ch

***Abstract*—The paper presents requirements related to a heterogeneous interoperable and transportable gateway called HitGW that is designated to connect different incompatible communication systems used by First-Responders (FRs) and to make them interoperable with IP-based networks. The HitGW allows IMS type of services and applications over a broadband network composed of ordinary Public Protection and Disaster Relief (PPDR) networks and Commercial Communication Networks (CCNs). Such a system solution for FRs becomes indispensable for interoperable, effective, secure, and safety communications especially in national and cross-border crisis scenarios. The innovativeness in the chosen approach to design and implement such a system gateway primary consists in the user needs and current trends of PPDR networks toward 4G LTE type of networks and services.**

I. Introduction

THE purpose of this document is to specify and describe some gathered user requirements in scope of the Hit-Gate (Heterogeneous Interoperable Transportable GATEway) project and to present a system concept how to use IP Multimedia Sub-System (IMS) as a component of the gateway for First-Responders. The main objective of the Hit-Gate project is to develop a generic gateway that allows communications across networks currently used by FRs in Europe especially during crisis events. Nowadays PSC (Public Safety Communication) systems use a large number of different and incompatible technologies therefore compromising or even disabling efficient coordination of combined operations (such as cross-border or crisis management when many services/utilities are involved) [1], [8], [13].

System solutions targeted by the Hit-Gate project are intended to be used by FRs and public safety entities, that is why they should be desirable and approved by the end-users and domain experts. To specify Hit-Gate user requirements, research in the area of general requirements for PSC systems have been performed first. Next, basing on identified and gathered requirements, they have been presented and consulted with end-users and domain experts in the form of questionnaire, brain storm discussion, and available references (state-of-the-art).

To prepare a list of end-user requirements we have applied VOLERE ([2], [3]) methodology that was however empowered by interactions and feedback from end-users. It allows the requirements to be more objective. The requirements have been defined based on consortium knowledge and experience, end-users/FRs and domain experts' opinions, available knowledge about general requirements for PSC systems, as well as literature survey. A collecting process included desktop research in the area of Public Protection and Disaster Relief (PPDR) systems, questionnaire for end-users and domain experts, and internal consortium discussions. It allowed grouping and prioritizing identified requirements.

For development of an interoperable gateway that meets end-users' and other system requirements the proposed solution is based on the 3GPP IMS architecture where Session Initiation Protocol (SIP) is used for control and signaling and Real-time Transport Protocol (RTP) for delivering voice, messages, and video over IP-based networks.

The paper is organized in the following manner. In Section II a short description of the Hit-Gate project and main features of HitGW are presented. The methodology of requirement collection and identification is described in Section III. Section IV includes general objectives of the Hit-Gate project which have great influence on requirements specification, while Section V comprises main requirements for the Hit-Gate system. The system architecture is demonstrated in Section VI. A trial implementation of the IMS platform for GSM network based on the proposed system architecture is presented in Section VII. In Section VIII conclusions are given.

This work is done within EU 7th Frame Programme - FP7 SECURITY, no. 284940

II. HIT-GATE PROJECT

The interoperability problem was perceived by EC a few years ago [19]. The main emphasis was put on spectrum interoperability across Europe ([20], [21], [24]) for broadband PPDR networks [22] that meet users' requirements [23]. These efforts have been also taken by many EU FP7 projects which aim at developing solutions that can deliver the interoperability among different PPDR systems. An objective of some projects is also to define a roadmap towards interoperability for PPDR agencies in Europe to support any cross-border events.

The Hit-Gate project is an EU FP7 SECURITY two-year small or medium-scale focused research project. Project Consortium is composed of eleven organizations from eight European countries ([18]). It includes large companies, Research and Technology Organizations (RTO), Small and Medium Enterprises (SME), and an external community of end-users. Consortium experience and knowledge in the area of development of PSC and security (mission critical) systems together with end-users support enables to develop solution, which is adequate to today's needs and covers lacks of interoperable technologies that offer interoperable IMS services. The main project goal is to develop a novel system solution for interoperability between access network elements (NE) (mobile terminals, base stations) of the First-Responder Networks (FRNs) and access network elements of the existing CCNs without modifications of handset devices or communication infrastructure of both network types. Because at both European and national/domestic levels, public-safety organizations have adopted a variety of systems, equipment, and incompatible technologies (different technologies, standards, and proprietary solutions of the system manufactures). To answer the FRs needs, HitGW is developed to support all those incompatible technologies ranging from legacy-PMR (Professional Mobile Radio), TETRA (TErrestrial Trunked RAdio) and TETRAPOL, to next-generation networks (e.g. LTE and LTE-Advanced). Other types of networks like GSM (2G), UMTS (3G), 4G LTE, WiFi and WiMAX can be used together with FRNs. Moreover, current security and emergency activities frequently involve multi-national FR teams (e.g., natural crisis response and cross-border operations) [10], [11], [12], thus the HitGW will enable efficient coordination of operations involving more than one nation. The main goal of this system gateway is to allow interoperability among NEs of the FRNs and CCNs used for PSC. It should:

- ensure mission critical requirements of PSC applications, i.e. high-availability, dependability, and security;
- be rapidly deployable over mobile, highly-dynamic, and unpredictable environments where existing infrastructures may be degraded and/or destroyed [4].

The Hit-Gate (HG) framework also includes system requirements and specifications, system architectures, development of the HitGW subsystems, system integration and validation, demonstration of the Hit-Gate Network (HGN), and standardization.

III. METHODOLOGY FOR GATHERING REQUIREMENTS

In order to gather the requirements from the consortium and end-users involved in Hit-Gate project, it has been considered necessary to use a suitable methodology. Thus, the VOLERE methodology has been chosen as a starting point for this work, adapting its requirements specification template to the particularities and needs of Hit-Gate.

Each of the partners involved in the project follows their own processes in the requirements definition phase performed within their activities. However it was important to propose a common way to formalize the requirements that is easy to use and adapt to the needs of the project by all partners. In this context, VOLERE is a straightforward methodology that does not require a complex analysis to be applied. Furthermore, it guarantees the participation of all relevant actors, who are further involved in the design and development that have to meet the requirements defined.

The adapted methodology allows for identification and formalization of unambiguous requirements, as well as assessment the correctness of the requirement in order to avoid of a lack of completeness and coherence.

The use of a common methodology to gather, classify, and assess the requirements a priori was important. The management of the requirements depends on this common methodology, providing the means to trace the identification, definition, assessment, formalization, and if necessary improvement of the requirements gathered.

On the other hand, the requirements should be the key to evaluate the entire project at the end of the development phase. A set of well-defined and unambiguous requirements is needed, not only as input for any further specifications and development, but also as part of the evaluation framework.

VOLERE defines the gathering process and the shell to register the requirements, classified in 27 categories in 5 main groups:

1. Project drivers, the business-related forces. For example, the purpose of the project is a project driver, as are all of the stakeholders — each for different reasons.
2. Project constraints, restrictions on how the product must be designed. For example, it might have to be implemented in the hand-held device being given to major customers, or it might have to use the existing servers and desktop computers, or any other hardware, software, or business practice.
3. Functional requirements, the fundamental or essential subject matter of the product. They describe what the product has to do or what processing actions it is to take.
4. Nonfunctional requirements, the properties that the functions must have, such as performance and usability. These requirements are as important as the functional requirements for the product's success.

5. Project issues, the conditions under which the project will be done. The reason for including them as part of the requirements is to present a coherent picture of all factors that contribute to the success or failure of the project and to illustrate how managers can use requirements as input when managing a project.

VOLERE methodology is an efficient way to describe requirements. In this document focus is on both the user requirements from the perspective of end-user needs and the system requirements that are assumed to be derivative of these former ones as well as other identified ones during the architectural framework. We assume that it is not possible to present only user view on the Hit-Gate project because users are not aware of many important system requirements, constraints, and other network requirements for developing a heterogeneous Hit-Gate Network of Networks (HG NoNs). Here, we will deliver an assemble of overall requirements collection consisting of end-user requirements and some high level functional and nonfunctional requirements in groups (chapters) more related to HitGW and its architecture. They can be divided into the following groups:

- Communication Requirements
- Data Requirements
- Usability Requirements
- Performance Requirements
- Operational Requirements
- Security Requirements
- Legal Requirements.

VOLERE specifies description of particular requirement with following characteristic:

A. Requirement Numbering — give each requirement a unique identifier to make it traceable throughout the development process. It is suggested that the numbering should include:
- Requirement number — the next unique requirement number;
- Requirement Type — the section number from the template for this type of requirement. This field serves as a reminder of what this requirement relates to and helps to compare requirements of the same type leading to detection of contradictions and duplications.

B. Event/use case number — the identifier of a business event or use case that contains this requirement. There might be several event/use case numbers for one requirement because the same requirement might relate to a number of events/use cases.

C. Customer Value — a measure of how much end-user cares about each requirement. It indicates the level of customer (end-user) satisfaction, if a given requirement will be implemented. It may help in prioritizing requirements and then system/solution's functions knowing which of them are the most important and desirable from customers (end-users) point of view. It helps looking at from different perspectives, and to uncover what they care about most deeply.

D. Priority — importance of the requirement.

E. Dependencies — keep track of other requirements that have an impact on this requirement.

F. Conflicts — keep track of other requirements that disagree with this one.

For this research we use simpler requirement cards with elements most suitable in this project. The fields in a questionnaire of the requirement template are:

- 'Id' is an identification of the requirement. It also includes a requirement type that precedes the requirement number. The requirement type is created as an abbreviation of <level 2> and <level 3> of sections (linked with a dash) that group requirements into types.
- 'Importance (priority)' (High/Medium/Low) tells how a requirement is important for end-users. "High priority" requirements are direct requirements of Hit-Gate, "Medium priority" ones are only desired requirements, and "Low priority" ones are optional.
- 'Source' indicates to references to a given requirement.
- 'Version' shows the evolution of the requirement.
- 'Speed' (Normal/Urgent) tells about a need of availability of the requirement in time during a mission.
- 'Reliability' (Normal/High) specifies the reliability of implementation of a given requirement. "High" means that there should be a confirmation of receipt of communication.
- 'Auditable' (Normal/High) denotes that the requirements should be tracked during operation.
- 'Security' (Standard/Enhanced) gives a level of security that the requirement should have.
- 'Quality' (Normal/High) tells about the quality of performing the requirement.
- 'Title' is a requirement name.
- In the 'Description' field the requirement is explained.

The process of defining end-users' requirements for the Hit-Gate project is split into four stages. Using existing different sources delivered by partners and found in available bibliography resources the preliminary list of requirements is collected in the stage I (from e.g. [1], [4], [5], [6], [7], [9]). These requirements are then updated by partners basing on their knowledge and experience in the stage II. In the stage III end-users are involved to review existing requirements and to define new ones. It is done during meetings at national level organized by consortium's partners. Then the collected information is merged in the stage IV.

IV. GENERAL OBJECTIVES

The Hit-Gate project defines the following general objectives for the HitGW that can be treated as baseline requirements [4]:

O1 it shall enable communications between First-Responders' heterogeneous networks used at European Level;

O2 it shall provide set of services required to meet First-Responders' needs across First-Responders' heterogeneous networks;

O3 it shall provide cross-network services to First-Responders in a seamless way;
O4 it shall be secure and shall comply with pre-established security policies;
O5 it shall be transportable, rapidly deployable, and autonomous;
O6 it shall comply with mission critical requirements;
O7 it shall automatically integrate new networks of known type;
O8 it shall require minimal or no changes to existing public safety communication (PSC) infrastructures;
O9 it shall provide a modular architecture allowing incorporation of future network types;
O10 it shall provide open-interfaces and will provide recommendations for standards;
O11 it shall address recommendations for interoperability dealing with operational, organizational, and legal aspects.
O12 it should ensure information exchange among First-Responders' networks.

The underlying requirements are [2]:

- Communications connectivity,
- Communications reliability,
- Communications deployability,
- Communications transparency: i.e. end-users are blind to the presence of communications, and only interact amongst each other at the information layer on the basis of "send information and forget how it gets there".

V. End-user requirements and needs

Identified functional requirements are shown in Table I.

Identified nonfunctional requirements (e.g. quality, efficiency, etc.) related to provided services are:

- for voice: Clarity of communications (no noise); High quality voice codecs;
- for data: Broadband data access; Enabling carry peak data rate.

The most important requirements identified in the process described in Section III are:

- PSC system in a typical operation should ensure connectivity Scene Command (tactical level) with Incident Command and with data centers (operational level) following the chain of command;
- There is a need to establish standard operating procedures;
- The interoperability gateway should not limit the coverage of communication provided by interoperating radio systems used on site during mission to an incident;
- The gateway should support from legacy to current to next-generation broadband networks used today by FRs;
- End-users of all wireless systems should have mobility as offered by a given system used on the scene;
- Interoperability gateway shall be developed in a way that allows adding new network (of known type) in a modular way, therefore ensuring an always up-to-date gateway;
- Interoperability gateway shall provide open-interfaces and a wide range of services across networks, based on widely adopted standards. These include IP (Internet Protocol) as basis protocol (for level 3 OSI interoperability), IPTV, VoIP, (Voice over IP), and SIP/SDP (Session Description Protocol) for signaling;
- PSC system in a typical operation should ensure inter communication between different entities, and these entities with Scene Command;
- Interoperability gateway should enable communications to FRs during operations across heterogeneous networks, using their current receiver equipment and with no impact on currently used PSC base-stations and infrastructures other than providing a connection-point to interoperability gateway;
- Interoperability gateway should automatically integrate new networks of known type using such functionalities like: (i) 'plug'n'play', being the ability for new networks (of known type) to be connected to interoperability gateway and become ready to use; and (ii) 'on-the-fly', being the interoperability gateway ability to automatically adapt and work with newly connected networks (of known type) without need to temporarily stop services;
- Interoperability gateway should be developed adopting methodologies defined for critical security systems, shall use secure technology and shall use proven security techniques for comprehensive end-to-end security across all networks;
- PSC system in a typical operation should ensure intra communication in particular entities — in Local Headquarters and between Local Headquarters and Remote Headquarters;
- There is a need for a set of standards for public safety services.

The most important requirements with their score are shown in Fig. 1.

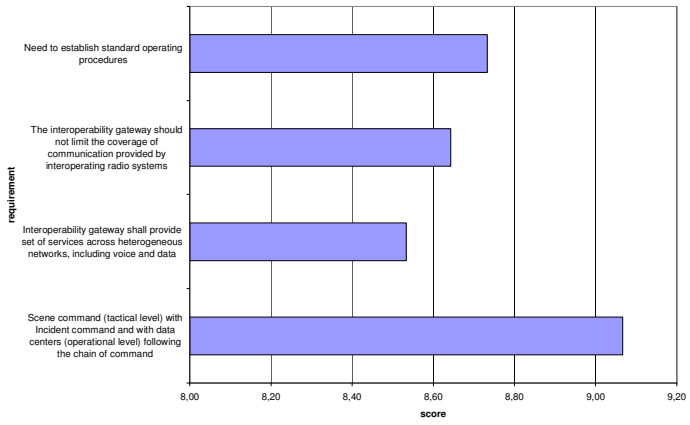


Fig. 1. Score for identified preliminary needs

Table I. Functional requirements

Domain	Functional requirements
Voice communication	Internal voice communications such as between closed user groups and end-to-end communications, etc.; External voice communications with other public safety forces or emergency services, at local level; Group communication: • ensures one to many communication to all registered in the group, regardless of the number of networks involved in the transmission, as well as the number of active users • must prevent the unauthorized access to a group • provides automatic answer of all calls within the group • prevents simultaneous voice transmission of many users within the same group • should notify the user about crossing the border of configured in the system group's activity area • ensures broadcasting capabilities to all registered users of the group • should provide a user ID presentation for all users of the group, regardless of the network used; Emergency calls to the number 112; Calls for end-user's PBX network; Call recording in dispatching communication system; Capability of assigning priorities to the individual users and type of connection; Capability of connection queuing
Data	Data access to radio channel; Data access for querying or others databases; Data transmission for decision making; Data exchange between different First-Responders; Image transmission if possible; Video transmission if possible; Metadata transmission for smart surveillance
Other	Roaming capabilities for cross-border operation if systems support this feature; Wide range of possible compatible communication technologies operating; Interoperable ad-hoc mesh network for several First-Responders operating in the same coverage area including other mesh networks deployed; Remote management of terminal by systems administrator (radio and telephone); Retransmission

VI. The Hit-Gate Architectural Concept

The Hit-Gate architecture (Fig. 2) is based on the basic concept of a Network of Networks ([15], [16]) to create a comprehensive system that supports Cross-Network Services (CNSrv) among FRNs. The Hit-Gate system is based on IP-based transmission, routing, signaling and communication. The system is compatible with any IMS type of network [17] that had been adapted in some key standards, e.g., 3GPP (3rd Generation Partnership Project), 3GPP2 and ETSI (European Telecommunications Standards Institute) TISPAN (Telecommunications and Internet converged Services and Protocols for Advanced Networking). IMS was selected to be implemented in the core of the Hit-Gate system because it performs all functions needed for call switching, user authorization and development of new services. Instead of development of an own switching and control node that manages calls and value-added services, it would be more efficient to use an existing and proved solution like IMS because it provides flexibility and ease for software developers.

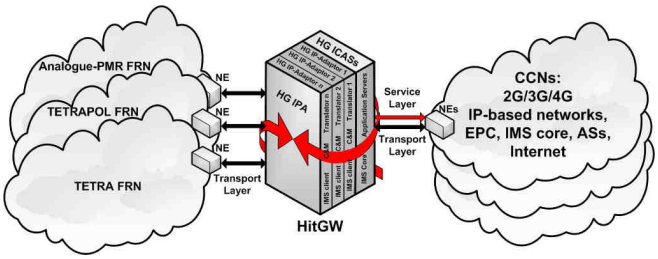


Fig. 2. The Hit-Gate architectural concept

One of the main components of the system gateway is HG IMS Core and Application Servers (HG ICASs) subsystem that is responsible for control, signaling and the CNSrv. The HG IMS Core functions are related to the generic communications services and standardized IMS interfaces, including IMS-based services. This component provides the basis for interoperability since it defines a common framework that ensures the interworking of all system nodes. The HG Application Servers (ASs) in the HG ICASs allow seamless access to CNSrv (e.g. presence function and conference connections) across whole HG NoNs even under high-dynamic conditions when additional NEs of the FRNs are attached using 'on-the-fly' and 'plug'n'play' features. The HG ICASs supports also transport network management (configuration, translation of data and protocol as well as discovery of network nodes) and transmission security.

The FRNs are connected to HG ICASs with HG IPA that consists of a set of n HG IP-adaptors ($i = 1, \ldots, n$). To have an access to a FRN over its access point, a FR Mobile Terminal (FR-MT) or Base Station (BS) and an HG IP-adaptor is required. A FR-MT is connected using manufacturer's specific interface by wire with HG IP-adaptor. FR BS can be also used instead of a FR-MT. The HG IPA can be done as one compact device that even allows connecting several incompatible technology types of mobile terminals or base stations, or it can be deployed as a distributed sub-system of n HG IP-adaptors and their attached FR-MTs or BSs. In this case the HG IPA and HG ICASs can be implemented on rugged PC tablets or laptops, respectively. An HG IP-adaptor consists of: an IMS Client representing unique user identification number; FR technology Translator that is responsible for translating the information from FR-MT into appropriate IMS messages as

well as from IMS messages into commands understood by FR-MTs; and Control and Management (C&M) component needed by an dispatcher who can handle CNSrvs among FRNs.

The "clouds" of incompatible FRNs and CCNs are interconnected with the HitGW. Thus a HG NoNs is created for CNSrv between FRs, its agencies, and other entities involved in a disaster scenario that used IP-based networks. The application components can be distributed and provided from HG ICASs or any CCN compatible with IMS, e.g. Evolved Packet Core (EPC).

The HG IPA constitutes a form of "bridge" between HG ICASs and specific radio access technologies that use manufacturer proprietary protocols which are accessible with an Application Programmable Interface (API) of the FR-MTs (Fig. 2). A HG IP-adaptor of HG IPA is used to adapt a NE of one FRN to an IMS type of network where services are controlled by HG ICASs. Due to this adaptation the IP connectivity is achieved between the FR-MTs and other IP NEs of the HG NoNs. Two types of interfaces are used to connect a HG IP-adaptor and HG ICASs:

i) SIP with extended header between HG IMS Client and HG ICASs (IMS interface); and
ii) HTTPS (Hyper Text Transfer Protocol Secured) for control and management commands (C&M interface).

The HG IP-adaptor has two additional interfaces:

i) External: to the mobile terminal; this interface has to meet radio manufacturer's specification (Translator specification).
ii) Internal: between HG IMS Client, C&M and Translator (Commands interface).

The HG IP-adaptor was designed to be compatible with several communication technologies listed in section II. It was developed comprising the common interfaces and functions used by the IMS type of network. For development of HG IP-adaptors for a FRN a software template in C++ for Visual Studio is available that can be extended to have a new HG IP-adaptor.

VII. Hit-Gate High Level Architecture

A. Development framework

In order to meet the basic requirements of creating an interoperable HG NoNs (see section IV), and all end-users requirements that are very diverse in application domains and configurable quantities, the HitGW should be based on a solution of significant flexibility. To achieve it several modern techniques have been selected to create a holistic development framework. The innovativeness of the gateway under the design makes use of the IP domain operating with SIP, IMS, and other open source solutions like On-Demand Intelligent Network Interface (ODINI) [14] that allow integrating e.g. trunking/dispatching analog/digital systems as well as data transmission ones. Such an approach accelerates a process of establishing the communication among different emergency services and other governmental organizations during crisis because the configuration procedures can be well-defined and versatile.

For the development framework of the Hit-Gate High Level Architecture (HG HLA) the Open IMS Core (www.openimscore.org) by Fraunhofer FOKUS and the application server Mobicents (www.mobicents.org) were selected. It is open source software that offers a flexibility to develop the IMS services and extensible solution and enables to make a proof of concept. The VoIP communicators were softphones OpenIC_lite for Ubuntu and X-Lite for Windows / Mac OS X. The Mobicents application server used JBoss server and SIP Servlets technology.

The HG HLA developed in Hit-Gate is shown in Fig. 3. This architecture has two layers: transport and service. The control/signaling sub-layer of the service layer consists of three Call Session Control Function (CSCF) modules and Home Subscriber Server (HSS) where subscribers' profiles are stored. The HG ASs lie on the application/service sub-layer of the service layer where the services are executed to provide them to the end-users.

The HG IP-adaptors (a distributed system solution of the HG IPA) connect access NEs of FRNs or CCNs to the HitGW, and jointly with HG ICASs they create a HG NoNs of incompatible NEs from FRNs and CCNs that are capable to exchange and use the information carried among them and thus offering the CNSrv for FRs. It ensures flexible and quick attachment of PSC systems.

Proxy CSCF (P-CSCF) component is a server that is the first point where a call is received. It also serves a function of firewall at the application level. Serving CSCF (S-CSCF) component is a central node responsible for handling the signaling. Interrogating CSCF (I-CSCF) deals with incoming SIP calls. Its address is published via DNS of the domain. It allows other servers to find this node and use it as a router for SIP packets. HSS is a main database about subscribers that supports call handling.

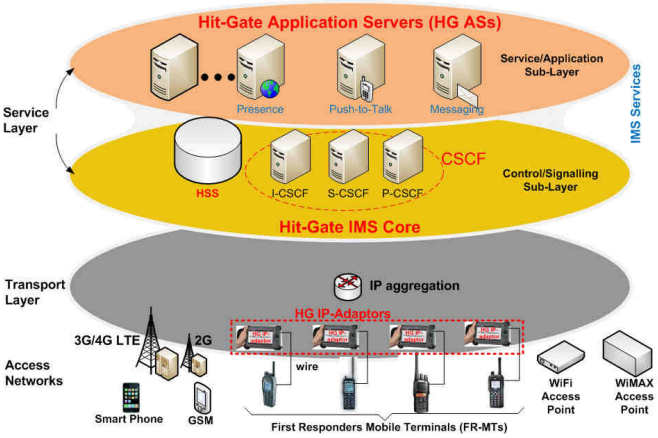


Fig. 3. The Hit-Gate High Level Architecture view

The architecture presented in Fig. 3 allows creating a wide scope of different services. FRs' services were primary targeted. The following services have been implemented: registration and authentication of NEs, presence, full-duplex and half-duplex private and group calls, emergency call, messaging, calling/talking party identification, broadcast call, retrieving a group list, group advertisement and session modification. Usage of other services (e.g., images, video streaming, IPTV) that are typical for 2G, 3G and 4G LTE type of networks can be performed as well between access NEs of FRNs and NEs of CCNs.

B. GSM Testbed

In a testbed developed we implemented the IMS platform to work with GSM (Global System for Mobile communication) network and to deliver voice communication and SMS (Short Message Service) between mobile terminals and a dispatcher. We used Nokia N900 mobile phone as a gateway terminal (access NE) to a GSM network. The analog voice signal from/to the terminal was processed in AD/DA converter and forwarded digitally with USB port to RTP driver and then to SIP softphone client. The mobile terminal was controlled using another USB port and by a software being an extension of the SIP client.

The testbed architecture is shown in Fig. 4. In such a testing environment we could set up voice calls between end-users' 2G/3G terminals and dispatcher's softphone client. Both call parties could also send SMS texts each other.

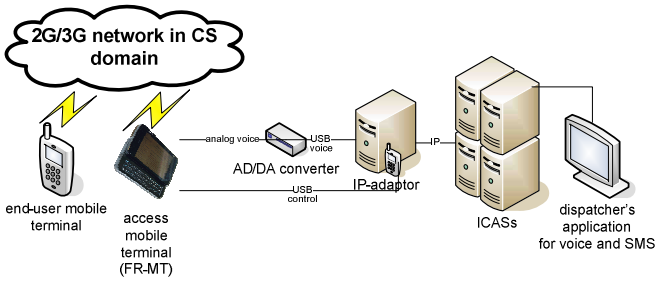


Fig. 4. Block diagram of the IMS testbed

VIII. Conclusions

In Fig. 5 one can see the coverage of end-user requirements with requirement types. The majority of requirements concerns communication aspects, and then performance ones.

Hit-Gate user requirements can be divided into five main categories of requirements:

C1 Communication services and performances requirements,
C2 Mobility and connectivity requirements,
C3 Interoperability requirements,
C4 Security and safety (i.e. confidentiality, availability, integrity) requirements,
C5 Management as well as monitoring needs requirements,

from which category C1 and C2 are described with the greatest number of requirements during the process of requirement collection.

Fig. 5. Coverage of end-user requirements with requirement types

The HitGW that offers interoperability between incompatible NEs of the HG NoNs has to meet end-user needs and expectations. Identified and defined requirements will be used for the future Hit-Gate product developments and as the guidelines for the Next Generation Networks (NGNs) and Future Mobile Convergence (FMC) as well as for ongoing standardization activities. They will be also helpful in defining overall and detailed system architecture, preparing system specification, as well as describing scenarios and use-cases, e.g., for the demonstration purposes.

The HG HLA is based on the NoNs concept. The system gateway HitGW is composed of two sub-systems: HG IPA and HG ICASs. Both sub-systems can be used remotely connected as network elements of HG NoNs or integrated as one compact system solution that interconnects incompatible NEs of the FRNs and IP-based networks. Within a HG IP-adaptor each FR technology (that is accessible via a FR-MT) should have a suitable technology translator that constitutes an interface between an access NE of the FRN and HitGW. The other HG IP-adaptor components — IMS Client and C&M — provide service as well as configuration and management solutions for FR-MT attached to it as well as for itself. Such an approach facilitates interoperability between heterogeneous systems because one just needs to develop a new translator in order to allow connectivity of a new FR-MT to its FRN.

The HG HLA was applied to elaborate a demonstrator that allowed developing a HitGW prototype as an interoperability gateway that offers FRs services and other services provided by 2G/3G/4G networks. The proposed system solution based on IMS occurred easy to install, run and configure. Development of new services to integrate other

communication systems needs Java programming and knowledge about API commands and hardware interface to control the communication via the terminal attached to the gateway. New services based on the IMS platform can extend the communication capabilities among PPDR entities offering them even ad-hoc integration and management.

REFERENCES

[1] "D2.1, Case studies and scenarios report," Hit-Gate project, 2012.

[2] "Deliverable D2.2, Analysis of crisis management system requirements, ver. 3," SECRICOM project, July 2009.

[3] "Requirements specification template," Edition 15 March 2010, James & Suzanne Robertson principals of the Atlantic Systems Guild.

[4] "Description of work of the Hit-Gate project," 2011.

[5] "Report of the workshop on 'Interoperable communications for safety and security' ", Workshop jointly organized by DG ENTR and DG JRC with the support of EUROPOL and FRONTEX, Gianmarco Baldini, 28-29 June 2010 – Ispra, Italy, EUR 24540 EN.

[6] "PPDR spectrum harmonisation in Germany, Europe and globally," WIK-Consult, Final Full Public Report, Bad Honnef, 6 December 2010.

[7] "LTE as potential communications technology for public safety operations – a usability study taking into account the aspect 'private vs. shared network' ", Baccalaurean Thesis II, Ing. Manfred BLAHA, 2011.

[8] "Meeting the challenge: the European security research agenda," European Security Research Advisory Board, September 2006 – http://ec.europa.eu/enterprise/policies/security/files/esrab–re-port–en.pdf.

[9] S. O'Neill *et al.*, "User requirements for mission-critical application – the SECRICOM case," *Technical Sciences*, No. 15(1)/2012, 2012.

[10] W. Wojciechowicz *et al.*, "Seamless communication for crisis management," *Technical Sciences*, No. 15(1)/2012, 2012.

[11] W. Wojciechowicz *et al.*, "Information and communication technology and crisis management", *Technical Sciences*, No. 15(1)/2012, 2012.

[12] ETSI TS 102 181, "Emergency communications (EMTEL): Requirements for communication between authorities/organizations during emergencies," 2008.

[13] Network Centric Operations Industry Consortium, "Findings and recommendations for mobile emergency communications interoperability (MECI)," 2007.

[14] Rohill, "Whitepaper, On-Demand Intelligent Network Interface," 2009.

[15] "D3.4, Hit-Gate High-Level Architecture and Interface Control – Final Version," Hit-Gate project, 2013.

[16] "D4.5, Generic Interoperability Framework: Interfaces and Functions – Final version," Hit-Gate project, 2014.

[17] "D4.10, Hit-Gate System (HW and SW): Design, Source Code, Binaries and Unit Test Results Iteration 2 – Final," Hit-Gate project, 2014.

[18] www.hit-gate.eu

[19] Thomas Weber, "Public Safety Future Networks – European Update," Future Network & Mobile Summit 2013, 4 July 2013

[20] ECC Recommendation (08)04, "The identification of frequency bands for the implementation of Broad Band Disaster Relief (BBDR) radio applications in the 5 GHz frequency range," October 2008

[21] ECC Recommendation (11)09, "UWB Location Tracking Systems," October 2011

[22] WGFM, "Implementation Roadmap for the Mobile Broadband applications for the Public Protection and Disaster Relief (PPDR)," Draft

[23] ECC Report 199, "User requirements and spectrum needs for the future European broadband PPDR system (Wide Area Network)," Draft

[24] ECC FM49(12)017, "FM 49 Radio Spectrum for PPDR," 20–21 March 2012

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 941–946

DOI: 10.15439/2014F437
ACSIS, Vol. 2

Throughput Improvement by Adjusting RTS Transmission Range for W-LAN Ad Hoc Network

Akihisa Matoba
Graduate School of Informatics,
Tokyo University of
Information Sciences,
4-1 Onaridai, Wakaba-ku,
Chiba, 265-8501 Japan

Masaki Hanada
Department of Informatics,
Tokyo University of
Information Sciences,
4-1 Onaridai, Wakaba-ku,
Chiba, 265-8501 Japan

Moo Wan Kim
Department of Informatics,
Tokyo University of
Information Sciences,
4-1 Onaridai, Wakaba-ku,
Chiba, 265-8501 Japan

***Abstract*—The W-LAN Ad Hoc network tends to cause problems called "Hidden Node" and "Exposed Node". RTS/CTS mechanism has been introduced to mitigate Hidden Node and most of existing researches assume that RTS and CTS are sent at the same transmission range. This paper describes a new method to improve the network throughput by adjusting the RTS transmission range. The simulation result showed that the proposed method achieved higher throughput in some degree.**

I. Introduction

THE W-LAN Ad Hoc network tends to cause problems called "Hidden Node" and "Exposed Node" [1]. Fig.1 shows an example of Hidden Node and Exposed Node. In Fig.1, transmission range and receive range are assumed to be equal. Assuming a sender node and a receiver node, a Hidden Node is located near to the receiver node and can hear the transmission from the receiver node while it cannot hear the transmission from the sender node as it is far enough from the sender node. The Hidden Node and the sender node can send a frame respectively at the same moment and can cause a collision at the receiver node. An Exposed Node is located near the sender node and its transmission reaches to the sender node but cannot reach to the receiver node as it is far enough from the receiver node. The Exposed Node doesn't cause a collision at the receiver node when it sends a frame at the same moment when the sender node sends a frame. But due to carrier sense mechanism of IEEE802.11, the Exposed Node detects a transmission of the sender node and has it suspend the transmission. This may cause unnecessary transmission suspensions and degrade network performance [2].

RTS/CTS mechanism as shown in Fig. 2 has been introduced to mitigate Hidden Node since the first version of IEEE802.11 [3].

First a sender node does carrier sense. If channel is idle, it further waits DIFS (DCF Inter Frame Space) period and random back off period. Then it sends RTS (Request To Send) and any nodes which hear the RTS reserves air time during

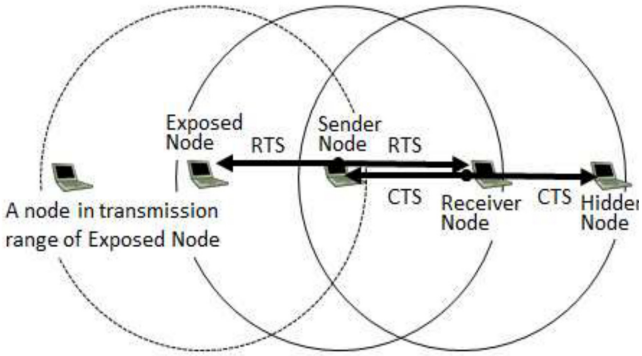


Fig.1 Example of Hidden Node and Exposed Node

NAV (Network Allocation Vector) period. The receiver node receives RTS and sends back CTS (Clear To Send) after SIFS (Short Inter frame Space) period.

CTS also have NAV and suspend transmission of nodes which hear the CTS until the receiver node sends ACK. Then the sender node sends a data frame after it receives CTS and the receiver node sends ACK after it receives the data frame. In this mechanism, Hidden Nodes around the receiver node can suspend transmissions by the CTS and the receiver node can avoid collisions to receive the data frame. But this mechanism creates Exposed Nodes around the sender node as they hear the RTS and suspend their transmission.

In this paper a new RTS/CTS method to reduce the number of Exposed Nodes has been proposed. This method assigns different transmission range to RTS and CTS respectively to reduce the number of Exposed Nodes. Simulation results shows that the proposed method improves the entire network throughput compared to the standard RTS/CTS mechanism, and also has effect to equalize variation of throughput among each node.

978-83-60810-58-3/$25.00

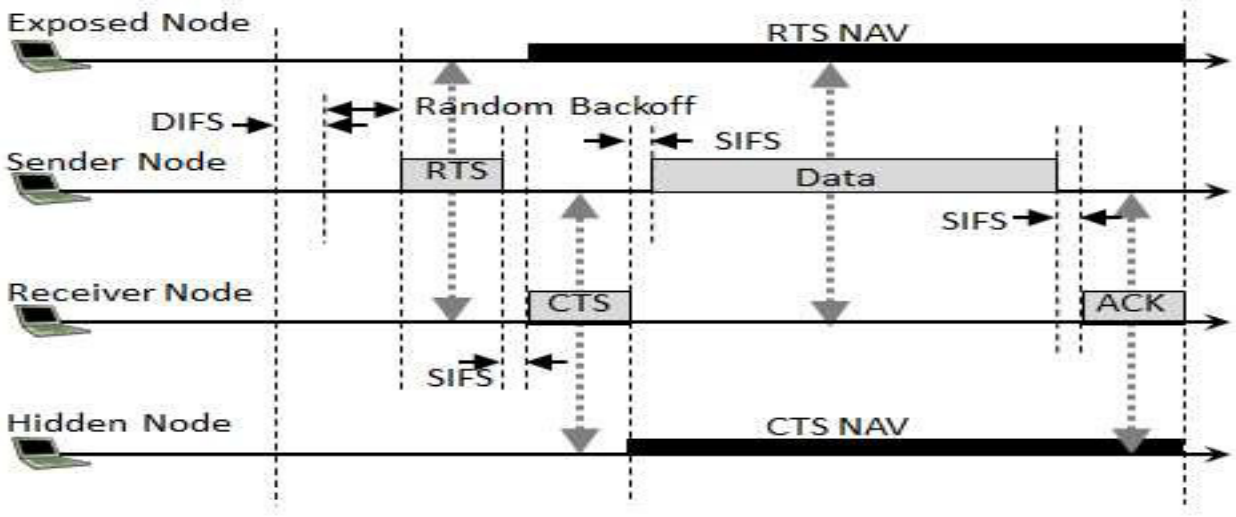


Fig. 2 Standard RTS/CTS Mechanism

II. Proposed Method

A. Related Work

Various researches have been conducted to mitigate Exposed Node and Hidden Node [4]. The most of existing researches assume that RTS, CTS and Data frame are sent at the same data transmission rate and have the same radio coverage. This was true with the first version of IEEE802.11, but now Data frame is sent at multi rates (54Mbps with 11a/g at maximum) while control frames such as RTS/CTS remain the lowest basic rate (1Mbps with 11g, 6Mbps with 11a). There are some researches to consider multi transmission rate, but no researches assume different transmission rate to RTS and CTS.

B. Basic Idea

In our research we intentionally allocate different transmission rate to RTS and CTS in order to proactively control the radio coverage and mitigate effect of Exposed Node. We have redefined the objective of RTS only to provoke CTS. RTS needs to reach to receiver node but it doesn't need to reach to any other nodes. So the RTS should be sent at the maximum data rate as data frame. This strategy introduces risk to lose CTS and ACK at the sender node by collisions from surrounding nodes (Exposed Nodes). But we assume this risk is minimal as CTS and ACK have short length in comparison to data frame.

Fig.3 shows the basic idea of our proposal. Here we define Sender Node as S, Receive Node as R, Hidden Node as H and Exposed nodes as Ei, Nj. Radius of RTS range and CTS range by standard method are defined as Rrts and Rcts. Radius of proposed higher transmission rate of RTS (i.e. RTS') is defined R'rts. As shown in Fig.3, higher transmission rate of RTS makes the RTS coverage range smaller than CTS and reduces Exposed Node. In order to avoid collision of data frame at the receiver node, CTS should be sent at lowest data rate to be heard by Hidden Node as many as possible. In case if RTS range is completely included in CTS range, there is no Exposed Node. The shaded area of Fig.3 contains the eliminated exposed nodes (i.e. Nj) by proposal.

The following steps show the procedure of the proposal;

Step1: S sends RTS' to R with possible highest transmission rate. This is to minimize the RTS' coverage range and to reduce the number of Ei.

Step2: R receives RTS' and sends back CTS with basic transmission rate. This is to ensure all potential hidden nodes to receive CTS and to suspend their transmission.

Step3: S receives CTS and sends data frame to R with maximum transmission rate. If the RTS' range is completely included in the CTS range, there is no exposed node.

Step4: R receives data frame and sends back ACK with highest transmission rate.

By the way above discussion, to reduce radio coverage is not applicable to CCA (Clear Channel Assessment). Any frames have PLCP preamble and header with 1Mbps, and the payload portion is in higher date rate (e.g. 54Mbps). The effect of CCA may need to be investigated further.

III. Simulation

A. Simulation Condition

We have assumed Wireless LAN standard of 5GHz band, IEEE802.11a. The system parameter for the simulation is shown in Table 1.

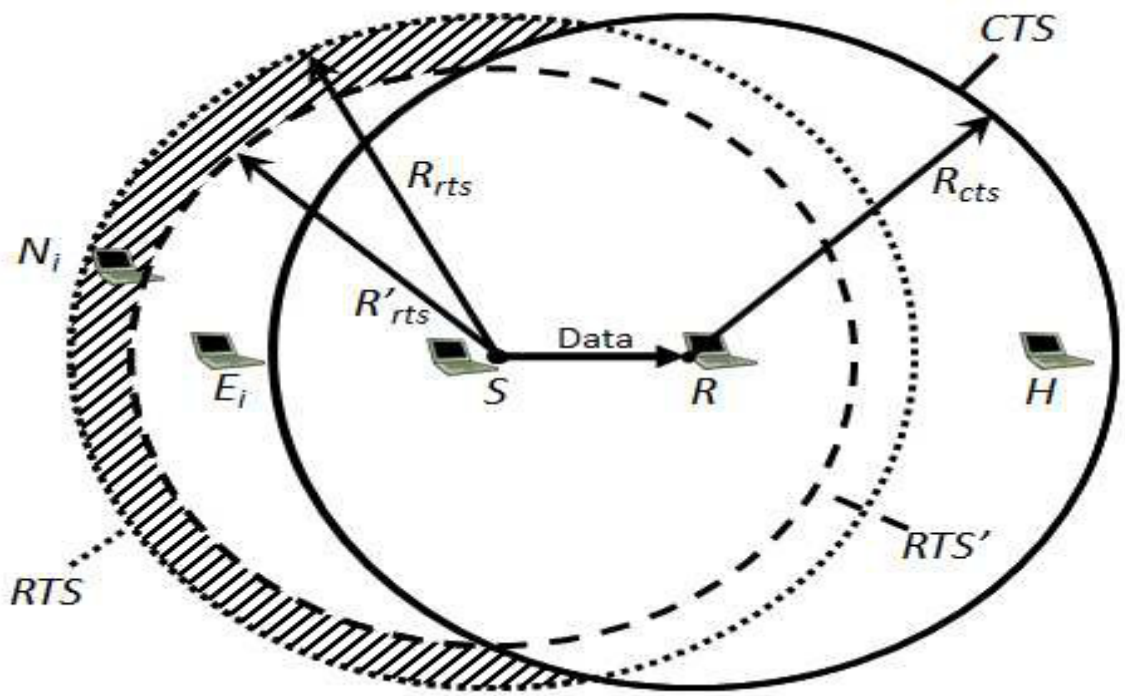


Fig. 3 Basic Idea of Proposal

In IEEE802.11a, 8 transmission rates are defined as 6Mbps, 9Mbps, 12Mbps, 18Mbps, 24Mbps, 36Mbps, 48Mbps and 54Mbps. For simplicity, in this simulation, RTS is sent at 18Mbps and CTS is sent at the minimum basic rate as 6Mbps. Data and ACK are sent at the same as RTS (i.e. 18Mbps).

As the simulated network topology all nodes are located in a grid with 70m interval. Seven cases are assumed for grid size from 3x3 with 9 nodes to 15x15 with 255 nodes. Nodes can be randomly distributed, but in practical deployment distribution of nodes is often governed by artificial objects such as walls, furniture, partitions and structures of building and it follows geometric arrangement. So we assumed the grid distribution as the initial research stage.

The 5x5 grid of nodes is shown in Fig.4. The node 13 is the sender node and receiver node is selected among node 8, 12, 14 and 18 at random. In Fig.4, node 14 is selected as the receiver node.

TABLE 1.

SYSTEM PARAMETERS FOR SIMULATION

Frame	Type	Transmission Rate	Range	
	RTS	18Mbps	88m	1 hop (70m)
	CTS	6Mbps	140m	2 hops (140m)
	Data	18Mbps	88m	1 hop (70m)
	ACK	18Mbps	88m	1 hop (70m)
Load	3Mbps per node with exponential distribution			
Data Size	1,000 bytes			
Distance	Nodes are located at 70m interval in a grid.			
Other	DIFS=34μs, SIFS=16μs and Slot time=9μs. Other parameters follow 802.11a standard.			

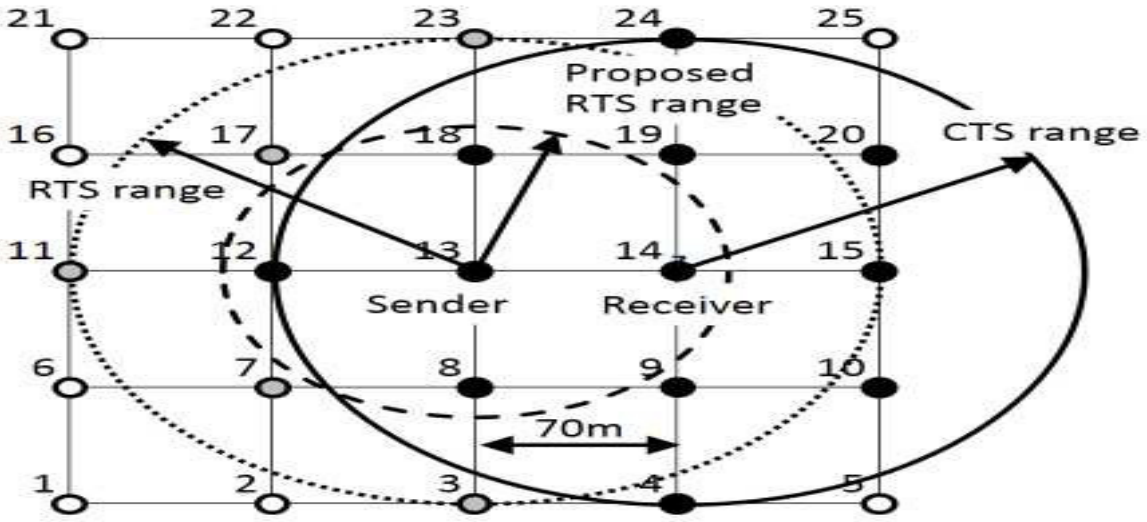


Fig. 4 25 Nodes (5 x 5) Example

An RTS with standard method reaches up to a node at two hop distance and totally 12 nodes excluding sender node are in transmission range. An RTS with proposed method reaches only to nodes at one hop distance and totally 4 nodes are in transmission range. As the CTS transmission range is two hops, RTS range of the proposed method is completely included in CTS range.

In Fig.4, black nodes are in the CTS transmission range and white nodes have no influences with the transmission from note 13 to node 14. Gray nodes would be Exposed Nodes if standard method is applied. As you see in Fig.4, in case of standard method with 5x5 grid, gray nodes (i.e. Exposed Nodes) are very often located at the boundary of the network. It is anticipated that boundary condition would affect throughput improvement ratio especially to small size grid. Therefore we have simulated up to 15x15 grid of 255 nodes.

B. Simulation Result

As shown in Fig.5, throughput per node goes lower as the size of grid goes bigger for both standard and proposed method. The simulation result showed that the proposed method achieved higher throughput per node in all grid sizes.

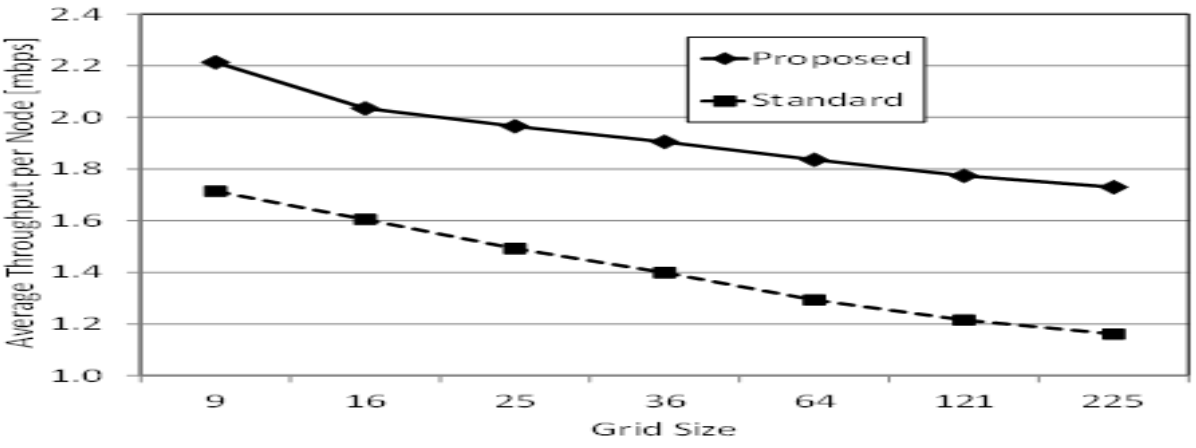


Fig.5 Average Throughput per Node

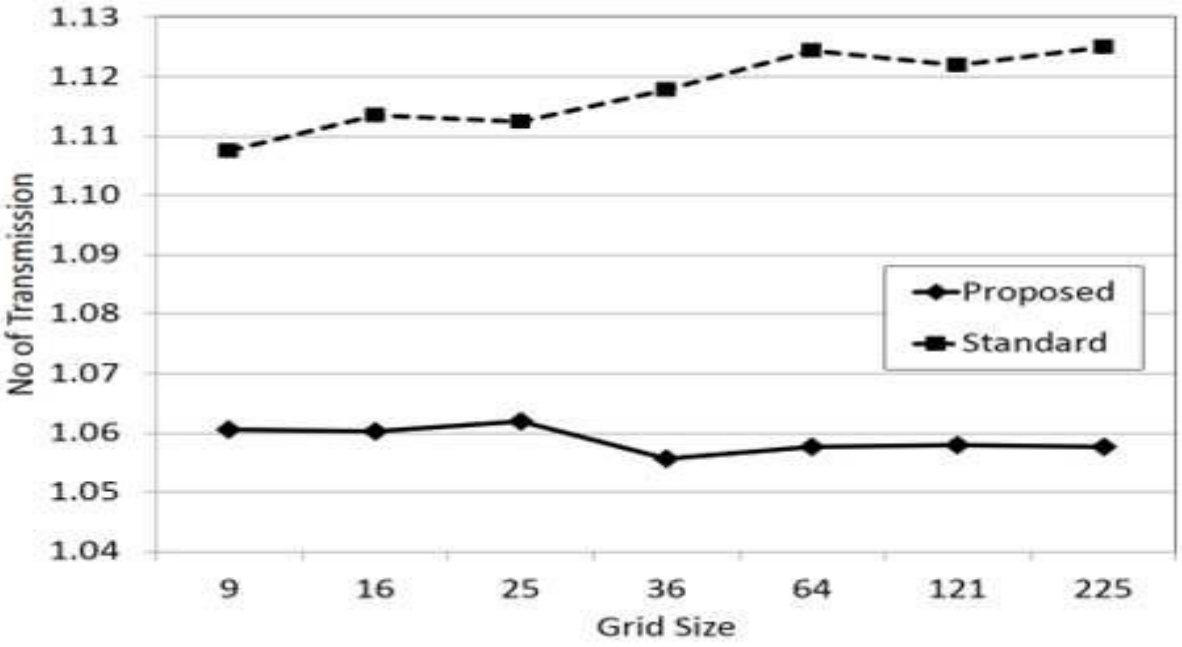


Fig.6 Average Number of RTS Transmission

Fig.6 shows average number of RTS transmission per data frame for each grid size. The number greater than 1.0 imply the occurrence of RTS retransmission. With standard method, 11% to 13% of RTS were retransmitted due to collisions. With the proposed method, only 5% to 6% of RTS were retransmitted.

Fig. 7 shows the average throughput dispersion. The proposed method has smaller dispersion than standard method, and this tendency is more ostensible with smaller grid size.

IV. Conclusion

In this paper we have proposed the new method to adjust the transmitting rate of RTS to the same as data frame in order to control its transmission range proactively.

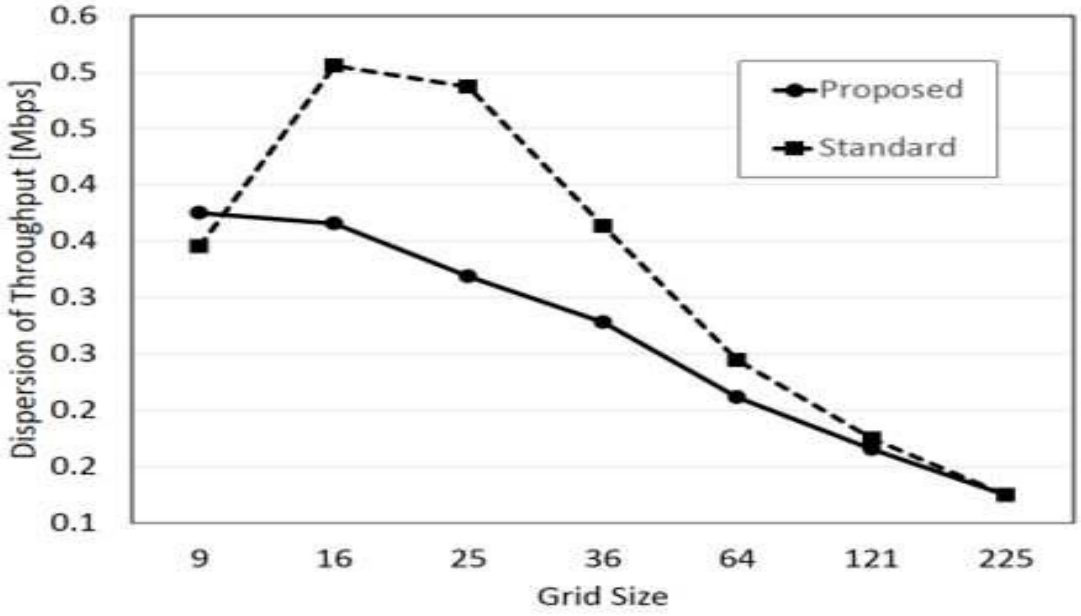


Fig.7 Dispersion of Throughput

We have showed by simulation that the proposed method can improve throughput. We need to investigate further to validate effect of proposed method and to find method of selecting appropriate parameters as well as theoretical explanation.

References

[1] K. Xu, M. Gerla, and S. Bae, "How effective is the IEEE802.11 RTS/CTS Handshake in Ad Hoc Network?" Proc. IEEE Globe Com'02, pp.72-76,Nov. 2002.

[2] S. Ray, J. B. Carruthers. and D. Starobinski, "RTS/CTS-Induced Congestion in Ad Hoc Wireless LANs," WCNC 2003 IEEE, vol. 3, pp.1516-1521, March 2003

[3] IEEE-SA Standards board, "Part11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," IEEE Std 802.11-2012

[4] K. Nishide, H. Kubo, and R. Shinkuma, "Detecting Hidden and Exposed Terminal Problems in Densely Deployed Wireless Networks," IEEE Transactions on Wireless Communications, vol. 11, no. 11, pp.3841-3849, November 2011

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 947–954
DOI: 10.15439/2014F159
ACSIS, Vol. 2

The procedure for monitoring and maintaining a network of distributed resources

Tomasz Malinowski
Military University of Technology,
Faculty of Cybernetics
ul. Kaliskiego 2, 00-902 Warsaw,
Poland,
Email: tmalin@ita.wat.edu.pl

Artur Arciuch
Military University of Technology,
Faculty of Cybernetics
ul. Kaliskiego 2, 00-902 Warsaw,
Poland,
Email: a.arciuch@ita.wat.edu.pl

Abstract—**In this article, we propose an algorithm for the management of the structure of the tunnel connections in large, distributed nature, corporate network. Management is here understood as monitoring the availability and health of key elements and dynamic reconfiguring the system in situations of failure symptoms. It was assumed that the algorithm can be used in cases where the correct functioning of the network (providing the appropriate level of quality of service) means an access to distributed resources, usually redundant, which may be subject to reallocation. It was shown that developed algorithm may be used for maintenance of VPN network with dynamic tunnels (DMVPN).**

I. Introduction

In present-day's communication networks, which are networks of different services integration and networks giving access to a distributed resources, the efforts of the engineers, protocols and network hardware designers, and in particular, the information and communication systems administrators, focus on ensuring high-reliability of the system, continuous availability of network resources and the desired level of quality of services. Efficient and effective diagnosis of inappropriate states (incompatible with expected) of system functioning is unquestionably the most important challenge for administrators of complex ICT environments.

The algorithm used to oversee the operation and dynamic reconfiguration of the system of dynamic tunnels, raised between border routers of company's branches networks was proposed. It's not too hard to indicate many examples of systems, in which the continuous and reliable access to distributed network resources (for example resources within the cloud computing system, within distributed data centres, etc.) seems to be critical (especially important) and for which different optimizing algorithms of structure of connections and reallocating resources of servers and network nodes, with the requirement of achieving state of convergence as soon as possible are proposed [8], [9], [10].

The proposed algorithm for system of dynamic tunnels is based on some elements of system-level diagnostics and self-diagnosable systems [1] - [7], [13], [14].

Corporate network of dynamic tunnels (DMVPN - Dynamic Multipoint Virtual Private Network) [11] is seen by authors as a client-server system in which a significant problem is to provide reliable client-server communication, where server plays role of the tunnel broker, and reliable client-client communication through dynamically created tunnels.

The DMVPN network, from the point of view of the authors, contains certain number of clients, DMVP servers and, introduced by authors, the management station. Transmission links (tunnels) between clients and servers form a logical structure of connection links and the primary function of tunnel brokers (servers) is to provide communication between network clients by informing clients about tunnel parameters. Tunnels between clients are rise dynamically on transmission time only, so the logical structure is subject to continuous modification.

It is assumed that the client can forward the query about parameters of tunnel leading to another client to one of the many assigned (known) servers (it's system with tunnel broker redundancy).There are the primary server and backup servers in the group of servers. In the absence of the ability to provide client-server communication, or in case of communication parameters deterioration, the client should have the possibility to appeal (to send service order) to the other server from a server group. In addition, the client should send inaccessibility notification to the management stations (suspicion of primary or/and backup server failure notification). Quick server failure detection will be possible through periodic availability testing of the servers, that are assigned to the client.

Management station should have the possibility of reconfiguring logical network connections, which means the ability to assign clients to servers (primary and backup).It is assumed that the servers will have the opportunity to test the quality of the connections to clients (source to destination latency, packet loss, jitter, etc.), and the result of testing will have an impact on the allocation of clients to the server.

Quickly responding to the unavailability of tunnel brokers (detection of unfit brokers) will be carried out using IP SLA probes [12], available within operating systems of network devices, automating the login process (telnet/ssh) from the

978-83-60810-58-3/$25.00

management station to network nodes (clients and servers) and changing their configuration.

Authors propose the procedure of dynamic tunnels network management based on linking the utility functions of network devices with diagnostic functions. Methods of connecting utility functions with diagnostic functions belong to the system diagnostic methods (system-level diagnosis), and systems capable to indicate unsuitable items and reconfiguration (auto-repair) belong to class of self-diagnosable systems.

II. Characteristics and Structures of DMVPN Network

DMVPN is a technology that allows to build scalable VPN, combining the advantages of GRE (Multipoint GRE), IPSec and NHRP (Next Hop Resolution Protocol). DMVPN connects branch offices through the Internet (WAN) infrastructure based on tunneling and assuming that the tunnels will be "lifted" dynamically, as needed, and so will not be required to create persistent connections and, worse still, arranging them in a structure of type "full-mesh". The technology is based on a logical topology of a star, with the highlighted nodes called Hub and Spoke. In the basic configuration, DMVPN offers communication between branches through Central Branch (Spoke-Hub-Spoke connections) and in the enlarged configuration gives the ability to directly connect branch office networks with Spoke routers (Spoke-to-Spoke connections).

The general shape of the corporate network with dynamic tunnels was shown in figure 2.

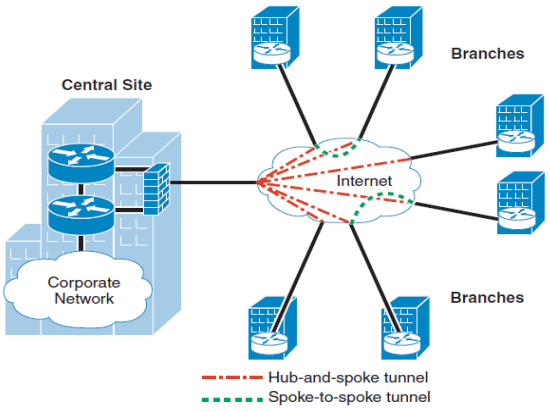


Fig. 2 Overall shape of DMVPN network[11]

Implementations of Dynamic Multipoint VPN networks

A particularly important feature of DMVPN is the ability to dynamic "lifting" tunnels, which encapsulate packets of any type (unicast, multicast, broadcast), IPv4, IPv6, and others. It follows directly from the application of GRE (Generic Routing Encapsulation) protocol. So, for example, DMVP can be seen as one of the transitional mechanisms for connecting IPv6-capable networks through IPv4-based network infrastructure. Importantly, packets transmission in dynamic GRE tunnels can be secured by IPSec, and one tunnel interface supports multiple IPSec sessions. It is also worth noting that due to the possibility of organizing tunnels between branch offices, although a little bit difficult, configuration of QoS policies, for example in system with VoIP calls, becomes more transparent.

DMVPN network in basic configuration was presented in figure 3.

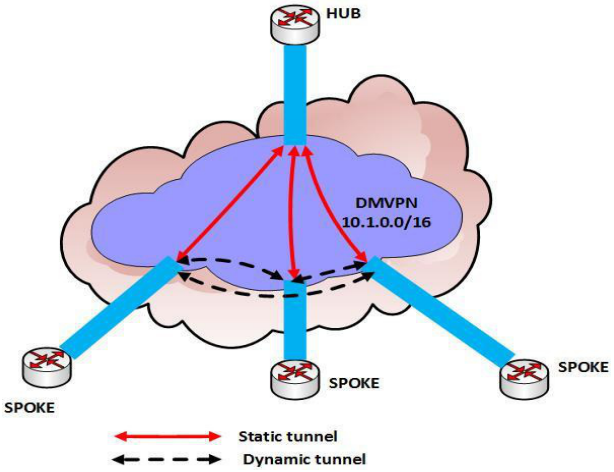


Fig. 3 DMVP Network with one Hub router

The router called Hub operates within the Central Branch network, while in remote branches act Spoke routers (Hub's clients). Remote branches have a persistent tunnel connection to Central Branch and access to Central Branch's network resources (static tunnel between a Spoke-Hub pair). In the case of realizing transmission between remote branch offices, transmission is preceded by sending an NHRP (Next Hop Resolution Protocol) initiator Spoke's request for tunnel address of Spoke router in a distant location. The Hub in a central location answers to NHRP query and is called NHS Server (Next-Hop Server), and NHRP protocol allows to get information about actual addresses of Spoke's interfaces. Tunnel connection between the branch offices can be implemented via the Hub, or it can be made a direct tunnel between Spokes. The main requirement of tunnel creation between the branches (direct or via the Hub) is Spoke's registration in NHS server and cyclic refreshing of registration. Hub's and Spoke's configuration details are discussed in the technical documentation [11] and will not be discussed here.

In the extended configuration (as in figures 4 and 5), to increase the degree of reliability of the network, additional Hub nodes are implemented or another DMVPN network is created.

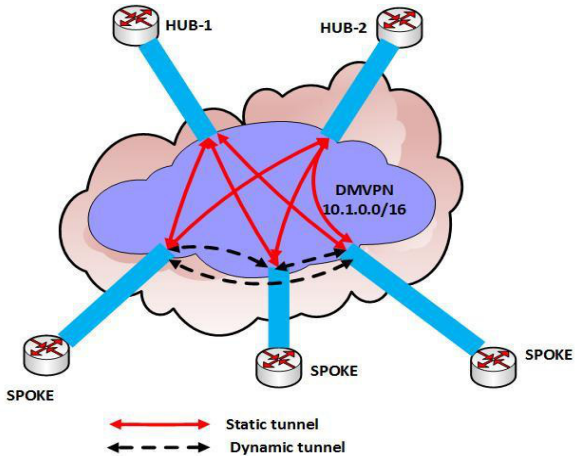


Fig. 4 DMVPN network with backup NHS server

In the case of single DMVPN cloud (figure 4), each Spoke uses a single mGRE tunnel leading to two or more Hubs, acting as the NHS servers.

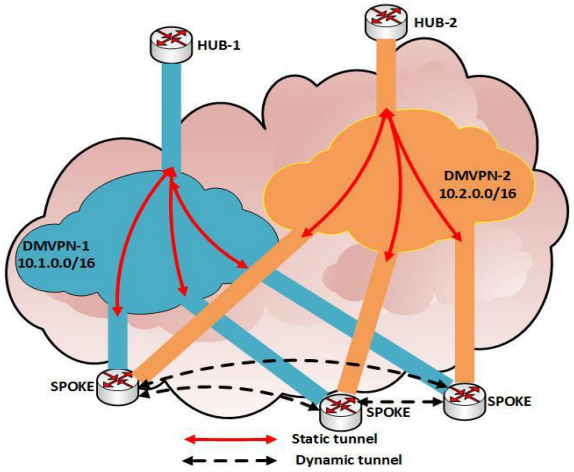


Fig. 5 Dual-cloud DMVPN

In dual-cloud DMVPN network (figure 5), each Spoke uses two tunnels, leading to two different Hubs.

Sample problems in DMVPN networks

As mentioned, the basic condition of successful tunneling between the different locations is proper functioning of the Spoke in Hub signing up mechanism, the polling mechanism about physical addresses of the final points of tunnels, but also disconnecting calls in case of failure (for example in case of demise Hub's network interface, its temporary loss, too much load of the Hub, too long a response time, etc.).

In some cases, such as realization of tunnel connections protected by IPSec, it is necessary to continuous monitoring Hub's availability and removing IPSec session after a period of temporary unavailability of the Hub (cleaning of IPSec Security Associations). In case of unavailability or even reduce the effectiveness of primary Hub, handling the entire NHRP process should be taken over by another (secondary) Hub.

DMVPN technology is refined through the years, but it's still possible to indicate some scenarios in which an administrator's intervention is necessary.

An illustration of problems appearing in the network DMVPN (besides the obvious problem of the physical unavailability of critical nodes) may be the following example with IPSec tunnels. In a simulation environment, DMVPN network with structure shown in Figure 6, has been configured.

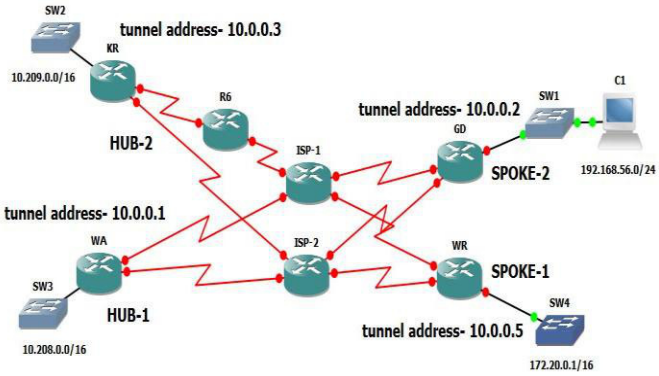


Fig. 6 Simulation environment of DMVPN network

In accordance with the assumption of DMVPN network, SPOKE-1 and SPOKE-2 maintain IPSec tunnels with theirs primary HUB-1. The sample configuration of Spoke's tunnel is as follows (for SPOKE-1):

```
interface Tunnel0
 bandwidth 1024
 ip address 10.0.0.5 255.255.255.0
 no ip redirects
 ip mtu 1400
 ip nhrp authentication dmvpn
 ip nhrp map 10.0.0.1 1.0.0.1
 ip nhrp map multicast 1.0.0.1
 ip nhrp network-id 99
 ip nhrp holdtime 300
 ip nhrp nhs 10.0.0.1
 ip tcp adjust-mss 1360
 delay 1000
 tunnel source Serial0/0
 tunnel mode gre multipoint
 tunnel key 232323
 tunnel protection ipsec profile dmvpn
```

Active IPSec connection can be listed after the commands: show dmvpn and show crypto session, as below (for SPOKE-1).

```
WR#show dmvpn
Legend: Attrb --> S - Static, D - Dynamic, I - Incompletea
      N - NATed, L - Local, X - No Socket
      # Ent --> Number of NHRP entries with same NBMA peer

Tunnel0, Type:Spoke, NHRP Peers:1,
 # Ent  Peer NBMA Addr Peer Tunnel Add State  UpDn Tm Attrb
 ----- --------------- --------------- ----- -------- -----
    1         1.0.0.1         10.0.0.1    UP 00:01:54 S
```

```
WR#show crypto session
Crypto session current status

Interface: Tunnel0
Session status: UP-ACTIVE
Peer: 1.0.0.1 port 500
  IKE SA: local 3.0.0.1/500 remote 1.0.0.1/500 Active
  IPSEC FLOW: permit 47 host 3.0.0.1 host 1.0.0.1
   Active SAs: 2, origin: crypto map
```

It's easy to demonstrate that a temporary loss of connection between Spoke and Hub, causes transmission rupture between two Spokes.

The following sequence of events is an illustration of this problem:

```
WR#ping 192.168.56.2 repeat 10000000

Type escape sequence to abort.
Sending 10000000, 100-byte ICMP Echos to 192.168.56.2, timeout is 2 seconds:
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
!!!!!!!!!.....
```

In this moment, the temporary collapse of the HUB-*1's interface* occurs

```
*Mar 1 00:07:23.679: %DUAL-5-NBRCHANGE: IP-EIGRP(0) 100: Neighbor 10.0.0.1 (Tunnel0) is down: holding time expired.....................................................
.....................................................................
Success rate is 37 percent (79/209), round-trip min/avg/max = 156/365/844 ms
```

When we observe messages on HUB-1's console, just after lifting the interface, we see syslog messages, announcing incorrect SPI IPSec session ID.

```
WA#
*Mar 1 00:08:31.703: %CRYPTO-4-RECVD_PKT_INV_SPI: decaps: rec'd IPSEC packet has invalid spi for destaddr=1.0.0.1, prot=50, spi=0x741CDD33(1948048691), srcaddr =3.0.0.1

*Mar 1 00:09:32.139: %CRYPTO-4-RECVD_PKT_INV_SPI: decaps: rec'd IPSEC packet has invalid spi for destaddr=1.0.0.1, prot=50, spi=0x1085029(17322025), srcaddr=2.0.0.1
```

Unfortunately, the remedy in this case is "manual" breaking IPSec session, as illustrated below.

```
WR#clear crypto session
*Mar 1 00:11:59.223: %DUAL-5-NBRCHANGE: IP-EIGRP(0) 100: Neighbor 10.0.0.1 (Tunnel0) is up: new adjacency

WR#ping 192.168.56.2 repeat 10000

Type escape sequence to abort.
Sending 10000, 100-byte ICMP Echos to 192.168.56.2, timeout is 2 seconds:
!!!!!!!!!!!!!!!!!!.
Success rate is 94 percent (18/19), round-trip min/avg/max = 224/411/624 ms
```

The case discussed above helped the authors to test the procedure of monitoring and reconfiguration of the nodes (routers, Hubs and Spokes) and it should be noted that in multiple IPSec implementation, including Cisco device's implementation, a mechanism for detecting unreachability of nodes communicating by cryptographic tunnels and resolving the problem of "dead" IPSec session, known as the Dead Peer Detection, was introduced.

Beside the mentioned case, problems caused by the failure of nodes and network links were identified. Scenarios in which reconfiguring the nodes was necessary (because of the failure of the primary route leading to the HUB) were considered. The necessary reconfiguration included changes of static routing paths in the routing table and ensuring availability of one of two NHS servers by the Spoke node.

III. General Model of Self-Diagnosable Client-Server System

The procedure for oversee a network of dynamic tunnels, presented in chapter IV, allows to treat a managed DMVPN network as a client-server and a self-diagnosable system ([3], [4], [7]).

The self-diagnosable system determines the structure of mutual testing of elements, the method of using tests results, and the model of inference, based on the results of tests, about the reliability state of the system [3]. The structure of self-diagnosable system is described using a testing graph. Depending on the method for interpreting the results of tests, distributed systems and centralized systems can be distinguished. Inference about the state of the distributed system take place on the basis of results of parts of all the results of tests. Inference about the state of the centralized system take place on the basis of results of all the results of tests. In addition, stands out heterogeneous and homogeneous systems. Due to the reasoning model, which defines the relationship between results of tests and the reliability state of system, in centralized systems stands out PMC model [4] and BGM model [5], in distributed systems stands out MM and MM* models [1],[2].

In the self-diagnosable systems inference about the state of the system is implemented under certain conditions and on the basis of the results of tests obtained by fault-free and faulty elements. One of the specific conditions necessary for the reasoning about the reliability state of the system is the requirement for the maximum number of faulty elements of the system (called diagnosability) within a given number of all elements of the system. The diagnosability of the system is defined as the maximum number t, such that the system is t-diagnosable as long as the number of the faulty elements is not greater than t.

If a testing graph of a self-diagnosable system is a such edge induced subgraph of the system, which describes the t-diagnosable system, has minimal number of tests, then is called t-optimal testing graph of the t-diagnosable system. The t-diagnosable system has an irreducible testing graph if none of its edge induced subgraph does not describe a testing graph of t-diagnosable system. An irreducible testing graph that is not t-optimal is a t-quasi-optimal testing graph.

Sometimes, the system design must take into account that the costs of mutual testing of elements are not the same. If the arcs of the testing graph have assigned a generalized cost of testing, then such a graph can be called economic testing graph. The edge induced subgraph, that describes t-diagnosable system for which the generalized cost of testing shall give the minimum value, describes the cheapest t-diagnosable system. The t-diagnosable system, in the general case, may has several (many) t-optimal testing graphs. The cheapest testing graph is generally one of the t-optimal or t-quasi-optimal testing graphs.

General model of self-diagnosable client-server system

Assume that an organization has a computer network O, $O = \langle V, E\rangle$, of a certain logical structure, in which V denotes a set of computers ($v \in V$), E – a set of communication links ($v', v'') \in E \wedge v'' \in V(v') \wedge v' \in V(v'')$, where V(v) denotes set of nodes adjacent to node v. In the network O stands out computers of client type $k \in K$, computers of server type $s \in S$ and a computer of manager type z. It is assumed, for simplicity, that the manager is reliable. Also $(V = K \cup S \cup \{z\}) \wedge ((K \cup S) \cap \{z\} = \emptyset)$. A client k has assigned an ordered pair of servers $k(S) = (s', s'') \in S$, where s' is a primary server and s'' is a backup server. A client communicates with the server to invoke a given service on a server. A client also sends diagnostic messages (so-called traps) to the manager z. Similarly, each server sends traps to the manager z. The manager z stores information about logical structure of network O and about the status of clients and servers. Each server s stores information about adjacent clients: $s(K) = \{k \mid k \in S(s)\}$.

Set S of servers is a such subgraph of O, $\langle S\rangle_V$, that it is a testing graph of the PMC model.

It is known that if the system with |V| nodes is t-diagnosable for the PMC model, then [4]:

$$(|V| \geq 2t + 1) \wedge (\forall(v \in V) \mid \mu^-(v) \geq t), \qquad (1)$$

where $\mu^-(v)$ is indegree of node v.

A system with |V| nodes that satisfies the formula (1) is t-diagnosable if and only if [7]:

$$((\forall(0 \leq p \leq t-1) \wedge \forall(V' \subset V) \mid |V'| = |V| - 2t + p)) \Rightarrow |\Gamma(V')| > p, \qquad (2)$$

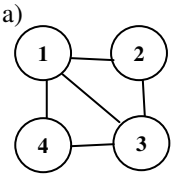
where $\Gamma(V') = \{v \mid \exists(v' \in V')\ v \in V'(v') \wedge v \notin V'$ means set of successors of elements of a set V', and $v \notin V'$.

Assume that subgraph $\langle S\rangle_V$ is described by a testing graph for 1-diagnosable system for the PMC model - it is assumed that the probability of damage of more than one server at the same time is low. From the formula (1) follows that $|S| \geq 3$. It is known that strongly connected graph, which has at least 3 nodes, is a testing graph for 1-diagnosable system of the PMC type.

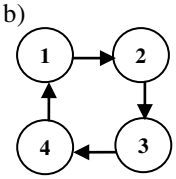
The figure 1 shows an example of a testing graph (1b) for subnet of servers, which has a logical structure like in figure 1a. A such pattern of test d_{st} where server s is testing server t was shown in figure 1c. Symbols: n(e), $n_0(e)$ and N = [n(4), n(3), n(2), n(1)], n(i) = x, i = 1,...,4, $x \in \{0,1\}$ denote: a reliability state of node e, state where node e is fault-free and vector that describes a reliability states of system, wherein, if n(i) = 0, then the i-th node is fault-free. If N = [0001], then result of test [d_{12}, d_{23}, d_{34}, d_{41}] can be [0001] else [1001], which corresponds to the pattern [x001].

In PMC model all tests are performed between two adjacent nodes, and it was assumed that a test result is reliable (respectively, unreliable) if the node that initiates the test is fault-free (respectively, faulty).

a)



b)



c)

$$[n(e_s) = n_0(e_s)] \rightarrow \left[d_{st} = \begin{cases} 0 \text{ dla } n(e_t) = n_0(e_t) \\ 1 \text{ dla } n(e_t) \neq n_0(e_t) \end{cases} \right]$$

$$[n(e_s) \neq n_0(e_s)] \rightarrow [d_{st} = x],\ x \in \{0,1\},$$

d)

N \ d_{st}	⟨1,2⟩	⟨2,3⟩	⟨3,4⟩	⟨4,1⟩
[0000]	0	0	0	0
[0001]	x	0	0	1
[0010]	1	x	0	0
[0100]	0	1	x	0
[1000]	0	0	1	0

Fig. 1 The PMC model illustration

IV. ALGORITHM FOR MONITORING AND MAINTAINING DMVPN NETWORKS

This chapter presents a general algorithm for oversee client-server networks, applied in the construction of self-diagnosable DMVPN network.

The fundamental objectives of proposed algorithm are as follows:

a) ensuring high reliability of the system on the basis of the continuous servers availability. We assume that the client should be able to communicate all the time with one/two servers. At the entrance, each client has assigned two servers, from a pool of servers (initial pool consists at least 3 servers to provide 1-diagnosability using the PMC model). In case of failure (unavailability) of one of the server, reconfiguration of the client is needed. Reconfiguration involves interfering in the configuration file and changes the settings of tunnel interfaces, which should indicate a new set of servers. So, also reconfiguration of the set of servers is needed.

b) shortening the convergence time of the system, which can be, in the opinion of the authors, achieved (in dynamic routing based and timers controlled DMVPN network) by forcing fast reconfiguration of nodes (servers and clients).

c) servers load balancing, which means that servers support nearly equal (comparable) number of clients.

We assume that the DMVPN network of an organization has a logical structure (figure 7) described by a graph O = ⟨V, E⟩ | (V = K ∪ S ∪ {z}) ∧ ((K ∪ S) ∩ {z} = ∅), where S denotes a set of servers (Hubs), K denotes a set of clients (Spokes) and z is reliable manager of the network. Servers s ∈ S form a testing graph for the PMC 1-diagnosable system, The testing graph has a logical structure which is the Hamiltonian cycle, which has a directed Hamiltonian cycle. As mentioned earlier, a client k communicates with its servers k(S) = (*s'*, *s''*), *s'*, *s''* ∈ S, where *s'* is a primary server and *s''* is a backup server. A client maintains constant tunnel with its servers and asks servers (primary at first turn) about tunnel parameters leading to other clients. Each client has the ability to send diagnostic messages about fault detection (inability to communicate with server) to the manager z. Similarly, each server can inform about problems within the set of servers by sending traps to the manager z. The manager z knows logical structure of network O and the role played by each network node (client or server). Each server s has a specific set of assigned clients s(K) = {k | k ∈ S(s)}.

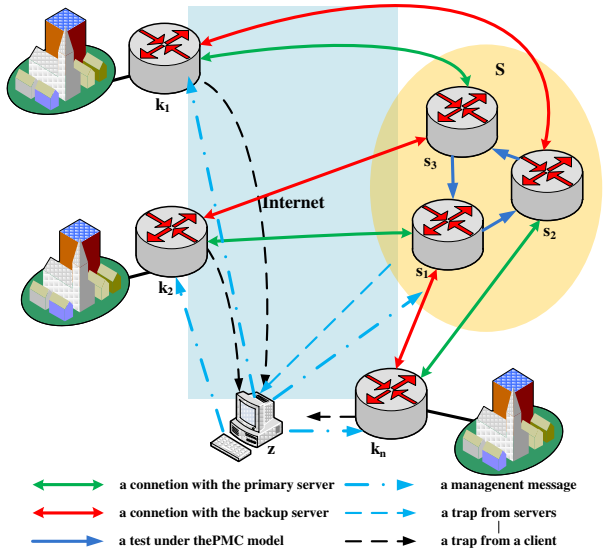


Fig. 7 Logical structure of connections and relations within DMVPN network

We assume that DMVPN system is running and has initial configuration, which means that each server supports some set of clients and the set of servers contains at least 3 servers. The set of potential servers (subset of all nodes) is also known.

The algorithm for monitoring and maintaining consists of the following steps:

1) The client k to perform a task (establish communication with the other client/server), invokes a service on a primary server s'. If the server s' can not do the job then:
   a) the client k invokes a service on a backup server *s''*,
   b) the client k generates the trap informing about server *s'* unavailability, which is sent to the management station (manager) z.
2) The manager z checks whether other traps, concerning server *s'*, from others clients k '| k' ≠ k were received, which would be a confirmation of the server's crash:
   a) if not, it is assumed that a failure of connections only between s' and k is occurred,
   b) otherwise, the manager z modifies a set of servers S, by removing the server s' and adding to this set a new server k * (from the pool of other network nodes) and by promoting it to the role of the server *s'*.
3) The manager z instructs the servers of S to execute functional tasks. Functional tasks are:
   a) the server s' instructs servers S(s') to perform the task of checking the connection parameters S(s')(K), where (S(s')(K) denotes set of clients adjacent to servers of S(*s'*)),
   b) servers S(*s'*), after checking of connections parameters with clients S(*s'*)(K), send these parameters to s'; sending responses to s' confirms the absence of failure of servers S(s'),
   c) if there is such s ∈ S(*s'*), which did not respond, the server s' sends to the manager z a trap, which inform about the not-responding server; a trap service is performed like in step 2b),
4) The manager z instructs servers s ∈ S to send back testing results, which are parameters values of connections between specific clients and servers.
5) After receipt of the test results from servers, the manager z allocate to clients k pairs of servers k(S). The method of allocation should result in an equal servers load and assuming that servers should support customers with the best connection parameter values.

In a DMVPN network, in relation to the general model of client-server, Hubs play a role of servers, Spokes play the role of clients and a manager manages a reconfiguration of Hubs and Spokes.

V. TEST ENVIRONMENT AND THE TECHNICAL FACILITIES USED IN THE COURSE OF EVALUATING THE FEASIBILITY OF THE ALGORITHM

Evaluating the feasibility of the algorithm was done in testing environment shown in Figure 8.

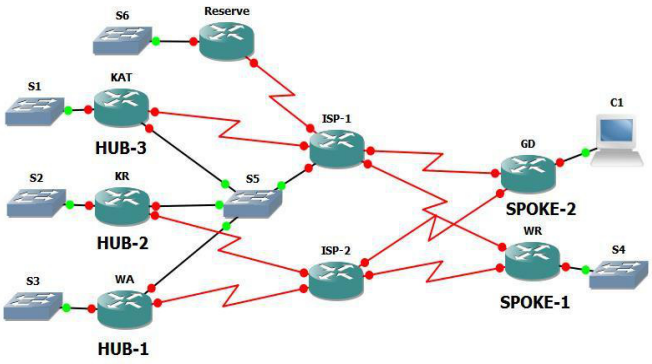


Fig. 8 Test environment of self-diagnosable DMVPN network

The set of potential Hub routers includes HUB-1, HUB-2, HUB-3 and Reserve. Routers called SPOKE-1 and SPOKE-2 register in the primary and backup server NHS (HUB-1 and HUB-2 routers). HUB-3 router is dynamically configured NHS server when the primary and/or backup server is unavailable.

Hub called Reserve was dynamically included to the set of hubs in case of failure of one of the hubs, which causes that testing in accordance with the PMC model assumptions was possible.

Simulated scenarios with symptoms of failure were as follows:

- loss of one NHS servers, caused by the failure of the primary link (leading to a primary NHS server) or simply by switching off NHS server,
- momentary loss of the primary server within IPSec connections.

Test environment has been built on the basis of Cisco equipment. Thus, offered by Cisco IOS system technical facilities for testing the quality of connections between nodes, with the possibility of sending a trap messages, generated by the device in the case of the unavailability of another tested device, were used. The logging from management station (C1 in Figure 8) on devices which require reconfiguration and the necessary changes in the configuration file was also used. At the moment, the procedure is not yet fully automated. Full automation of our system still requires writing shell scripts, woven into the configuration file of network devices (in the case of the Cisco, Tcl scripts).

The principal mechanisms of detecting the unavailability of network nodes and the state of the route's transmission parameters deterioration were IP SLA probes [12].

Sample configuration of two IP SLA probes is given below.

```
ip sla 1
 icmp-echo 10.0.0.1 source-interface Serial0/0
 timeout 1000
 threshold 1000
 frequency 30
ip sla schedule 1 life forever start-time now

ip sla 10
 udp-jitter 192.168.56.2 5000 codec g729a
 frequency 30
 ip sla schedule 10 life forever start-time now
```

The probe No. 1 was used to test reachability of a selected node (HUB router with the address 10.0.0.1), while probe No. 10 was used to verify the quality of the connection (transmission delay, packet loss, jitter).

In addition to the probes, the ability to track the status of the connection between nodes (up/down), with sending the syslog trap message to the management station, was used. A simple example of node's reachability monitoring, along with the illustration of sending messages to the management station when the Spoke loses connection with the Hub and the connection is restored, is given below.

```
track 1 rtr 1 reachability
 delay down 90 up 90

event manager applet track_SLA_1
 event track 1 state any
action 1.0 syslog msg "IPSLA collector 1 time out"


Dec 30 11:28:34.181: %TRACKING-5-STATE: 1 rtr 1 reachability Up->Down
Dec 30 11:28:34.217: %HA_EM-6-LOG: track_SLA_1: IPSLA collector 1 time
out

Dec 30 11:31:34.185: %TRACKING-5-STATE: 1 rtr 1 reachability Down->Up
Dec 30 11:31:34.245: %HA_EM-6-LOG: track_SLA_1: IPSLA collector 1 time
out
```

The result of the measurement of the quality of the connection between the network nodes is a series of statistics, as shown in the listing below. The most important are marked in bold.

```
WR#show ip sla statistics

Round Trip Time (RTT) for      Index 1
Latest RTT: 52 milliseconds
Latest operation start time: 14:09:33.018 MyZone Mon Dec 30 2013
Latest operation return code: OK
Number of successes: 23
Number of failures: 1
Operation time to live: Forever


Round Trip Time (RTT) for      Index 10
Latest RTT: 195 milliseconds
Latest operation start time: 14:09:03.222 MyZone Mon Dec 30 2013
Latest operation return code: OK
RTT Values:
Number Of RTT: 990    RTT Min/Avg/Max: 20/195/635 milliseconds
Latency one-way time:
  Number of Latency one-way Samples: 409
Source to Destination Latency  Min/Avg/Max: 2/72/240 milliseconds
Destination to Source Latency  Min/Avg/Max: 2/224/453 milliseconds
Jitter Time:
      Number of SD Jitter Samples: 987
      Number of DS Jitter Samples: 982
Source to Destination Jitter Min/Avg/Max: 0/17/211 milliseconds
Destination to Source Jitter Min/Avg/Max: 0/25/324 milliseconds
Packet Loss Values:
      Loss Source to Destination: 0          Loss Destination to Source: 0
      Out Of Sequence: 0     Tail Drop: 9
      Packet Late Arrival: 0  Packet Skipped: 1
Voice Score Values:
      Calculated Planning Impairment Factor (ICPIF): 13
MOS score: 4.00
Number of successes: 21
Number of failures: 2
Operation time to live: Forever
```

The availability of NHS servers was supervised (in accordance with the proposed algorithm) by Spoke nodes with reachability testing probes. In the case of the unavailability of the NHS servers, the management station, on the basis of received messages-traps, decides to make Spokes and Hubs reconfiguration (firstly, resets the

connection between the Hub and Spoke—*clear dmvpn session*, secondly, indicates a new set of primary and secondary NHS server) or/and to start the procedure of mutual testing Hub nodes to the designation of a new set of NHS servers and the new allocation of clients to servers.

In the case of removing one of the Hub from the set of Hubs, Hub named Reserve was included to the set of Hubs. Test procedure, consistent with the PMC model, was initiated by the manager and the Reserve's job was to initiate testing of all Hubs. It was assumed that Reserve is a reliable node (reachable by all the other nodes).

Functional test performed by a single Hub relied on the use of probe like No. 10, although only the reachability test was taken into account. Hub was treated as capable of realize its function if it was able to communicate with all of its Spokes (Spoke-1 and Spoke-2).

Specifying a new allocation can be carried out on the basis of analysis of the results of testing the quality of the connection between potential NHS servers and Spoke nodes (probe like No. 10). It should be noted that the test results can be used to determine the new allocation taking into account the response times of nodes, packet loss, jitter, but also current load related to the number of supported Spokes, CPU utilization, memory consumption, etc.

Currently, the new allocation was implemented on the basis of assigning Spokes to the Hub that hosts the least Hubs.

In the case of temporary unavailability of the NHS server, with tunnels protected by IPSec, the management station, on the basis of received message-trap from the Spoke, has decided to log on the Spoke and to break active IPSec session (clear crypto session).

VI. Conclusion

The experiment had the hallmarks of a “manually” controlled experiment (supervised). The development work on fully automate the procedure of DMVPN network management is underway. The authors have a preconception about the effectiveness of the proposed solution. The tests (manual inspection and reconfiguration of the system) did not allow to assess the impact of the procedure on convergence time of large DMVPN network. It seems that a good means of verification and comparison proposed algorithm with system without modification or with other solution are simulation studies, which the authors intend to accomplish in the near future. Also, an interesting issue seems to be develop effective procedure for load balancing of Hubs, which will be based on the results of testing the communication parameters in the DMVPN network.

References

[1] J. Maeng, , M. Malek,”A Comparison Connection Assignment for Self-Diagnosis of Multiprocessor Systems”. Digest Int′l Symp.FTC, 1981, pp. 173–175.

[2] M. Malek,”A Comparison Connection Assignment for Self-Diagnosis of Multiprocessors”, Systems, Proc. Seventh Int′l Symp. Computer Architecture, 1980, pp. 31–35, http://dx.doi.org/10.1145/800053.801906

[3] A., Sengupta, A. T. Dahbura,”On Self-Diagnosable Multiprocessors Systems: Diagnosis by the Comparison Approach”, IEEE Trans. Comput., 41, 11, 1992, pp. 1386–1396, http://dx.doi.org/10.1109/12.177309

[4] F. P. Preparata, G. Metze, R. T. Chien, R.T,”On the Connection Assignment Problem of Diagnosable Systems” IEEE Transactions on Computers Vol. EC-16 No. 6, 1967, pp. 848–854, http://dx.doi.org/10.1109/PGEC.1967.264748

[5] F. Barsi, F. Grandoni, P. Maestrini, ”A Theory of Diagnosability of Digital Systems”, IEEE Transactions on Computers, Vol. C-24, Np. 6, 1976, pp. 585–593, http://dx.doi.org/10.1109/TC.1976.1674658

[6] A. Arciuch, “Reliability state of connections in a microprocessor network with binary hypercube structure”, Electrical Review, R.86 No. 9/2010, pp. 154–156.

[7] S. L. Hakimi, A.T.Amin, “Characterization of Connection Assigmanet of Diagnosable Systems”, IEEE Transactions on Computers 1, 1974, pp.86-88, http://dx.doi.org/10.1109/T-C.1974.223782

[8] L. Zang, D. Ardagna, “Sla based profit optimization in autonomic computing systems”, Proceedings of ICSOC’ 04, 2004, http://dx.doi.org/10.1145/1035167.1035193

[9] Y. Hamadi, “Continuous resources allocation in Internet data centers”, CCGrid 2005. IEEE International Symposium on, 2005, pp. 566-573, http://dx.doi.org/10.1109/CCGRID.2005.1558604

[10] K. Lu, R. Yahyapour, P. Wieder, C. Kotsokalis, E. Yaqub, A. I Jehangiri, Y. Hamadi, “QoS – Based Resource Allocation Framework for Multi-Domain SLA Management in Clouds”, International Journal of Cloud Computing, Vol. 1, No. 1, 2013.

[11] „Dynamic Multipoint VPN (DMVPN) Design Guide”, Cisco Systems, Inc. 2006.

[12] „Cisco IOS IP SLAs Configuration Guide”, Cisco Systems, Inc. 2008.

[13] J. Chudzikiewicz, K. Murawski, ”Wyznaczanie bezkolizyjnych dróg przesyłania danych w sieci teleinformatycznej o strukturze typu hipersześcianu”, Diagnostyka 3(39), pp. 131-136 (in Polish).

[14] J. Chudzikiewicz, Z. Zieliński, “Resources placement in the 4-dimensional fault-tolerant hypercube processors network”, Studia Informatica Universalis, Volume 11 (2013), Number 1, pp.1-20

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 955–960
DOI: 10.15439/2014F177
ACSIS, Vol. 2

Anonymization of data sets from Service Delivery Platforms

Radosław Naumiuk (1,2)
(1) Orange Labs CBR
ul. Obrzeżna 7
02-691 Warsaw, Poland
(2) Military University of Technology
ul. Gen. Sylwestra Kaliskiego 2
00-908 Warsaw, Poland
Email: radek92n@gmail.com

Jarosław Legierski
Orange Labs
CBR
ul. Obrzeżna 7,
02-691 Warsaw, Poland
Email: jaroslaw.legierski@orange.com

***Abstract*—The paper presents an anonymization of telecommunication data sets collected through Service Delivery Platforms (SDP), and describes an example tool SDPAnonymizer to make such operation. Information from SDP are processed in form of log files, consisting data sets, which show activity of users of APIs (Application Programming Interfaces). Data sets which should be anonymized contain sensitive data, for example: Names, MSISDN numbers (Mobile Station International Subscriber Directory Numbers) or IP addresses processed by Service Delivery Platforms..**

I. Introduction

IN ICT infrastructure of telecommunication operators, there is processed a large number of data related to subscribers activity and data from flow of information between different systems (call and messaging flows, billing, payment etc.). One of the systems taking part in processing of those information flow, is a Service Delivery Platform (SDP), which is e.g. used for exposure telecommunication services in Internet in form of API (ang. Application Programming Interface), for external programmers. In SDP, a large number of information is being processed, which is related to activity of users and used APIs. Among this data, there are both: information which can be public accessible (for example, the global number of API calls), and as well sensitive data, which publishing is impossible without the anonymization data sets for example the number of MSISDN, ID of localized people, content of SMS, MMS or USSD (Unstructured Supplementary Service Data) messages, Names etc.). This publication focuses on information processed by SDP and proposing a way of their anonymization, based on processing single data sets and presenting an example of application for processing them.

The article is organized as follows: Part I is a short introduction. Part II includes a review of some of the different works related to anonymization of telecommunication operator's data. Part III describes concept of Service Delivery Plaforms. Part IV concentrates on anonymization methods used in the application prototype. Part V presents examples of data sets undergoing the anonymization. Part VI describes the architecture of a solution and part VII shows plans of its further development. The last VIII part contains a short summary.

II. Telecommunication Data Anonymization

In literature, there can be found some examples of anonymization of information used in telecommunication systems. In position [1], the authors give examples of anonymization of IP addresses, made in real time, used during passive monitoring of networks based on TCP/IP protocol. Many works concentrates on issues of anonymization of information related to localization of mobile terminal (Location Based Services - LBS) [2], [3] made in order to provide privacy and security of information about subscriber's location (for example in case of unauthorized use of LBS). Many of data repositories on telecommunication operator's side has the information stored in form of records in databases. Due to that, the most safe way of anonymization is data processing on level of whole repositories, for both single records as well as their sets (data sets) according to k-anonymity methods [4], proposed in following literature positions [5], [6], and widely discussed in part IV of this paper.

III. Service Delivery Platforms

SDP (ang. Service Delivery Platform) is a system present in architecture of telecommunication operator systems, that manages network enablers and open API to allow third users to use these enablers. Such services can be for example:

978-83-60810-58-3/$25.00

functionality prepaid or exposing of functions of the operator in form of API e.g. in order to build services (in the sense of applicationsproviding specific functions) outside the operator - [7], [8], [9], [10]. The role of the SDP is creating and control of sessions and protocols for chosen service. Versatility of SDP usage makes, that it is present in almost every interaction of user with telecommunication service. This fact makes the service platform the source of large amount of data, which can be used in business, statistic or maintenance purposes, however they must undergo anonymization similarly often, before they can be provided.

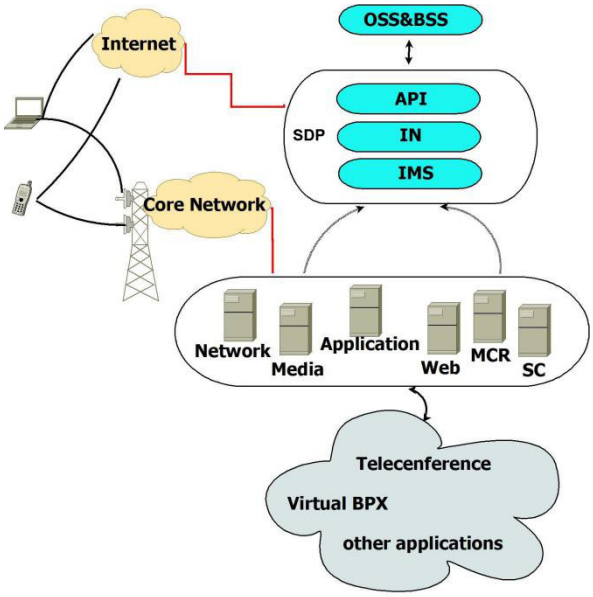


Fig. 1.Concept of Service Delivery Platform

IV. ANONIMIZATION METHODS

Anonymization of data is an operation of data processing, in way to disallow identification of individuals and sensitive data, existing in the processed data, in a way so that the data set remained as readable as possible, and so that it doesn't lose any of the content of general nature, and especially statistical data. Because in the information received from service platform, the basic unit of data is a single record, which represents - from the point of view of project-oriented programming - an event, in the following publication, the authors concentrated on processing of data in form of single records in data set, saved in logs of the SDP platform, provided in form of text files.

The operation of data processing focuses on methods based on data deletion and methods based on data disturbance [5], [6].

Among the most often used methods based on data deletion there are:

1) the deletion of variables - deletion of a record of direct identifiers, which allow identification (e.g. MSISDN number, name, surname, IP addresses etc.)

2) the deletion of records - deletion of individual records in data set. This method is advised only, when identification of sensitive data is possible despite the usage of different anonymization methods.

3) global recording - is based on aggregation of data in data set, according to predefined criteria (e.g. the number of terminals in given area defined by Local Area Code - LAC in various time intervals).

The methods based on data disturbance, in case of application presented in the following article, concentrate on methods from post-randomization group.In case of data from SDP, was implemented algorithms such as MD5 or SHA2 and replacement of sensitive data present in record, with character string generated by used hash function.

1) the method based on MD5 algorithm (Message-digest algorithm version 5). MD5 is an algorithm from cryptography field. It is a popular cryptographic shortcut function, which generates, from any data string its 128-bit hash .

2) the method based on SHA algorithm (Secure Hash Algorithm) - SHA is a family of connected with each other, cryptographic shortcut functions, designed by NSA (National Security Agency) and published by National Institute of Standards and Technology. It has several versions: SHA-1,SHA-256,SHA-384,SHA-512. In SDPAnonymizer application, was used the SHA algorithm, in SHA-256 version).

V. TEST DATA SET

As test data, was used a set of records from test Service Platforms exposing API to Orange network. Data sets were in form of text files in two versions:

Data partially aggregated - set of files with API users activities sorted on the basis of the MSISDN number in form of several files:

getlocation.csv - terminal location api function calls

getop.csv – which operator function calls

sendsms.csv –Send SMS function calls

sendussd.csv –Send USSD function calls

Every of presented above of files had following structure:

DATA HOUR|LOGIN=MSISDN

Where: DATA HOUR – date and time of the event, LOGIN number MSISDN of the API user e.g.:

2011.11.29 18:17:17|LOGIN=48500163047

The second type of data were raw data in txt format in form:
recordID Data Time Event GivenNameName MSISDN IP address deviceidhttpsessionID

2691 237633 2011-03-09 14:43:57 Authentication attempt (ussd pass sent) JaroslawLegierski48500163047192.168.20.124 null 1299678218892
2693 265971 2011-03-09 14:44:41 index.jsp - Authentication successfull JaroslawLegierski48500163047 192.168.20.124 null 1299678218892
2694 265971 2011-03-0914:45:20WebServ:http://10.255.240.50 :2006/tp/orangelabs/jslee/oc/webservices/sendRestUssdNotify?number=48500163047&text=Default+JSLEE+USSD+WebService+text&webSerName=sendRestUssdNotify Jaroslaw Legierski48500163047 192.168.20.124 null 1299678218892
2695 265971 2011-03-09 14:46:17 WebServ:http://10.255.240.50:2006/tp/orangelabs/jslee/oc/webservices/sendRestMMS?number=48508367971&subject=test+123&priority=Low&text=Default+MMS+text+sent+from+JSLEE+WebService+1233&webSerName=sendRestMMS JaroslawLegierski48500163047192.168.20.124 null 1299678218892
2696 265971 2011-03-09 14:47:20 WebServ:http://10.255.240.50:2006/tp/orangelabs/jslee/oc/webservices/getRestTS?number=48500163047&webSerName=getRestTS JaroslawLegierski48500163047192.168.20.124 null 1299678218892

All of the example presented above records, included activity related to API usage, provided through SDP platform by the users (developers) in form of sending SMS, MMS, USSD message, logging in to the system or usage of mobile terminal location function.

VI. SYSTEM ARCHITECTURE

This section of the publication contains the architecture and functionality of the SDPAnonymizer application. The SDPAnonymizer is a simple application operating on SDP platform files in form introduced in chapter V. The application process the anonymization of data sets included in the files.

Input file containing database, which is the data set to anonymize, is loaded to application, and then modified by one of the user-chosen methods. The final result of the program activity is the output file containing anonymized data. The functionality of application is based among other things on algorithms already existing in Java libraries, such as MD5 [11], [12] or SHA2, and on mathematical operations (e.g. summation, deletion etc.).

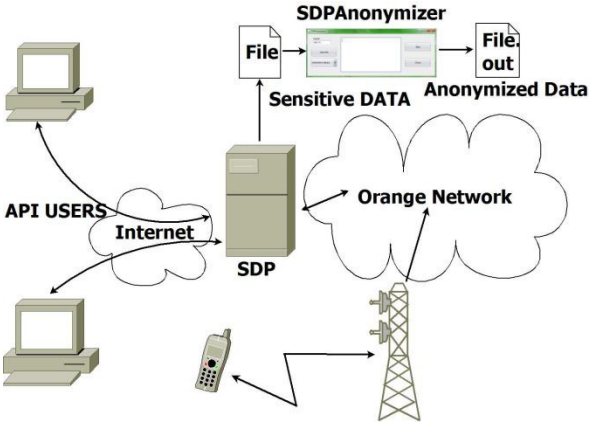


Fig.2. SDPAnonymizer– application concept

A) Programing environment

The application for data anonymization was developed and tested for Java environment JRE in 1.6 and 1.7version. In creation of the SDPAnonymizer application, was used Eclipse environment in "Juno Service Release 2" version. Application user's graphical interface was created with the usage of standard SWING libraries included in JDK.

B) Application

SDPAnonymizeris an application, dedicated for anonymization of data sets from operator service platforms.Tool was developed as easy to use and simple stand-alone solution not integrated with any ETL (Extract, Transform, and Toad)data management framework.

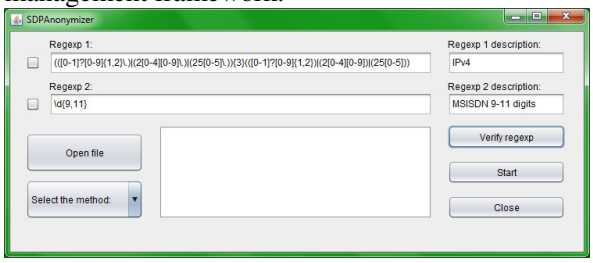


Fig.3. Application window.

In the program, there were implemented methods such as: SUM, DELETE RECORD, CATCH RECORD, DELETE NUMBER, MD5, SHA-256. Program identifies variables for anonymization using regular expressions defined in GUI in two textboxes.
The SUM method is a method from a group of methods based on deletion of data as a part of

global data set recording. In the version implemented in SDPAnonymizer, summation is based on summing of numbers of API usage by one user identified by MSISDN number.

Fig. 4. Application window –SUM method

DELELE NUMBER – is a method from group of methods based on deletion of variables. The MSISDN, IP address or Name is identified in the processed file using regular expressions, and it's deletion, through replacement by XXXX string.

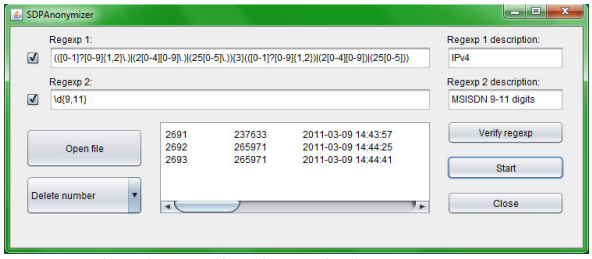


Fig. 5. Application window – DELETE NUMBER

DELETE RECORD method is based on deletion of records in data set. It works by deleting all records containing defined variables (MSISDN, IP address or Name etc.).

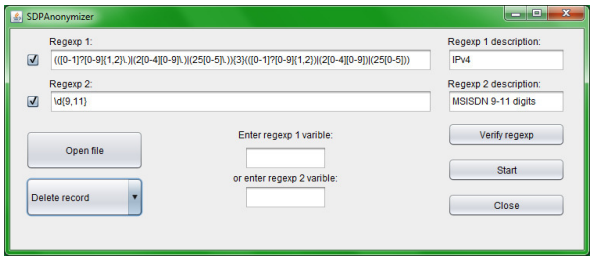


Fig. 6. Application window – DELETE RECORD

Complementary action to the method above is the CATCH RECORD method, which is based on deletion of all records in data set, which do not have defined variable.

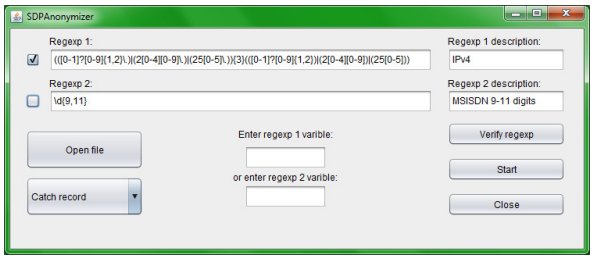


Fig. 7. Application window– CATCH RECORD

As it was already mentioned before, the application also implements two methods from Post-randomization group implemented. MD5 and SHA2 are shortcut functions, which transform a variable defined by regular expression in a hash string. Examples of records including data, for which hash functions were calculated are listed below (for MSISDN numbers anonymization):

Original record:
2691 237633 2011-03-09 14:43:57Authentication attempt (ussd pass sent)JaroslawLegierski485001630471192.168.20.124 null 1299678218892

MD5:
2691 237633 2011-03-09 14:43:57Authentication attempt (ussd pass sent)JaroslawLegierskiaf4f2db11d6a477aeed011a02aa9d549192.168.20.124 null 1299678218892

SHA2:
2691 237633 2011-03-09 14:43:57Authentication attempt (ussd pass sent)JaroslawLegierskid9e88a466dd7cc2527514581b2c73b0bffe34173ce60dd0ace42aa5751e79e19 192.168.20.124 null 1299678218892

Data marked in grey is the effect of encrypting, made with use of MD5 and SHA2 shortcut function.

The program also has the option to define another data than MSISDN, IP address or Name which can be anonymized, through the use of regular expressions(Table1) defined in SDPAnonimizer GUI, which gives us the possibility to anonymize data in different bases (e.g. consisting of telephone numbers of different length and stored in different formats, IPV6 addresses, or terminal location coordinates).

Table 1. Example regular expressionsused in SDPAnonymizer

Regular expressions	Identified variable
\d(9,11)	MSISDN (9-11 digits)
^(25[0-5]|2[0-4]\d|[0-	IPV4 Network Address

1]?\d?\d)(\.(25[0-5]|2[0-4]\d|[0-1]?\d?\d)){3}$	
^(?:[0-9a-fA-F]{1,4}:){7}[0-9a-fA-F]{1,4}$	IPV6 Network Address
[\s-]([A-Z][a-z]*)+[\s-]([A-Z][a-z]*)	GivenName Name

The handling of events and errors, from the application's end user point of view was implemented through display of a dialog windows (pop-ups).

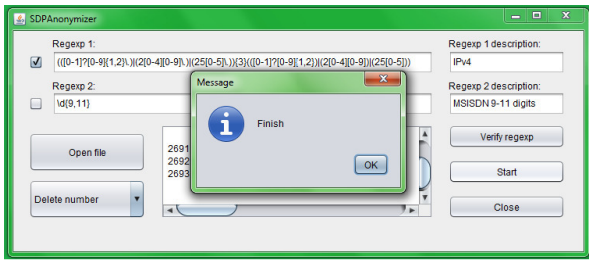


Fig. 8. Application window – dialog window, finish of file processing

VII. FUTURE WORK

In the future, the SDPAnonzmiyer program will be upgraded by functions, doing the process of microaggregation. The Micro-aggregation replaces the anonymized value - with medium value calculated for small group of records (micro-aggregate), which will for example allow to present the usage of communication functions, shown by API for example in hour intervals. Also additional method of encryption, which is the SHA3 method, will be added. In the next step the performance, and load test of application SDPAnonymizerare planned.

Target task is to transform the application into web service form, and exposure of anonymization methods in form of RESTlikeAPI, in order to increase the availability of anonymization methods for users - programmers.

V. SUMMARY

The SDPAnonymizer application is dedicated for Service Delivery Platform administrators, which, for example, share data from SDP system logs with external parties (e.g. the suppliers, which are responsible of the Level 3 maintenance of service platforms). Anonymization of information processed by service platforms in form of dedicated application, working on files solves following problems:

Legal (sensitive data such as MSISDN aren't shared) and data are anonymized as close as possible to the data source (anonymization is done by SDP administrator).

Computational - the SDP platform does not process data, so it's not overloaded with additional, computationally expensive tasks (data processing is done by additional component outside the SDP environment)

Mathematical - system administrator receives dedicated tool, with proper algorithms and mathematical functions already implemented (e.g. MD5, SHA).

Prototype of application SDPAnonymizer was made as part of the Open Middleware 2.0 Community by Orange Labs program[13]

REFERENCES

[1] Ubik, S.; Zejdl, P.; Halák, J., "Real-time anonymization in passive network monitoring," Networking and Services, 2007. ICNS. Third International Conference on , vol., no., pp.100,100, 19-25 June 2007

[2] Heechang Shin; Atluri, V.; Vaidya, J., "A Profile Anonymization Model for Privacy in a Personalized Location Based Service Environment," Mobile Data Management, 2008. MDM '08. 9th International Conference on , vol., no., pp.73,80, 27-30 April 2008

[3] Gedik, B.; Ling Liu, "Protecting Location Privacy with Personalized k-Anonymity: Architecture and Algorithms," Mobile Computing, IEEE Transactions on , vol.7, no.1, pp.1,18, Jan. 2008

[4] Latanya Sweeney K-Anonymity: A Model For Protecting Privacy, Journal International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems archive Volume 10 Issue 5, October 2002,pp 557 - 570

[5] Alpa K. Shah,State-of-art in Statistical Anonymization Techniques for Privacy Preserving Data Mining International Journal of Computer Science & Engineering Technology (IJCSET) Vol. 3 No. 7 July 2012

[6] L. Jaganraj, S. Balamurugan, Empirical Investigation on Certain Anonymization Strategies for Preserving Privacy of Social NetworkInternational Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 10, October 2013

[7] Podziewski, A.; Litwiniuk, K.; Legierski, J., "Emergency button — A Telco 2.0 application in the e-health environment," Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on , vol., no., pp.663,677, 9-12 Sept. 2012

[8] Wawrzyniak, P.; Korbel, P.; Borowska-Terka, A., "Student information delivery platform using telecommunications open middleware APIs," Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on , vol., no., pp.871,874, 8-11 Sept. 2013

[9] Korbel, P.; Wawrzyniak, P.; Grabowski, S.; Krasinska, D., "LocFusion API - Programming interface for accurate multi-source mobile terminal positioning," Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on , vol., no., pp.819,823, 8-11 Sept. 2013

[10] Korbel, P.; Skulimowski, P.; Wasilewski, P.; Wawrzyniak, P., "Mobile applications aiding the visually impaired in travelling

with public transport," Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on , vol., no., pp.825,828, 8-11 Sept. 2013

[11] Junwei Zhang, Jing Yang, Jianpei Zhang,Yongbin Yuan KIDS:K-anonymization data stream base on sliding window

[12] http://md5calculator.chromefans.org/

[13] Open Middleware 2.0 Community portal – http://www.openmiddleware.pl [20.05.2013]

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 961–968
DOI: 10.15439/2014F175
ACSIS, Vol. 2

MonSamp: A Distributed SDN Application for QoS Monitoring

Daniel Raumer, Lukas Schwaighofer, and Georg Carle
Technische Universität München, Department of Computer Science, Network Architectures and Services
{raumer|schwaigh|carle}@in.tum.de

***Abstract*—Software Defined Networks are intended to be less complex, more flexible, and free of vendor-lock-ins. Therefore the Software Defined Networking (SDN) instantiation OpenFlow has been designed according to these properties. The efforts are expected to result in lower expenditure and operational costs. To reach these objectives, mechanisms of classical networks that provide established functionalities have to be revalued and either transformed or redesigned from scratch to take advantage from SDNs. In this paper we describe our vision on flow sampling suitable for traffic monitoring in those networks. Without the loss of generality our approach was specifically created to monitor the quality of service for flows. We describe monitoring as one of possibly many applications that communicate with the SDN controller via the SDN Northbound API. We implemented a prototype SDN application called MonSamp and performed tests to demonstrate the feasibility of our concept.**

***Keywords*—SDN, QoS Monitoring, SDN Application, Flow Sampling, Northbound API**

I. Introduction

TRIGGERED by OpenFlow [1] (OF) the rise of SDN breaks one of the basic network design principles that was valid since the first days of the Internet [2]: the avoidance of central instances as these can be a single point of failure that compromise the robustness of the network. The main idea of SDN is the separation of control and data plane. The most prominent SDN instantiation, OpenFlow is managed by the Open Networking Foundation (ONF [3]). It provides access to the forwarding plane of network devices. Flexible controller software that performs the control plane tasks runs on commodity hardware while switches serve the data plane according to the rules received via the OpenFlow protocol. Uniform data plane hardware avoids vendor lock-ins and is expected to reduce capital expenditure as less functionality is required from it. The wide area network of Google, B4, demonstrated the feasibility [4] of SDN deployments in large and productive wide area networks. While the OpenFlow protocol itself is qute stable, real world SDN applications and their requirements for the SDN Northbound API (NB), allowing the combination of multiple, modular SDN applications into a single consistent controller, are still research in progress [5].

At the time of writing, SDN is still an active research topic. Only very few known SDN installations used for productive, non-research networks exist. One example is the B4 network [4]. With this paper we want to add techniques for sampling traffic suitable for quality of service analysis to the available SDN tools, as the OpenFlow protocol itself only provides little feedback (similar to SNMP) about the network traffic, which is unsuitable for an in-depth quality analysis. Thus, we introduce an SDN application for traffic sampling designed with the requirements for a later quality of service (QoS) analysis in mind. Our goal is to extract a subset of the packets handled by the switch, without limiting the selected traffic to a few ports or suffering from seemingly random drops because of an over-utilized monitoring link. We achieve this using the capabilities that are already present in SDN migrated networks, i.e. commodity servers and OpenFlow switches. We do not introduce a new technique for data analysis or traffic sampling. Our focus is on moving sampling away from dedicated hardware into cheap standard network hardware using SDN techniques.

The paper is structured as follows: In section II we give an overview to the state of the art in network monitoring. Section III describes the challenges of QoS monitoring and discusses ways of integrating these into SDNs. Based on these we infer properties for the sampling and present our architecture. In section IV the prototype, the setting in which we tested our prototype, and first measurement results that demonstrate the capabilities of our model are presented. Afterwards, in section V we discuss the state of the art and related publications. We conclude with by highlighting our contributions, giving a wrap up of open research questions and providing an outlook to our intended future work in section VI.

II. Excursus to Network Monitoring

Gaining information about the network performance is a worthwhile objective that is not unique to SDNs but has been addressed in classical networks since the first days. It is important to differentiate between the productive and the monitoring functionality. The latter could be omitted without having any direct effect for the information delivery. In this chapter we provide a short overview to the state of the art in QoS-monitoring and discuss what distinguishes QoS monitoring from other monitoring objectives.

A. State of the Art in QoS-Monitoring

Network monitoring techniques can be classified according to many criteria: They can be active or passive and require end host control or just access to the network. They may run on a single node or require aggregation of information gathered on different locations, require dedicated hardware and software or be part of existing network equipment. The techniques can be vendor specific, standardized or open solutions.

The value of the information provided by the different techniques for QoS monitoring is highly diverse. Active mea-

978-83-60810-58-3/$25.00

surement tools like SmokePing are used to monitor the latency and connectivity of a specific path but are impractical for monitoring each possible connection in a network. Passively collecting information about the network traffic is a service provided by most forwarding network devices. Information about flows can be pushed to a monitor using protocols like IPFIX (formerly NetFlow) or actively pulled from a monitor via device management protocols like SNMP, NETCONF, or OpenFlow. Theoretically almost any information can be exported via these protocols. However, each network device needs the capability to collect the required network information and to be able to export it to a collector or monitor where information of different networking devices can be combined or analysed. Although OpenFlow is intended to provide vendor independent access, the heterogeneity of the networking hardware available at the market leads to a situation where only selected devices are able to provide the desired information with an acceptable performance. Thus, specific requirements may limit the choice of vendors which again creates a dependency similar to vendor lock-ins. So, in practice, the usefulness of QoS information exported by the network devices is very limited. Additionally, pulling information too frequently causes high load to the network devices which may have a negative impact on the performance of the productive network (c.f. [6]).

The other widely used monitoring technique we will briefly introduce is the collection of local network information by so called sniffers. Sniffers may be hardware components or software tools; e.g. Wireshark (former Ethereal) or tcpdump. They are able to collect and log network traffic and usually require truncating, filtering, and aggregation of the collected packets to extract useful metrics. Software tool kits like VERMONT [7] contain a set of modules that allow configuring a sniffer with adequate post-processing functionality tailored to almost any problem. Other solutions like the intrusion detection system Snort [8] can also be described as sniffers with post-processing functionality focused on detection of network attacks. In state of the art approaches sniffers recognize packet drops by looking at higher level information like TCP retransmits and round trip times. Because the sniffers only have access to traffic of a certain network link or node, the information is necessarily incomplete: This approach is unable to determine where a drop occurred and is not applicable to UDP traffic used by most time-critical applications. To allow monitoring nevertheless, the analyser needs to understand each higher level protocol (e.g. SIP) that is to be monitored. The downside of this approach is that the monitoring has to be implemented for every protocol on top of UDP and may not be available for some protocols used in the network (especially proprietary protocols can pose problems). While it is theoretically possible to combine the information from multiple sniffers, the requirements to do so are quite high. The exactly same traffic has to be monitored on multiple locations (which may be a problem with sampling). Too much aggregation will diminish the usefulness but without aggregation the load on the analyser is very high as all the monitored packets from the whole network need to be correlated.

B. Sampling for QoS-Monitoring

In opposite to security monitoring where we need to find traffic patterns outlining an attack, QoS monitoring can be tackled by sampling. Whenever we detect bad performance of a single flow, we assume that other flows that are in some ways similar, e.g. they also traverse the same overloaded path, will suffer as well. Note that this assumption holds only to performance problems that are caused by the network and not necessarily to those caused by an application running on an end host. If monitoring is applied to all network nodes the performance degradation can be quantified per hop and thus per physical link or network device. Based on our assumption information about the performance gained from the sampled network traffic can be extrapolated to the whole network traffic using the sampling rate.

The IETF working group on Internet Protocol Performance Metrics (IPPM) focuses on developing and maintaining "standard metrics" that can be applied to the quality, performance, and reliability of Internet data delivery services and applications running over transport layer protocols [9]. The QoS metrics can be gathered on almost all layers of the ISO OSI model: e.g. connection establishment time, connectivity, (maximal/minimal) round trip time, delay, jitter, out of order delivery, (maximal/minimal) throughput/goodput, or packet/information loss.

For the quality perceived by the user the term quality of experience (QoE) is used. These metrics need to differentiate between different types of traffic, e.g. when using the network for telephony and the delay becomes greater than 150 ms the user experience will be bad (c.f. [10]) but same delay for a file transfer does not impact the user experience. Another property is the relation between metrics that may be gathered on different layers. Some indicators are highly dependent on the type of application and often difficult to quantify, but network problems detected on high levels are caused by lower levels. Higher level indicators are thus correlated to lower level metrics. For example layer 1 errors lead to retransmits on layer 2 which lead to more traffic causing a lower throughput. This may lead to congestion causing full buffers on layer 3 that cause higher delays or even drops of packets. The result are retransmits or missing packets on layer 4 impacting the performance of the application and thus the user experience. So it is desirable to gain information on a low protocol layer. For practical reasons the IP layer is used when a solution is intended to be applied to network traffic.

To allow detection of single lost packets, flow level sampling of network traffic is preferable to other sampling techniques as it allows the analysis of complete flows. For instance flow level sampling still allows determining performance metrics like the number of dropped packets and one way delays on the IP layer between two monitoring points. A drawback with state of the art monitoring approaches (c.f. section II-A) is that monitoring of newly deployed protocols is hard and requires adaptation of monitoring mechanisms, whereas the flow level sampling approach would instantly recognize lost packets for any IP-based traffic. Using the flow-tuple to match specific parts of the traffic allows (pre)classification by the network and dealing with it according to its needs. For example a VoIP flow constantly needs low latency and jitter, whereas a file transfer flow just needs high bandwidth on average (it does not even suffer much from frequent bandwidth variations). For scenarios where the flow-tuple is insufficient, the packets can be classified and tagged based on more complex features (c.f. [11], [12]). If different classes of traffic have been

introduced by tagging at the network edges the matching rules can also be refined to match only a subset of the tagged traffic.

Another desirable feature is selective monitoring of certain kinds of traffic, e.g. allowing to decouple statistic and network performance collection from the data plane. A network may have various monitoring systems that focus on different aspects. E.g. a system that measures the performance of telephony traffic is not required to receive any other traffic in the network, while a second monitoring system may not require the telephony traffic.

From a network operator's point of view it is desirable to recognize bad service quality as easily and universally as possible. Network operators are usually not in control of the end nodes of the connections. This condition differs from scenarios like the Google B4 [4] where traffic can be elastically delayed because of the control over the end-hosts.

III. Bringing Monitoring Functionality to SDNs

SDN, in particular OpenFlow, is expected to overcome main problems of classical networks: complexity, inflexibility, and vendor-lock-ins that cause high costs. Consequently we expect to benefit from these advantages when designing a monitoring solution for SDNs.

Software monitoring tool kits, like VERMONT [7], can easily be deployed on commodity servers, are flexible, and can thus be adapted for many different purposes including security monitoring, network logging and gathering QoS metrics (c.f. section II). However, placing a monitor to every link in order to get the complete picture of the network does not scale. Thus it is common practice to place monitors only at central nodes in the network. Therefore, in a typical scenario, the monitors are connected to one or more switches. The bandwidth of this link is orders of magnitude smaller than the backplane capacity of the switch and may act as bottleneck limiting the monitored traffic. The switch is then configured to send a copy of the forwarded packets to the port connected to the monitor. The monitoring system analyses the traffic and extracts performance metrics. These can be visualized for an operator to take appropriate actions or, in a more advanced setup, directly given to the SDN controller and its applications. For example a traffic engineering application can react to congested links by preferring alternative paths. Such an approach also provides a solution to challenges of distributed monitoring by utilizing the central controller of SDNs (c.f. section II-B).

This approach differs from another way of transferring traffic information about the network to the controller that we want to discuss briefly: The initial authors of OpenFlow [1] already intended a configuration where the controller receives all packets not matched by an OpenFlow rule on the switch. However, handling all packets that way significantly impacts the performance. To mitigate they propose an architecture for monitoring where a subset of the traffic is redirected to a programmable packet processing system (e.g. NetFPGA) as installed into the OpenFlow rules. The advantage of this approach is more flexibility as it allows modification and filtering of the network traffic but introduces extra delays, costs, and potential bottlenecks. Therefore we consider it as attractive for security motivated scenarios. On the other hand, this overhead is unnecessary if only passive monitoring and no modification or filtering is desired.

Even in scenarios where the bandwidth is sufficient, other bottlenecks and the unpredictable CPU capacity limits the monitoring capabilities even further. Braun et. al. [13] mitigated the unpredictability and demonstrated that dynamic adoption is a suitable way to avoid loss of monitoring information as long as the whole monitoring system has free capacities. Still, exceeding the link or processing capabilities leads to uncontrolled and thus random loss of packets in the monitoring components. These cannot be distinguished from real packet loss occurring in the productive network parts and are affecting the real traffic. In order to optimize monitoring performance with minimal costs packets should only enter the monitoring network and connected monitoring systems if they are likely to be processed. Our approach utilizes OpenFlow capabilities for fine grained matching of the desired packets and allows balancing the monitoring load. It provides an intelligent alternative to classical monitoring ports that are either affected by tail dropping behaviour on the monitoring link or block productive traffic if the transmit queue for the monitor port is full.

A. Northbound API

In SDN the interface between the controller and the higher level applications and network functions is the Northbound API. It is the logical consequence of the divide et impera principle in software development to hide complexity that is irrelevant for the applications behind clear interfaces. Network functionality is separated from the controller software by the Northbound API. The importance of the Northbound API increases even more with the ongoing efforts of scaling controller architectures. E.g. Kang et al. introduced the idea of providing a "One Big Switch" abstraction to the application [14]. A network operator who wants to implement a network function (e.g. security policies) does not want to think about OpenFlow rules, the underlying controller hierarchy, or the combination of these policies with others. Resulting from our monitoring application we identified the following requirements for the Northbound API:

- **Shared contexts** with different levels of (topology) abstraction should provide access to the information that is required and influenced by different applications. These levels also have to hide information, protect related configuration options for faulty access, and keep application design as easy as possible. E.g. an application for BGP routing can handle the network as one big switch [15] while traffic engineering in the network requires differentiated views of the single switches in the network [14].
- **Conflict free controller behaviour** has to be guaranteed through intelligent failover techniques [16], composition techniques for different applications [17], and prioritization of rules [18].
- **Information transparency** is required in both directions. The Northbound API not only needs to translate abstract rules into OpenFlow rules for the switches but also requires the ability to react to incoming OpenFlow packets sent from switches because none

of the installed rules matched. For `proactive` SDN applications only the translation of the rules is required while `reactive` applications need to be aware of incoming OpenFlow packets too.

Until a standardized Northbund API is established, flexible and transferable SDN applications are out of reach. The relevance of the Northbound API has been noticed by the ONF which started to analyse existing approaches for the Northbound API in 2012. The ONF classified existing Northbound APIs according to their scope and application (from OpenFlow enabled networks to general networks) and their level of abstraction. The definition of the Northbound was so unfocused in terms of the abstraction level that the spectrum ranges from the actual Southbound protocol OpenFlow, addressing a concrete switch to high level virtual network management interfaces like OpenStack (c.f. [19]). Existing SDN surveys contain comparisons of different implementations of the SDN Northbound API [20], [21]. The definitions used in the surveys assume an abstraction level somewhere in the middle of the mentioned extremes. In 2013 the ONF created a new working group with the goal to collect requirements and to implement first use cases [19].

B. Architecture

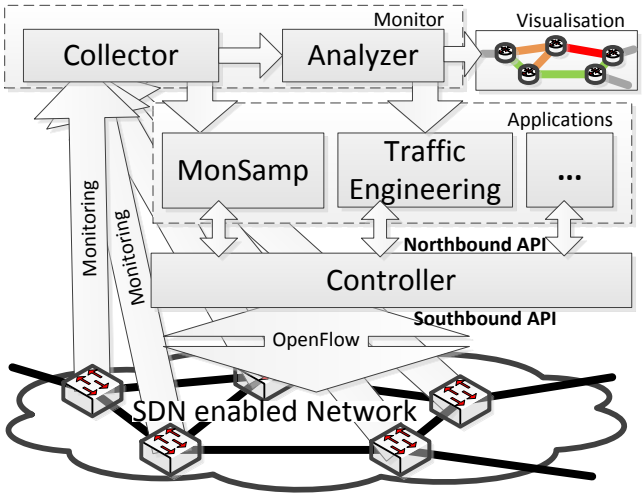


Fig. 1: SDN with integrated monitoring application

Fig. 1 displays an SDN with integrated monitoring functionality. The monitor receives parts of the network traffic for analysis. A sampling application called MonSamp decides which flows to duplicate and send to the monitor. The monitor consists of different modules (here collector and analyser) that are connected adequately for monitoring purposes. The reactive flow sampling application (MonSamp) receives information from the collectors regarding their current workload. Using this information the flow sampling can reactively decide to reduce the number of flows (to avoid random drops). If the links are not working to capacity and the collectors are not fully utilized, the amount of flows can be also be dynamically increased. The selection range for traffic replication is transformed into OpenFlow rules that are installed on the switches by the controller and the Northbound API. In order to allow different applications on top of a controller to interact within the same network slice the Northbound API has to provide abstraction and the ability to combine the behaviour of different applications to a consistent controller behaviour. For example, the monitoring rules must not overwrite or supersede rules required for forwarding of the productive traffic. On the other hand we define a specific interface for communication between the monitor and the flow sampling application. MonSamp uses thresholds to adjust the amount of monitoring load that is forwarded by the SDN enabled devices to allow monitoring without drops. Thereby it mitigates the tail drop behaviour of classical monitoring ports on network devices that occurs whenever a buffer is full. These thresholds are dependent on the concrete monitoring setup and have to be set according to it. MonSamp allows the configuration of monitoring destinations, so that the workload can be split amongst multiple monitoring instances, either though a dedicated physical monitoring network or through a virtual monitoring network. The sampling application also ensures that packets belonging to the same flow are replicated and monitored on all monitoring devices (capacity permitting), thus allowing meaningful correlation. The design supports horizontal and vertical scaling of monitoring systems: It can balance the load to both different monitors with redundant capabilities and different monitors with different objectives (e.g. to a security and a QoS monitor).

In general we cannot assume that all applications running on the Northbound API are reactive. In fact, reactive rules are expected to make the controller a bottleneck of the network [22]. So whenever rules are installed before they match flows or when rules use wildcards the flow sampling application needs information about which flows to replicate for monitoring. The use of event-based OpenFlow messages can again degrade the performance of the controller. To bootstrap the knowledge we imagine to have an ordinary monitoring link. The traffic on this link can be analysed and can serve as input for the flow sampling application. For gaining knowledge about existing flows the random drop behaviour is unproblematic. Another solution is the use of wildcard rules that may be problematic for the switch performance (c.f. section IV-E).

One of the design goals of MonSamp is to avoid any negative influence to the productive traffic. That includes the resources used on the OpenFlow enabled devices. Therefore MonSamp limits the number of installed OpenFlow rules for the monitoring. However, controllers currently do not have any awareness of the performance implications of the installed OpenFlow rules, because the OpenFlow protocol lacks this kind of information [23]. This limitation will be discussed more thoroughly in section IV-E on infrastructural requirements.

IV. Case Study

To provide a proof of concept and to demonstrate the feasibility of our proposed architecture we implemented MonSamp as a prototypic flow sampling application for SDNs and performed experiments for evaluation.

A. Prototype

As discussed in section III-A the definition of the Northbound API is an ongoing process in the community. For our

prototype implementation the Northbound API Pyretic seemed the most fitting. The domain-specific language Pyretic is a successor of Frenetic [24], was published in 2013 [17], [25], and rapidly gained attention. We decided to use it for the following reasons: Pyretic allows for automatic combination of different SDN applications. It provides the required level of abstraction and internally uses the POX controller to communicate with the OpenFlow switches. Pyretic supports both `reactive` and `proactive` applications. It can also be run in an `interpreted` mode, where every packet is sent through the controller. Both Pyretic and its applications are written in Python making things fairly easy to debug and extend in case of unexpected roadblocks. The downside is the relatively low performance, which is not critical for a prototype. We do not claim our choice of the Northbound API to be the silver bullet or a general best practice decision.

The flow sampling application is responsible for deciding whether a newly arriving flow is monitored. The flow sampling application maintains the currently free capacity of the monitor (and the link to it) in its knowledge base. Based on this information the flow sampling application limits the number of flows to be monitored. It does so by computing a direction-independent hash of the flow 5-tuple and selecting the flows within an adjustable hash-range. We achieve direction independence by lexicographically sorting of the destination and source (IP, port)-tuple before hashing. For each monitored flow MonSamp installs a rule into the OpenFlow switches that triggers the action to send a copy of the matched packets to the monitor.

The described prototype focuses on the evaluation of the MonSamp application. Therefore we implemented the monitor as a stub that runs on Linux. It gives feedback about the current level of received traffic to the flow sampling application. Currently we transmit this feedback on the same link that connects the monitor to the OpenFlow switch. There is no interference between the incoming monitoring traffic and the outgoing feedback as Ethernet provides exclusive channels in each direction. A permanent OpenFlow rule on the switch creates connectivity from the monitor to the flow sampling application.

B. Implementation of the Test Scenario

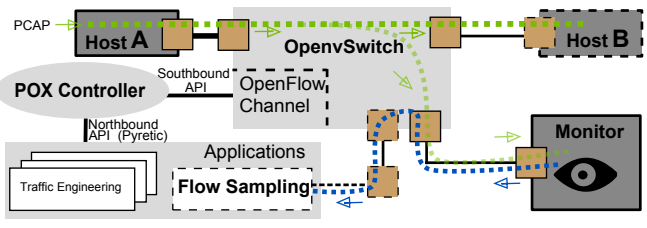


Fig. 2: Realization of the test setup with Mininet

For our case study we used the network emulator Mininet [26] as it allows execution of reproducible network experiments with real world traffic and functional realism [27]. The setup is illustrated in Fig. 2. We created a Mininet topology containing three nodes connected by an Open vSwitch. Mininet separates the hosts by using separate network name spaces, a feature provided by the Linux kernel. As the controller in the default Mininet setup is not part of separate namespace, we added an additional port to the switch and connected it to the root namespace. The flow sampling application, which runs on top of the Pyretic Northbound API, uses this port to communicate with the Monitor. To simulate an application for steering the productive traffic we implemented an application that simply forwards all traffic from Host A to Host B regardless of layer 2 addressing. This allows us to inject traffic containing multiple conversations between different hosts into our test setup. For simplicity we assumed that the productive application acts reactive. Thus the controller notifies the flow sampling application on the arrival of new flows.

C. Measurement Results

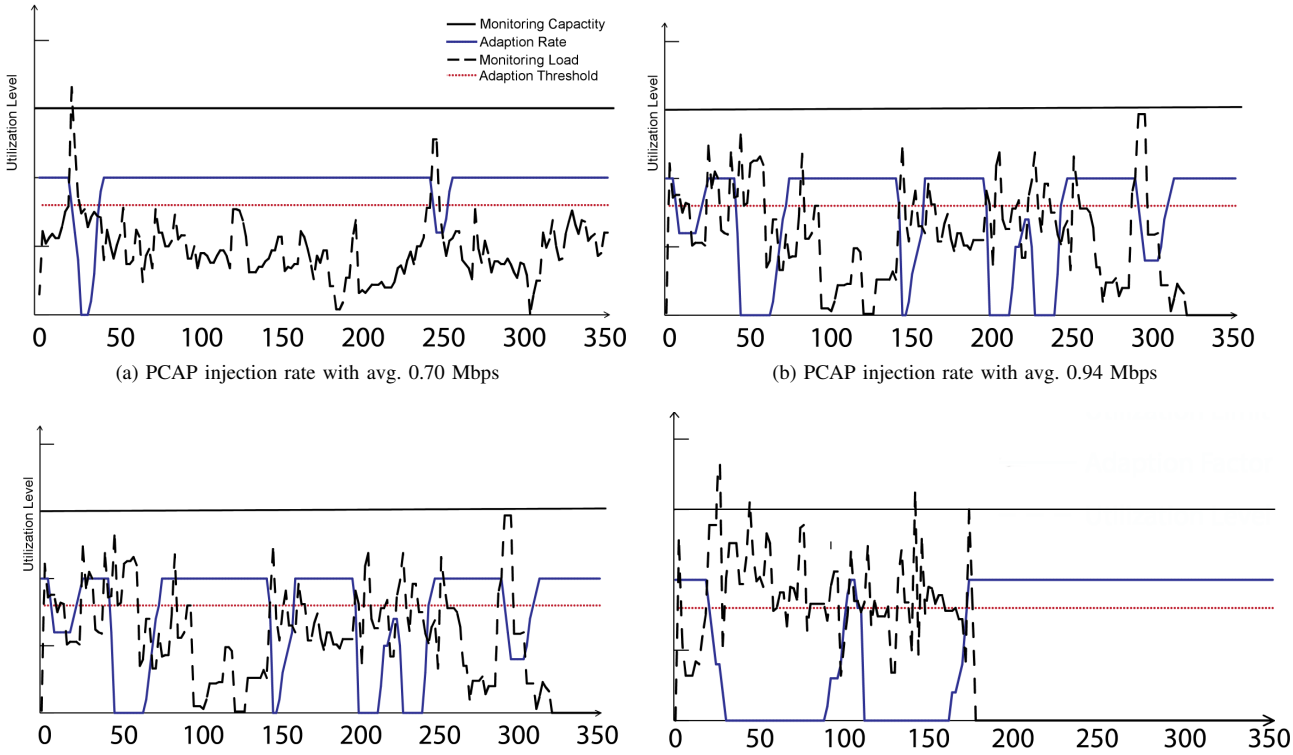
For a first concept evaluation we performed test runs with a real network trace taken from parts of our campus network. Although pseudonymized, we consider the traffic as realistic as the trace contains all connections passing our gateway node which is used by a group of more than 40 students and researchers. The PCAP trace contains 18,000 packets, distributed into 140 TCP and 53 UDP flows. Therefore we created a testing scenario where from the view of the switch and the controller not only two (host A and B) communicate with each other, but more than hundred pairwise different hosts. However, this scenario is representative for real world networks as host A and B represent next hop neighbours of the switch. The monitoring link has a bandwidth of 1.0 Mbps, which we chose as an exemplary link limitation that the Mininet tetbed still can serve without problems.

Figure 3 shows the monitoring utilization during the replay of previously recorded real world traffic through our test setup with four different average speed levels: 0.70 Mbps, 0.94 Mbps, 1.17 Mbps, and 1.40 Mbps. Our application is configured to keep monitoring utilization at 50 % of the theoretical capacity (red doted line). For these initial measurements we assume the limiting factor for monitoring to be a constant link bandwidth. Another relevant factor can be the dynamically changing CPU utilization [13]. The blue line represents the adaptation factor that is the input parameter for the decision function to determine the ratio of new flows to be monitored. The dashed line is the real bandwidth processed by the monitoring system. The Monitor reports this back to the MonSamp application.

The tests show that monitoring of flows without uncontrolled drops in the monitoring system is possible. The few areas where the monitoring load exceeds the capacity are mainly caused by the relative big impact of elephant flows, which are flows that contain a very large share of transferred bits of the overall traffic. We do not expect this to be a problem when applied to high-bandwidth scenarios where new flows are created and old flows are finished more frequently. Thus the impact of a single flow in relation to the total amount of traffic is lower.

D. Open Issues with the Northbound API

Unfortunately, Pyretic – the Northbound API we chose for our implementation – is currently not suitable for realising our



(a) PCAP injection rate with avg. 0.70 Mbps

(b) PCAP injection rate with avg. 0.94 Mbps

(c) PCAP injection rate with: avg. 1.17 Mbps

(d) PCAP injection rate with: avg. 1.4 Mbps

Fig. 3: Sampling adaptation for different bandwidth utilizations

monitoring vision in a meaningful way. The main limitation is the lack of a policy that can make a decision based on an incoming packet. The dynamic policy, as supported by Pyretic, only allows adjusting the policy for future packets, not for the one that is being processed. This means that a default policy decision for the MonSamp application is required – either the first packet of each flow is always monitored or it is never monitored. Because (unless running in interpreted mode) the default policy decision for each flow gets pushed to the OpenFlow enabled hardware, subsequent packets are handled directly on the OpenFlow hardware and not sent to MonSamp any more, rendering the policy-adjustment useless.

Furthermore, Pyretic and other reviewed Northbound APIs for SDNs lack control for the rules that are installed on the OpenFlow switch. Rules are automatically installed based on every decision and applied to the whole flow. It is not possible to define additional rules (e.g. the reverse flow) to be installed or changing the header-fields used for identifying a flow.

E. Infrastructural Requirements

The application of our vision requires both, hardware and software features that provide fields for research. Our monitoring concept requires a scalable controller architecture that provides a reasonably powerful Northbound API. We expect that the desired controller infrastructure will exist in the near future, since first steps were already taken [22], [19]. Thereby the development process of controller software and architectures benefit from the low market barriers, numerous players, and almost no proportional costs for additional features of software developments. These characteristics do not apply for hardware OpenFlow-switches. These switches may support OpenFlow 1.3 – at least after updates of firmware – but still suffer from processing in software on the general purpose CPU if certain OF rules are defined. This results in an extreme heterogeneity of switch behaviour [23] that may lead to poor forwarding performance and an unpredictable capacity for stored OpenFlow rules on the switch. As an example the HP OpenFlow Guide [28], which applies to switches in our testbed, only states forwarding to a single port, dropping, and modification of the IP TOS and VLAN-PCP field as actions that can be executed in hardware; all other actions are executed in software and therefore limit the processing capacity of the whole switch to a rate of around 10,000 packets per second. In virtual switches like Open vSwitch the replication of traffic only introduces a comparatively small and constant overhead per processed packet. The exact overhead is defined by the network stack that is used to send out the traffic to the monitoring system (e.g. to a VM, or to a physical connection). Better optimized hardware for OpenFlow and feedback channels from the OpenFlow switches via the Northbound API to the applications can solve these problems.

V. Related Work

Related work in network monitoring is primary motivated by security concerns and dates back some years. E.g. in 2005 Schaelicke et al. [29] discussed requirements for parallel network monitoring and proposed an architecture for adapting

load balancing of security monitoring traffic. Limmer and Dressler created an adaptive load balancer for NIDS systems. It uses sampling to cope with high bandwidths [30]. It dynamically maps flows to NIDS processes. To alleviate issues resulting from packet drops they use flow sampling, which guarantees loss-free analysis for selected flows.

When monitoring the QoS it is an advantage that a network problem resulting in degraded performance will affect a whole class of flows that share some properties (i.e. routed through an overloaded device and use the same QoS class). Thus, unlike for security purposes, it is sufficient to monitor a representative subset of the network flows which makes QoS-Monitoring eligible for sampling. As we do not provide a new sampling technique we just highlight the long history of sampling dating back to mathematical work on statistics over many decades. For a focused introduction we refer to work of Claffy et al. [31] (general overview), Carela-Español et al. [32] (study of sampling influence for traffic analysis), Braun et al. [13], who recently implemented an adaptive sampling within a monitoring system to mitigate tail dropping behaviour within the overloaded system, and referenced in there.

Using OpenFlow switches for load balancing services (e.g. web servers) in data centres was one of the first use cases of OpenFlow switches. In 2009 Handigol et al. used a reactive NOX controller to minimize response time for load balanced web servers without IP address rewriting [33]. Uppal and Brandon also described a reactive NOX-based load balancer that does address rewriting [34]. Wang et al. presented a NOX based, proactive load balancer that uses wildcard OpenFlow rules with the motivation that reactive approaches causes undesired load to the controller [35]. The controller assumed that the load of each server is proportional to the number of flows directed to it, but did not consider feedback from the servers. Due to their flexibility and the resulting capabilities, OpenFlow switches have also been recognized as a powerful tool for solving scalability issues in network monitoring. For instance big switch network already sells Big Tap, a network monitoring solution [36], [37]. Big Tap uses a separate OpenFlow enabled (monitoring) network equipped with monitoring systems to deliver, filter, and analyse traffic in a scalable manner. Recently Shirali-Shahreza et al. proposed a concept for OpenFlow based traffic sampling called FleXam [38], [39]. They demonstrated how OpenFlow functionality can help detecting attacks in the network [38].

VI. Conclusion

In this paper we described our vision on traffic monitoring in SDNs. Our architecture is scalable and cost efficient because utilizing the already existing OpenFlow functionality does not add any additional costs. All special monitoring functionality is provided by flexible software. We designed an SDN application for extraction and sampling of network traffic directly from the data plane that can be added to other applications via the SDN Northbound API. The application addresses unsolved and solved problems in (QoS) network monitoring with high flexibility and low costs. Thereby we also contributed to the open topic of the canon of SDN applications and requirements for them (c.f. [5]). We implemented a prototype and performed different measurements on it to demonstrate the feasibility of our approach.

Our future work comprises the transfer of our application into a real test environment which also provides more than one switch. Thus we can extend the value of our architecture evaluation beyond functionality (Mininet [27] only provides functional realism). We plan to perform a more sophisticated evaluation, transfer the application to other controller platforms like Ryu, and combining it with monitoring systems like the one proposed by Braun et al. [13]. We also plan further refining to support multiple different limits where each may dynamically interfere as bottleneck and consider sampling coordination together with vertical and horizontal balancing of monitoring loads. These steps should point out required features in future SDNs and demonstrate (e.g. technical) limitations of available SDNs.

Acknowledgments

This research has been supported by the EU as part of KIC EIT ICT Labs on Software-Defined Networking (SDN) and the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung; BMBF) under support code 01BP12300A; EUREKAProject SASER. We also would like to acknowledge the valuable contributions from our colleagues Peter Schaab, Florian Wohlfart, and Lothar Braun to the implementation and maturing of our vision.

References

[1] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "Openflow: Enabling innovation in campus networks," *SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 2, pp. 69–74, Mar. 2008. [Online]. Available: http://dx.doi.org/10.1145/1355734.1355746

[2] B. M. Leiner, V. G. Cerf, D. D. Clark, R. E. Kahn, L. Kleinrock, D. C. Lynch, J. Postel, L. G. Roberts, and S. Wolff, "A Brief History of the Internet," *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 5, pp. 22–31, October 2009. [Online]. Available: http://dx.doi.org/10.1145/1629607.1629613

[3] "ONF - Open Networking Foundation," https://www.opennetworking.org/, April 2014.

[4] S. Jain, A. Kumar, S. Mandal, J. Ong, L. Poutievski, A. Singh, S. Venkata, J. Wanderer, J. Zhou, M. Zhu, J. Zolla, U. Hölzle, S. Stuart, and A. Vahdat, "B4: Experience with a globally-deployed software defined wan," *SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 4, pp. 3–14, Aug. 2013. [Online]. Available: http://dx.doi.org/10.1145/2534169.2486019

[5] B. Munch, "Hype cycle for networking and communications," Gartner, Report, 2013.

[6] C. Rotsos, N. Sarrar, S. Uhlig, R. Sherwood, and A. W. Moore, "Oflops: An Open Framework for OpenFlow Switch Evaluation," in *Passive and Active Measurement*. Springer, 2012, pp. 85–95. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-28537-0_9

[7] "VERMONT - VERsatile MONitoring Toolkit," https://github.com/constcast/vermont/, April 2014.

[8] "Snort," http://www.snort.org/, April 2014.

[9] "IPPM - IP Performance Metrics," http://tools.ietf.org/wg/ippm/, April 2014.

[10] M. Hassan, A. Nayandoro, and M. Atiquzzaman, "Internet telephony: services, technical challenges, and products," *Communications Magazine, IEEE*, vol. 38, no. 4, pp. 96–103, 2000. [Online]. Available: http://dx.doi.org/10.1.1.43.1441

[11] A. Callado, C. Kamienski, G. Szabó, B. Gero, J. Kelner, S. Fernandes, and D. Sadok, "A survey on internet traffic identification," *Communications Surveys & Tutorials, IEEE*, vol. 11, no. 3, pp. 37–52, 2009. [Online]. Available: http://dx.doi.org/10.1109/surv.2009.090304

[12] T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *Communications Surveys & Tutorials, IEEE*, vol. 10, no. 4, pp. 56–76, 2008. [Online]. Available: http://dx.doi.org/10.1109/SURV.2008.080406

[13] L. Braun, C. Diekmann, N. Kammenhuber, and G. Carle, "Adaptive Load-Aware Sampling for Network Monitoring on Multicore Commodity Hardware," in *IFIP Networking 2013*, New York, NY, May 2013. [Online]. Available: http://dx.doi.org/10.1.1.395.9415

[14] N. Kang, Z. Liu, J. Rexford, and D. Walker, "Optimizing the one big switch abstraction in software-defined networks," *Proc. ACM CoNEXT*, 2013. [Online]. Available: http://dx.doi.org/10.1145/2535372.2535373

[15] A. Vidal12, F. Verdi, E. L. Fernandes, C. E. Rothenberg, and M. R. Salvador, "Building upon routeflow: a sdn development experience," in *XXXI Simpsio Brasileiro de Redes de Computadores, SBRC'2013*, 2013.

[16] M. Reitblatt, M. Canini, A. Guha, and N. Foster, "Fattire: declarative fault tolerance for software-defined networks," in *Proceedings of the second ACM SIGCOMM workshop on Hot topics in software defined networking*. ACM, 2013, pp. 109–114. [Online]. Available: http://dx.doi.org/10.1145/2491185.2491187

[17] C. Monsanto, J. Reich, N. Foster, J. Rexford, and D. Walker, "Composing software-defined networks," in *Proceedings of the 10th USENIX Conference on Networked Systems Design and Implementation*, ser. nsdi'13. Berkeley, CA, USA: USENIX Association, 2013, pp. 1–14. [Online]. Available: http://dx.doi.org/10.1145/1384609.1384625

[18] A. D. Ferguson, A. Guha, C. Liang, R. Fonseca, and S. Krishnamurthi, "Hierarchical policies for software defined networks," in *Proceedings of the first workshop on Hot topics in software defined networks*. ACM, 2012, pp. 37–42. [Online]. Available: http://dx.doi.org/10.1145/2342441.2342450

[19] Open Networking Foundation, "Northbound interface working group charter," 2013.

[20] B. A. A. Nunes, M. Mendonca, X.-N. Nguyen, K. Obraczka, and T. Turletti, "A survey of software-defined networking: Past, present, and future of programmable networks," *COMMUNICATIONS SURVEYS AND TUTORIALS*, 2014. [Online]. Available: http://dx.doi.org/10.1145/505733.505735

[21] D. Kreutz, F. M. V. Ramos, P. Verissimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-Defined Networking: A Comprehensive Survey," 1414.

[22] A. R. Curtis, J. C. Mogul, J. Tourrilhes, P. Yalagandula, P. Sharma, and S. Banerjee, "Devoflow: scaling flow management for high-performance networks," in *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 4. ACM, 2011, pp. 254–265. [Online]. Available: http://dx.doi.org/10.1145/2043164.2018466

[23] P. Pereš íni, M. Kuzniar, and D. Kostic, "Openflow needs you! a call for a discussion about a cleaner openflow api," in *The Second European Workshop on Software Defined Networking, EWSDN 2013*, 2013. [Online]. Available: http://dx.doi.org/10.1109/EWSDN.2013.14

[24] N. Foster, R. Harrison, M. J. Freedman, C. Monsanto, J. Rexford, A. Story, and D. Walker, "Frenetic: A network programming language," *SIGPLAN Not.*, vol. 46, no. 9, pp. 279–291, Sep. 2011. [Online]. Available: http://dx.doi.org/10.1145/2034574.2034812

[25] J. Reich, C. Monsanto, N. Foster, J. Rexford, and D. Walker, "Modular SDN Programming with Pyretic," *USENIX ;login*, vol. 38, no. 5, pp. 128–134, Oct. 2013. [Online]. Available: http://dx.doi.org/10.1.1.403.3030

[26] "Mininet - An Instant Virtual Network on your Laptop," http://mininet.org/, April 2014.

[27] N. Handigol, B. Heller, V. Jeyakumar, B. Lantz, and N. McKeown, "Reproducible network experiments using container-based emulation." in *CoNEXT*, C. Barakat, R. Teixeira, K. K. Ramakrishnan, and P. Thiran, Eds. ACM, 2012, pp. 253–264. [Online]. Available: http://dx.doi.org/10.1145/2413176.2413206

[28] H. O. Switches, "Openflow configuration guide," 2012.

[29] L. Schaelicke, K. B. Wheeler, and C. Freeland, "Spanids: a scalable network intrusion detection loadbalancer." in *Conf. Computing Frontiers*. ACM, 2005, pp. 315–322. [Online]. Available: http://dx.doi.org/10.1145/1062261.1062314

[30] T. Limmer and F. Dressler, "Adaptive Load Balancing for Parallel IDS on Multi-Core Systems using Prioritized Flows," in *20th IEEE International Conference on Computer Communication Networks (ICCCN 2011)*. Maui, HI: IEEE, August 2011, pp. 1–8. [Online]. Available: http://dx.doi.org/10.1109/ICCCN.2011.6006063

[31] K. C. Claffy, G. C. Polyzos, and H.-W. Braun, "Application of sampling methodologies to network traffic characterization," in *ACM SIGCOMM Computer Communication Review*, vol. 23, no. 4. ACM, 1993, pp. 194–203. [Online]. Available: http://dx.doi.org/10.1145/167954.166256

[32] V. Carela-Español, P. Barlet-Ros, A. Cabellos-Aparicio, and J. Solé-Pareta, "Analysis of the impact of sampling on netflow traffic classification," *Computer Networks*, vol. 55, no. 5, pp. 1083–1099, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2010.11.002

[33] N. Handigol, S. Seetharaman, M. Flajslik, N. McKeown, and J. Ramesh, "Plug-n-serve: Load-balancing web traffic using openflow," in *ACM SIGCOMM Demo*, 2009.

[34] H. Uppal and D. Brandon, "Openflow based load balancing," Tech. Rep., 2010.

[35] R. Wang, D. Butnariu, and J. Rexford, "Openflow-based server load balancing gone wild," in *Proceedings of the 11th USENIX Conference on Hot Topics in Management of Internet, Cloud, and Enterprise Networks and Services*, ser. Hot-ICE'11. Berkeley, CA, USA: USENIX Association, 2011. [Online]. Available: http://dx.doi.org/10.1.1.227.7761

[36] "Big Tap," http://www.bigswitch.com/products/big-tap-network-monitoring, April 2014.

[37] big switch networks, "Sollution guide: Open sdn for network visibility - simplifying large scale network monitoring systems with big tap."

[38] S. Shirali-Shahreza and Y. Ganjali, "Efficient implementation of security applications in openflow controller with flexam," *High-Performance Interconnects HotI 2013, Symposium on*, pp. 49–54, 2013. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/HOTI.2013.17

[39] ——, "Empowering software defined network controller with packet-level information," in *Proceedings of the 1st IEEE Workshop on Traffic Identification and Classification for Advanced Network Services and Scenarios, TRICANS'13*. IEEE, 2013, pp. 1355–1359. [Online]. Available: http://dx.doi.org/10.1109/ICCW.2013.6649444

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 969–976
DOI: 10.15439/2014F293
ACSIS, Vol. 2

POI Explorer – A Sonified Mobile Application Aiding the Visually Impaired in Urban Navigation

Piotr Skulimowski, Piotr Korbel and Piotr Wawrzyniak
Institute of Electronics
Lodz University of Technology
Lodz, Poland
piotr.skulimowski@p.lodz.pl, piotr.korbel@p.lodz.pl, piotr.wawrzyniak@dokt.p.lodz.pl

***Abstract*—The paper presents POI Explorer mobile application aiding the visually impaired in spatial orientation and in urban navigation. A user equipped with a smartphone with accelerometer, electronic compass, mobile data transmission and positioning capabilities can access information on nearby points of interest. Maintaining data connection with dedicated system servers provides access to additional services facilitating the navigation in urban areas. The paper describes an overall architecture of the system. Then, the details of the user interface of the application are presented. The user interface of the application was designed to meet both the needs of visually impaired users and the requirements imposed by dynamic data changes. It employs a unique combination of text-to-speech and sonification techniques to ensure clarity of messages as well as high responsiveness of the application. The results of experiments performed in areas with different densities of points of interest proved the usability of the proposed approach.**

***Index terms*—Location-based services, mobile computing, pervasive computing, electronic aids, visually impaired**

I. Introduction

RECENTLY, a number of electronic travel aids (ETA) addressing the needs of the visually impaired have been developed [1-9]. Such systems can be used to overcome difficulties with spatial and geographical orientation and navigation, and facilitate access to various public services. Those difficulties become especially cumbersome in urban areas. Lack of good spatial orientation makes difficult to find a safe path among obstacles, and to locate and identify points of interest (POI) like bus stops, offices, restaurants, or even pedestrian crossings [1].

Dedicated electronic travel aids play a vital role in aiding the visually impaired in everyday activities. Such assistive devices are usually equipped with a GPS receiver to provide precise information on user terminal position, GSM/UMTS transceivers, inertial sensors, and a speech synthesizer enabling auditory form of presentation of various data to the user [2, 4, 8-9]. Such devices can usually be carried in a pocket [2, 4, 9] and because of dedicated user interfaces (tactile keyboards, text-to-speech and speech-to-text systems) are easy to use by visually impaired users [2, 4, 9]. Most of dedicated ETAs allow to lead visually impaired or blind users along previously recorded routes. They also store databases of POIs, which can serve as an aid in orientation. Due to the fact that such devices are produced only in short series, the build quality and functionality of such solutions may be far from that of massively produced electronic devices. The main disadvantages limiting the popularity of such systems include difficulties in upgrading software and high unit costs.

With the growth of popularity of advanced mobile phones, more and more applications aiding the visually impaired in navigation and travelling appear on the market [10-17]. Modern phones are usually equipped with advanced positioning capabilities as well as in a range of additional sensors, like accelerometers, gyroscopes or magnetometers. ETA solutions using mobile phones as user terminals may benefit from the use of detailed digital maps from different providers, like Google Maps, Bing Maps, OpenStreetMap, etc. in combination with satellite (GPS, Glonass) or network (cellular, Wi-Fi) based positioning techniques. Also, mobile phones usually are equipped with good quality speech synthesizers facilitating communication with visually impaired users.

Since most of the users use the built-in text-to-speech and sonification systems, such as VoiceOver [18] or TalkBack [19], navigation applications are adapted to work with these systems. For this reason, application user interfaces use standard system components like buttons, text boxes, lists, etc. The drawback of such an approach is the lack of adaptation of very small items (which you must specify for the sonification) to the screen reader systems, or long, scrollable lists with dynamic contents.

This article describes a novel approach to presenting information about the POIs using combination of text-to-speech and sonification techniques. The remainder of the paper is organized as follows. Section II provides an overview of the system architecture while Section III gives details of the POI Explorer mobile application functionality. A proposed approach to auditory presentation of POI data is described in Section IV, and the results of application tests are presented in Section V. Finally, Section VI summarizes our work.

II. System Architecture

The architecture of the proposed system for guidance of visually impaired in urban environment is shown in Fig. 1. The system consists of two subsystems: application servers and mobile user terminals with dedicated applications.

This work was partially supported by the National Centre for Research and Development of Poland under grant no. NR-02 0083-10 in years 2010-2013.

The mobile user terminal is an Android OS based smartphone equipped with a dedicated POI Explorer mobile application. Since the application uses Sensors API of the operating system, it can be run on devices with Android OS in version 2.3 or newer. The details of the application are provided in Section III.

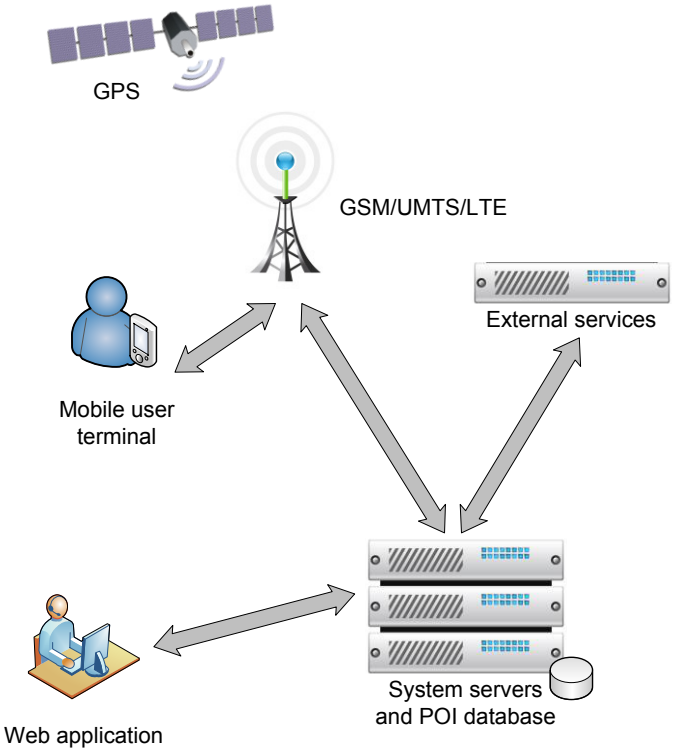


Fig. 1. Architecture of the electronic system for guidance of the visually impaired in urban environment.

The main function of the system server is storing a database of points of interest. The prototype solution uses MySQL relational database management system. POIs are organized into categories and subcategories, what allows to find necessary information easily. The POIs stored in the database can be both public and private. The public POIs are available to all the users of the system, while the private ones are accessible only to the owner. Users can also add additional personalized information to the points (text notes, voice recordings) to enrich the database.

The server also hosts web application for remote POI data management. The web based application (Fig. 2, Fig. 3) uses PHP and AJAX and is dedicated for sighted users assisting the visually impaired and blind. The management application allows to add and remove points of interest. It is also possible to define new categories of the POIs as well as hierarchy of the different categories of points. An important functionality of the system is the ability to define routes which the visually impaired can follow. This may be for example a route leading to the office, shop, etc. The routes can be created in one of two ways: from the POI Explorer mobile application and from the aforementioned web based management application. Examples of such predefined routes are shown in Fig. 2. The management application allows also to delete or change the order of the points in a route.

The data are exchanged between the POI Explorer mobile application and the system server in an XML format. Examples of XML system messages are presented in Fig. 4 and Fig. 5. Hierarchical structure of XML documents allows elastic and extensible POI category management. Moreover, such a solution allows to provide a universal application programming interfaces (API) to other, external platforms. For example, it is possible to import waypoints and routes from external sources like Loadstone database [11].

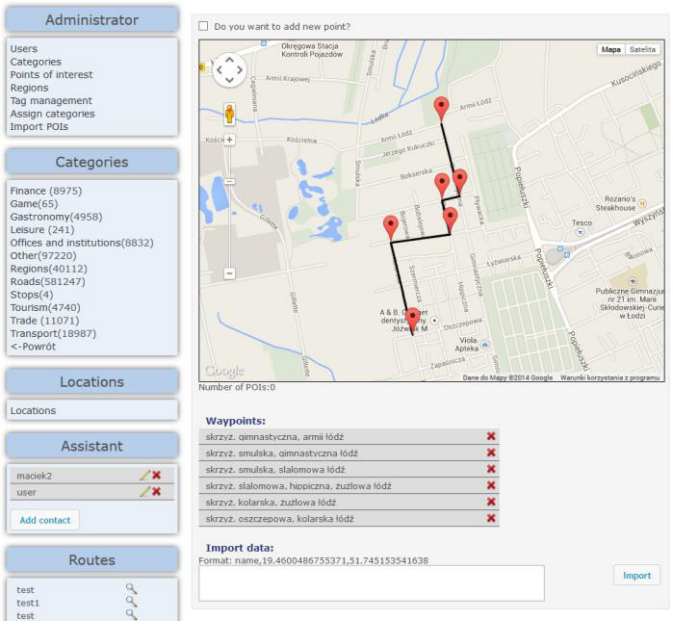


Fig. 2. The graphical user interface of a web application for the management of the POI database – path creation example.

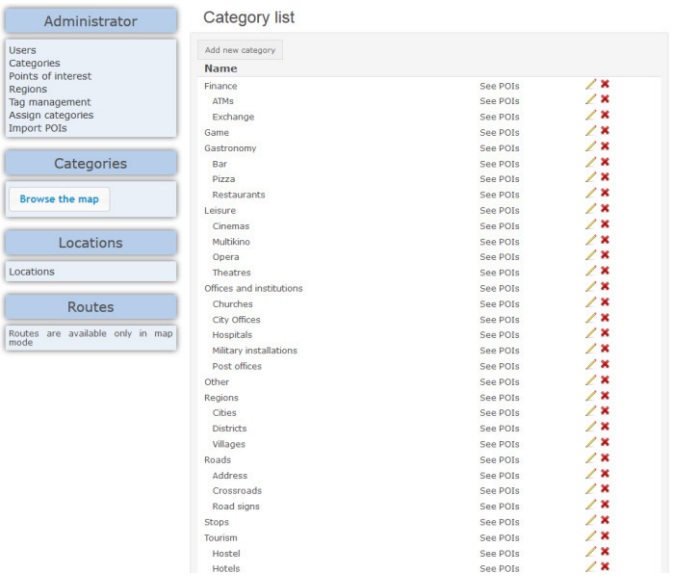


Fig. 3. The web application for the management of the POI database. The categories and subcategories of POIs are shown, each of the database entries can be edited or deleted.

```
<?xml version="1.0" encoding="UTF-8"?>
<main>
<poi name_pl="Adresy" name_en="Address"
     id="30" id_mam="27"/>
<poi name_pl="Apteki" name_en="Pharmacies"
     id="45" id_mam="7"/>
<poi name_pl="Bankomaty" name_en="ATMs"
     id="21" id_mam="20"/>
<poi name_pl="Bar" name_en="Bar"
     id="24" id_mam="23"/>
<poi name_pl="Drogowe" name_en="Roads"
     id="27" id_mam="0"/>
<poi name_pl="Dzielnice" name_en="Districts"
     id="37" id_mam="32"/>
<poi name_pl="Hotele" name_en="Hotels"
     id="46" id_mam="9"/>
<poi name_pl="Pizzeria" name_en="Pizza"
 id="26" id_mam="23"/>
<poi name_pl="pkp" name_en="trains"
 id="36" id_mam="35"/>
(…)
</main>
```

Fig. 4. The example XML server response contains list of categories. Each category has its name in Polish and English language, identification number and identification number of parent category.

```
<?xml version="1.0" encoding="UTF-8"?>
<main>
<poi x="19.46369" y="51.77139" r="1"
name="bankomat millennium ul. narutowicza łódź"
id_cat="21" id="128" id_user="15" active="1"
report="0" kind0="1" kind1="2"/>
<poi x="19.457" y="51.7647" r="1"
name="bankomat bos ul. piotrkowska 103/ 105
łódź" id_cat="21" id="129" id_user="15"
active="1" report="0" kind0="0" kind1="0"/>
<poi x="19.45733" y="51.76407" r="1"
name="bankomat i oddział bph male ul.
piotrkowska 109 łódź" id_cat="21" id="131"
id_user="15" active="1" report="0" kind0="1"
kind1="0"/>
</main>
```

Fig. 5. The example XML server response contains list of points in the given area. Each of the point of interest has GPS cooridnates, radius r in meters, name, category id (id_cat), point id (id), id of user who added this point (id_user), active status (active), number of reported bugs (report), number of text messages (kind0) and voice notes (kind1).

III. SMARTPHONE BASED URBAN NAVIGATION – POI EXPLORER

The first version of POI Explorer application aiding the visually impaired in moving in the urban area was developed for Symbian OS based mobile phones [20]. As Symbian based devices no longer play an important role on the market, the application had to be migrated to a new platform.

Nowadays iOS and Android based mobile phones are of a special interest for blind and visually impaired users. The reason for that is that both the systems offer built-in and well integrated with the operating system text-to-speech modules: Voice Over (iOS) [18] and TalkBack (Android) [19]. Availability of such high quality system modules allows developers to create their own applications using standard GUI elements which can be easily presented to the visually impaired users. Moreover, most of contemporary smartphones are equipped with touch screens and offer gesture-based screen readers supported by multitouch capabilities. For example moving a single finger over the list causes a list item to be read, while moving of two fingers can be used to scroll the list.

After market analysis and discussions with representatives of the target group of POI Explorer users, we have decided to select Android Operating System based devices as a target platform. According to ABI Research Android is the most popular operating system (77% share in Q4 2013) [21]. In 2013, according to Gartner Android market share was 78,4%, while iOS 15.6% [22]. Another advantages include lower costs of Android phones in comparison to iPhones as well as availability of more accessible low-end devices.

The development of the application was based on test results of previous releases for Symbian based mobile phones as well as initial releases for Android OS based devices. About 20 blind and visually impaired volunteers received the aforementioned applications for testing purposes. The volunteers from different regions of Poland were recruited through an advertisements on Loadstone [11] mailing lists. Although now deprecated, at that time Loadstone was one of most widely used applications for navigation of the blind and visually impaired. The users of competitive application positively rated the idea to use a compass to determine the direction of movements (instead of analyzing history of GPS values). There have been requests for the adding information on the number of satellites used to calculate GPS coordinates, which allows them to rate, how much they can trust the messages returned by the application. There were different opinions about the possibility of downloading lists of POIs around the current position of the user. On one hand, a visit to another city does not require prior preparation of a list of points of interest, on the other hand there were concerns that data for some reason would not be downloaded. There were a lot of positive comments about organizing POIs into categories and subcategories, because it allows to reduce the number of points to be presented to the user.

Graphical user interface layout and functionality of POI Explorer mobile application have additionally been consulted with blind and visually impaired users from the Polish Association of the Blind. As the result, the graphical user interface of POI Explorer (Fig. 6, Fig. 7, Fig. 9, Fig. 11, Fig. 14) uses large, high contrast characters aiding users with moderate visual impairment. Also, system requirements of our applications allow to run them on low cost devices.

To use the application, the user needs to set up his/her system account. This is an optional step, however it allows to store private points of interests and routes. It also enables the possibility to adding notes and voice description to POIs. When the user runs the application for the first time, a list of POI categories is downloaded from the server. Moreover, for a selected area (the user can define its radius) a range of POIs are downloaded and stored in local database. It allows the use of the application without maintaining Internet connection.

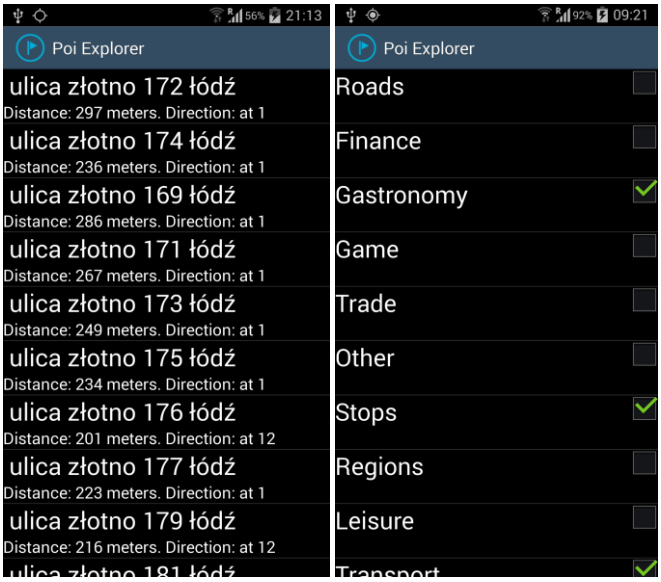


Fig. 6. Screenshots from the POI Explorer application captured by Samsung Galaxy S4 Android Phone with KitKat 4.4.2. The list of the points of interests in the front of the user (direction is calculated based on compass values) are presented in a standard listbox. On the right: list of categories, each of them can have unlimited number of subcategories.

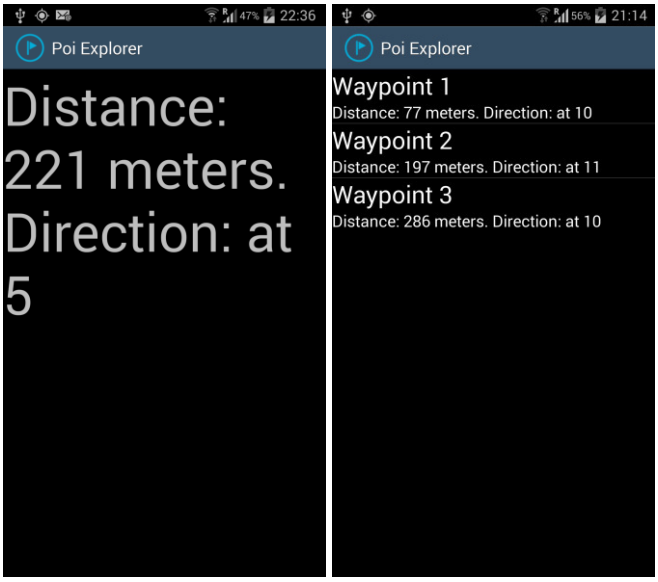


Fig. 7. Screenshots from the POI Explorer application captured by Samsung Galaxy S4 Android Phone with KitKat 4.4.2. On the left: mode distance/direction, on the right: list of waypoints obtained from the Google Maps engine.

The POI Explorer supports the navigation of the blind users in one of three different modes. The users can choose navigation: along a predefined path, to the selected point (e.g. a bus stop), or in a "look around" mode. The distance to a given POI is calculated based on readings from the built-in GPS receiver, while orientation in the area is calculated on the basis of the values returned by the compass module. The data for the user are being refreshed on demand and take into account the current readings of the two aforementioned sensors.

A. Navigation to the specific point

In this mode, the user picks from a list of points of interest any entry. Then the user receives the information about the distance of this point from the current position and direction to the point expressed in the "per hour" mode. Orientation is determined relative to the current position of the phone and the point. POI Explorer calculates the angle between the orientation of the phone (from the compass) and the orientation of the target point, which is then converted to the appropriate "hour". Example message is shown in Fig. 7. In addition, the user can designate the route to the point using Google Maps mechanism, which returns a list of waypoints.

B. Navigation along a recorded path

User have possibility to register the route and determine intermediate points like intersections, characteristic points etc. Example of such route is shown in Fig. 2. The route most often is created with an assistance of a sighted person. The user is guided to the each of the waypoints as described in section IIIA. When the waypoint is reached, POI Explorer changes the destination point to the next one from the list. This mode is most liked by the target group of users due to the possibility of creating waypoints in a supervised way.

Automatic determination of intermediate points is also possible. In this case, there is a risk that the algorithm will not take into account slight changes in direction of the route. Moreover, in this mode it is not possible to automatically add information on dangerous locations, such as defects in the roadway, etc.

C. Look around mode

Look around mode is one of the most important features of POI Explorer especially useful in unfamiliar locations, for example, in an unknown city. A list of points of interest within a certain radius with the indication of the distance and direction is presented to the user. The blind person can restrict presented points only to selected categories. An example of such filtered list is shown in Fig. 6.

POI Explorer was written taking into account the suggestion from the target group of users. Some application capabilities have been developed as the result of requests from our testers. These for example include the use of a compass to determine direction (usually applications determine direction based on the history of GPS values) or importing data from the Loadstone, very popular platform aiding blind and visual impaired people for Symbian OS [11].

Unfortunately, interaction with a touchscreen of a smartphone requires the use of both hands and may be especially uncomfortable for blind users, who at the same time are often using a white cane. Moreover, screenreaders are not very convenient for sonification of dynamic GUI controls.

The tests with the target group of users revealed that the main problem for them is the way of sonification by built-in screenreaders. Every change in the user interface (for example resulting from an update of a text field, or from appearance of a new item in the list) implies that the user is informed about that change. Finding the selected POI often requires reading of the entire list of points (Fig. 6). Any update of the interface contents may interrupt reading of the items from the list and force re-reading of the entire information about the screen elements. In addition to that, every change of position (due to changing user's position or fluctuations in GPS or compass readouts) causes, that the list frequently changes its contents. Moreover, because the list is sorted by distance from the user, the order of the points in the list can also change. The user is notified by the screen reader on every such change. Excessive information becomes hardly understandable and confusing for the user. This problem has become an inspiration to propose another approach to presentation of information on points of interest located in the vicinity of the user.

IV. Sonified Presentation of Information On Nearby POIs

When designing the new method of presentation points of interest, to avoid the previously described problems, the following assumptions have been made:

- the usage of the application interface should be possible with only one hand. Due to the safety of the blind, it is advisable that the phone should be hidden in a pocket, and voice messages could be transmitted via the hands-free set;
- amount of conveyed information should be limited to the minimum, and the number of voice messages should also be reduced. Previously conducted experiments showed that the excess of voice messages is very tiresome and discourage users to use such kind of applications;
- it should be possible to reduce the number of points from the vicinity presented to the user at a single time;
- a way to read detailed information about the point being currently sonified should be provided.

To fulfill of all of the above requirements, it was decided that a unique short sound will be assigned to the each category of the POIs. The points will be presented to the blind person in the order of increasing distance from the user. Only the points located along a straight line (within a predefined angle) from the user will be presented.

After starting of the scan, a short audio message is played. A virtual circle of search is moving away from the user and when it encounters a POI on its way, a sound associated with given category of the point is played. After completion of playing the sound, the radius of the circle of search is increased until it reaches the next point from the user vicinity (Fig. 9, 11). Fig. 9 and 11 show application screens displaying the POIs found in the user vicinity, while Fig. 8 and 10 show these points on the Google Map.

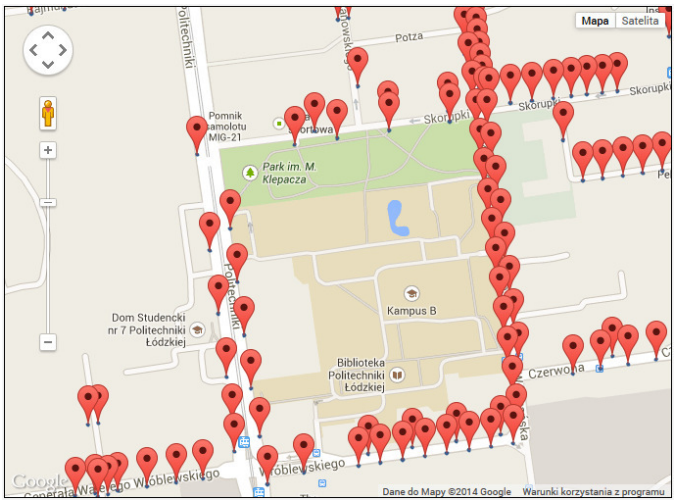


Fig. 8. POIs from the selected category Address shown on the Google Maps (screenshot from the web application). Number of POIs presented to the user is limited to 100. A fragment of the map corresponds to the area shown in Figure 9.

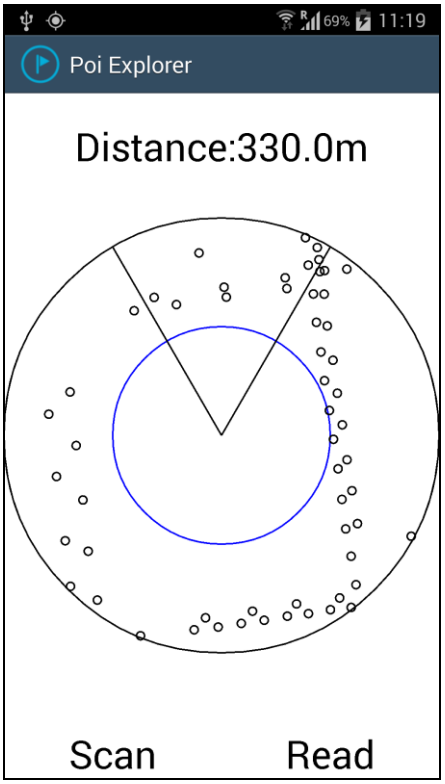


Fig. 9. Screenshot from the POI Explorer application captured by Samsung Galaxy S4 Android Phone with KitKat 4.4.2. Each point of interest from Figure 8 is marked with a circle.

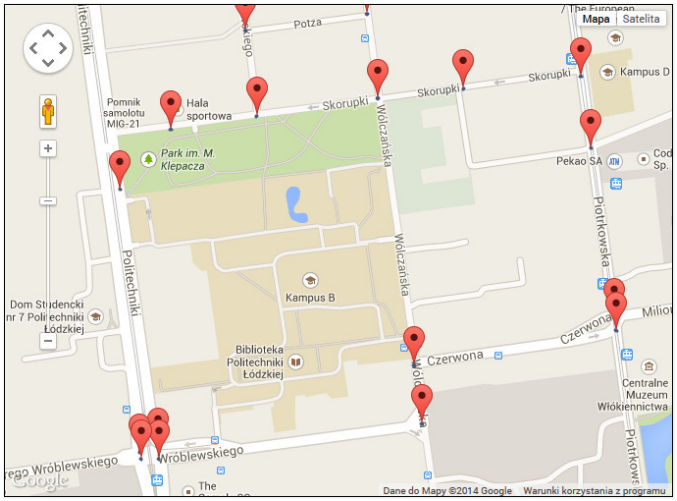


Fig. 10. POIs from the selected category Intersection shown on the Google Maps (screenshot from the web application). A fragment of the map corresponds to the area shown in Figure 11.

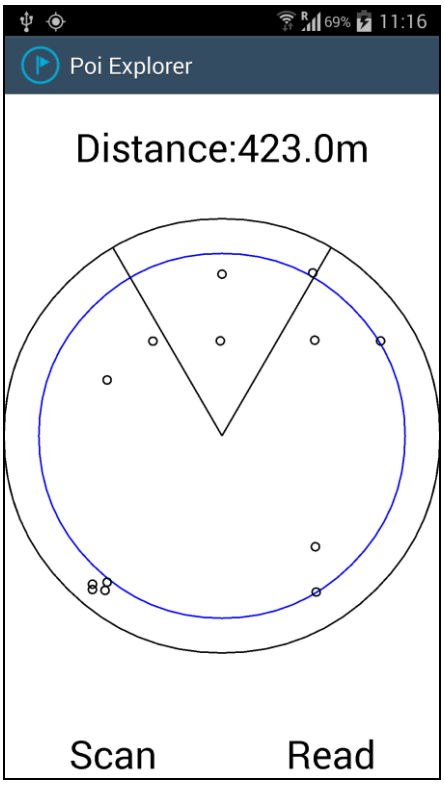


Fig. 11. Screenshot from the POI Explorer application captured by Samsung Galaxy S4 Android Phone with KitKat 4.4.2. Each point of interest from Figure 10 is marked with a circle.

Preliminary tests performed at early development stages revealed that updating the positions of points during the scan often causes confusion, because the scanning time strongly depends on the number of POIs and can take up to several seconds. For this reason, it is assumed that update of the positions of POIs in relation to actual user position occurs at the beginning of the scan.

In the application for navigation purposes we set out three control fields. The field at the top allows the user to change the scan range. The higher the x coordinate of a touch point, the larger the scan area. Change (releasing) the touch location on the screen sets the new value of the radius of the scan, the selected value is read to the user. Scan area at the bottom of the screen causes that scan starts from the beginning. It may be useful at a time when the user changes its orientation and wants to start scanning points after update their relative position. Field Read allows to read information about the recently sonified point. When the user selects this option, the distance of the point and its orientation relative to the user is read.

Internal tests have shown the system accuracy decreases with the growth of the number of points to be presented. An example of such a situation is shown in Fig. 9. Points arranged in lines represent neighboring buildings along the street. Fig. 8 shows the presented area on the Google Map (only addresses and intersections are shown).

To make it easier to read the information about points we introduced manual mode, which uses the readouts from the accelerometer to control the presentation of the POI list (Fig. 12).

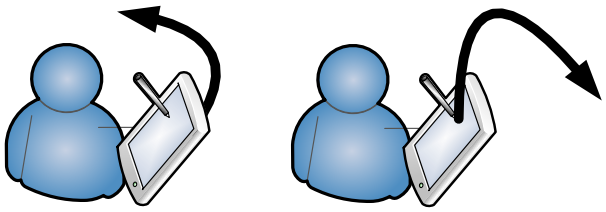


Fig. 12. Gesture based control of the device – Move to the previous POI (left), Move to the next POI (right).

A move of the phone downwards stops scanning and information about the next point on the list is read. Contrary, the upward gesture starts reading of the information about the previous point. The threshold values of data from the accelerometer readings were chosen to eliminate false gesture detection. Our solution allows to manipulate the phone hidden in the jacket pocket. At any time it is possible to finish manual mode and start automatic scanning.

A. The use of built-in motion sensors for gesture recognition

Motion sensors: gyroscope, accelerometer, gravity sensor, linear acceleration sensor can be utilized in finding user gesture patterns. They measure acceleration forces and rotational forces along three axes. Some of them are hardware-based, while others are software-based: they derive data from one or more hardware sensors. Accelerometer sensor measures acceleration force in SI units that is applied to the device on all physical axes. The values include the force of gravity. Gravity sensor (which can be a software sensor) measures only the force of gravity. Gyroscope returns device's rate of rotation around each of the three axes. Linear acceleration works like the acceleration sensor, but values exclude the force of gravity. Not all the sensors are available on every device. For example gravity sensor is not available on devices with API versions lower than 9. Android allows to determine the capabilities of sensors, like maximum range, resolution etc. These values can

differ between devices. It is also possible to get notifications on accuracy changes or when a sensor reports a new value. Each log file entry also contains a timestamp [23].

Block diagram of accelerometer based gesture recognition is shown in Fig. 13. All values (x, y, z components) returned by accelerometer are in SI units. Because a sensor measures acceleration applied to the device, the values returned by the device include gravitational acceleration. During gesture detection we take into consideration only values of y component. Initially, we expect to begin a gesture. It is a condition in which the user is holding the phone horizontally ($|y|<Th_{y1}$). Then we expect to perform the gesture. The user must tilt the phone up or down. If ($|y|>Th_{y2}$) an appropriate action is performed. To detect another gesture, the phone should return to the horizontal position. The algorithm works only when the gravitational acceleration is only applied to the device. All values not satisfying criteria (1) are ignored. During the tests with Samsung Galaxy S4 device, the Th_{y1} and Th_{y2} threshold values were set to 2 and 5 correspondingly.

$$8 < \sqrt{x^2 + y^2 + z^2} < 10 \qquad (1)$$

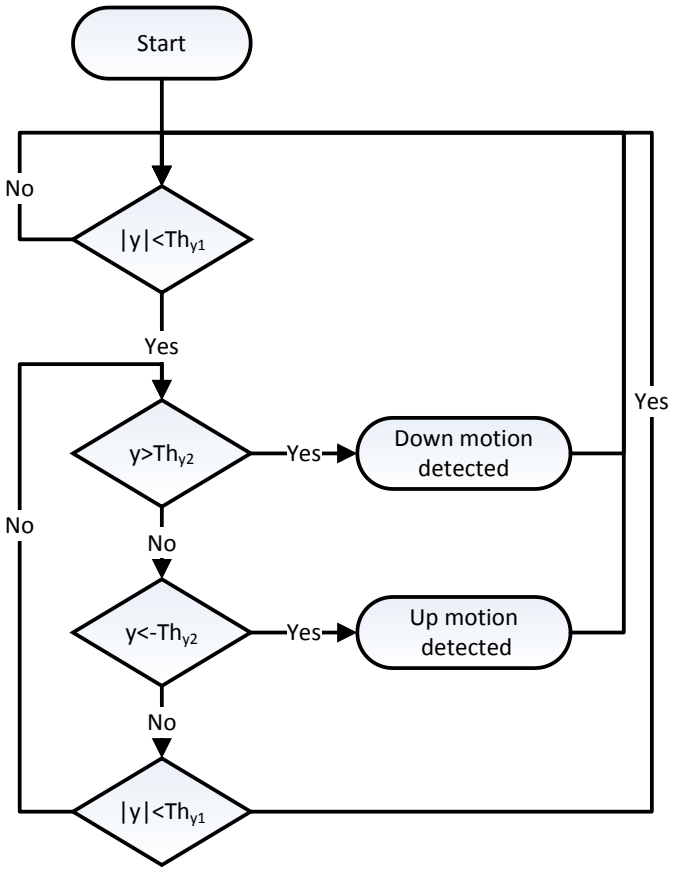


Fig. 13. Accelerometer based gesture recognition – block diagram of the algorithm.

V. APPLICATION TESTS

Application tests were carried out in selected areas of the city of Lodz, in places containing a large number of points of interest. As a test platform we have chosen Android 4.2 and 4.4.2. Fig. 14 shows POI Explorer running on Samsung Galaxy S4 mobile phone. A group of several sighted volunteers at different age and with different familiarity of mobile technologies was involved in the tests.

An option to specify the radius of the scan turned out to be a good solution. It allows to eliminate the problem of too high number of points to be presented at the time. Accelerometer-based control worked properly, there were no problems with reading information about the points in the vicinity. Testers pointed out that despite the possibility to specify the categories of presented points, there should be an option to hide unnecessary points.

During the tests of the application, time of a complete scan was set to about 15 seconds. In case of several points in the vicinity, it was the optimal time for selecting the chosen point of interest by the user.

The active fields (the slider for setting the distance, Scan and Read buttons) in the GUI seem to be well matched. We did not notice problems in locating these fields, even while the phone was kept in the user's pocket.

Fig. 14. POI Explorer launched on Samsung Galaxy S4 Android KitKat phone.

The tests of the application revealed the limitations of audio playback system on Android OS. Playing short audio samples indicating the points ran smoothly. The system allows to play several audio samples at the same time. Playing a voice message interrupts the previous message. Fortunately, it is possible to play sounds when reading messages by Text-To-Speech engine.

Since most of contemporary mobile phones are equipped with electronic compasses, these can be kept in a pocket and the user can use the compass readouts for the orientation

purposes. Text to speech TalkBack service of the phone can be used to read system messages and POI descriptions to the user.

VI. SUMMARY

Although many electronic travel aids have been developed so far, urban areas still remain hardly accessible to the visually impaired. Poor spatial orientation and inability to access textual information makes difficult to move around the cities. Development and growing popularity of new generations of advanced mobile devices opens up new opportunities to create inexpensive and efficient ETAs for the blind and visually impaired. Even inexpensive mobile phones are nowadays often equipped with reasonably good quality sensors (compass, accelerometer, A-GPS receiver). In addition, speech synthesis, text-to-speech and speech recognition systems allow to offer alternative communication interfaces to visually impaired users. Using general purpose mobile devices as platforms for development of ETAs lowers deployment costs of such systems and as a result the user can carry around only a single device instead of a range of devices dedicated to different purposes. Also cloud computing and storage solutions provide higher security and facilitate access to user data in the case of a change of the device.

The system described in the paper is primarily intended to aid the visually impaired in urban navigation. It uses a novel approach to presenting information about the points of interest in the vicinity of the user. The combination of text-to-speech and sonification techniques allows to effectively reduce the amount of excessive information presented to the user and overcome the limitations of popular screen reader systems.

Currently, our efforts are focused on preparation of the release candidate version of the application, that will be made available to a wider group of blind and visually impaired volunteer testers.

ACKNOWLEDGMENT

This work was partially supported by the National Centre for Research and Development of Poland under grant no. NR-02 0083-10 in years 2010-2013.

REFERENCES

[1] P. Strumiłło, "Electronic Interfaces Aiding the Visually Impaired in Environmental Access, Mobility and Navigation," Proc. 3rd International Conference on Human System Interaction, Rzeszów, Poland, 2010, pp. 17–24, http://dx.doi.org/10.1109/HSI.2010.5514595

[2] Kapten Mobility. http://www.kapsys.com/fr/en/products/kapten-mobility/. Accessed 22 April 2014.

[3] Talking Signs. http://www.talkingsigns.com/. Accessed 22 April 2014.

[4] Trekker Breeze handheld talking GPS, http://www.humanware.com. Accessed 22 April 2014.

[5] J. Marski, P. Bajurko, K. Radecki, and T. Buczkowski, „Miniaturowe radiolatarnie i terminale z sygnalizacją RSSI do wspomagania orientacji osób niewidomych," Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne, 6, 2010, pp.320-323. (in Polish)

[6] PAVIP. http://bones.ch/. Accessed 22 April 2014.

[7] Step-Hear. http://www.step-hear.com/. Accessed 22 April 2014.

[8] P. Barański, M. Polańczyk, P. Strumiłło, "A Remote Guidance System for the Blind," Proc. 12th IEEE International Conference on e-Health Networking, Applications and Services HealthCom, Lyon, France, 2010, http://dx.doi.org/10.1109/HEALTH.2010.5556539

[9] NaviEye (Nawigator). http://www.migraf.pl/. Accessed 22 April 2014.

[10] Ł. Kamiński, K. Bruniecki, "Mobile Navigation System for Visually Impaired Users in the Urban Environment," Metrology and Measurement Systems XIX (2), 2012, pp.245-256.

[11] Loadstone GPS – Free GPS Software for Your Mobile Phone. http://www.loadstone-gps.com/. Accessed 22 April 2014.

[12] P. Korbel, P. Skulimowski, and P. Wasilewski, "A Radio Network for Guidance and Public Transport Assistance of the Visually Impaired," Proc. 6th International Conference on Human System Interaction HSI 2013, Sopot, Poland, 2013, http://dx.doi.org/10.1109/HSI.2013.6577869

[13] P. Korbel, P. Skulimowski, P. Wasilewski, P. Wawrzyniak, "Mobile applications aiding the visually impaired in travelling with public transport," Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on , vol., no., pp.825-828, 8-11 Sept. 2013.

[14] P. Eslambolchilar, A. Crossan, R. Murray-Smith, "Model-based target sonification on mobile devices," Proceedings of International Workshop on Interactive Sonification, 8 January 2004, Bielefeld, Germany, 2004.

[15] B. Taylor, Dah-Jye Lee, Dong Zhang, Guangming Xiong, "Smart phone-based Indoor guidance system for the visually impaired," Control Automation Robotics & Vision (ICARCV), 2012 12th International Conference on, pp.871-876, 5-7 Dec. 2012, http://dx.doi.org/10.1109/ICARCV.2012.6485272

[16] K. Koiner H. Elmiligi, F. Gebali, "GPS Waypoint Application," Broadband, Wireless Computing, Communication and Applications (BWCCA), 2012 Seventh International Conference on, vol., no., pp.397,401, 12-14 Nov. 2012, http://dx.doi.org/10.1109/BWCCA.2012.71

[17] Wang Guolu; Qiu Kaijin; Xu hai; Chen Yao, "The design and implementation of a gravity sensor-based mobile phone for the blind," Software Engineering and Service Science (ICSESS), 2013 4th IEEE International Conference on, pp.570-574, 23-25 May 2013, http://dx.doi.org/10.1109/ICSESS.2013.6615373

[18] VoiceOver. http://www.apple.com/accessibility/voiceover/. Accessed 22 April 2014.

[19] Google TalkBack, https://play.google.com/store/apps/details?id=com.google.android.marvin.talkback . Accessed 22 April 2014.

[20] M. Polańczyk, P. Skulimowski, B. Sujecki, and D. Sulmowski, "Personal Navigation System for the Blind based on Points of Interest," Proc. II Forum Innowacji Młodych Badaczy 2011, Łódź, Poland, 2011.

[21] Q4 2013 Smartphone OS Results: Is Google Losing Control of the Android Ecosystem?, https://www.abiresearch.com/press/q4-2013-smartphone-os-results-is-google-losing-con Accessed 22 April 2014.

[22] Gartner Says Annual Smartphone Sales Surpassed Sales of Feature Phones for the First Time in 2013 http://www.gartner.com/newsroom/id/2665715 Accessed 22 April 2014.

[23] Android developer's guide on sensors http://developer.android.com/guide/topics/sensors/sensors_overview.html Accessed 22 April 2014.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 977–984
DOI: 10.15439/2014F104
ACSIS, Vol. 2

Characterizing webpage load from the perspective of TCP connections

Luis Miguel Torres, Eduardo Magaña, Mikel Izal and Daniel Morato
Departamento de Automática y Computación, Universidad Pública de Navarra, Pamplona, Spain.
Email: [luismiguel.torres, eduardo.magana, mikel.izal, daniel.morato]@unavarra.es

***Abstract*—Over the last years websites have evolved rapidly incorporating new content types and becoming more and more dynamic. Users today are able to access a wide variety of content and services through their web browsers. As a consequence, web traffic has become increasingly complex and, from a network perspective it can be difficult to ascertain which websites are being visited by a user, let alone which part of the user's traffic each of them is responsible for.**

Although there is an extensive literature on the new characteristics of web traffic, few works have focused on a connection level perspective even if this kind of data is easily available for network administrators. In this paper we offer a characterization of webpage download using connection level metrics. This description is a first step in developing techniques able to identify individual webpage downloads in real traffic.

We have captured an extensive dataset of more than 20,000 webpage downloads that we study in order to provide different connection level based metrics. We study how these metrics vary between different webpages of different popularity and complexity. In the end, we attempt to provide a general modelling of a normal webpage download.

I. Introduction

THE web is probably the classic Internet application that has grown and evolved the most during the past two decades. The simple and mostly static webpages of the 1990s have given way to much more complex sites. This complexity is represented, in the first place, by adding a wide variety of content types (like videos, scripts or interactive media) to the text and images that classic webpages traditionally hosted. Nevertheless, modern websites not only offer these new content types, but they do so in a dynamic way, keeping their content current and tailoring their offer to each specific visitor. The network requirements introduced by all this and the ever-increasing popularity of the web have also pushed for improvements in the web application protocols and the development of new techniques, like content distribution networks (CDNs) or analytics services, that help in its operation. As a consequence, the web application has achieved a remarkable flexibility that allows it to provide a huge range of different services aside from traditional web browsing.

All these changes have obviously affected the profile of web traffic. Recent studies [1]–[3] show that its characteristics have greatly changed from the (simpler) ones described thoroughly in the 1990s [4]. This is partially the result of the introduction of HTTP 1.1: persistent connections and pipelining have made obsolete the notion that every connection comprises a single request/response pair. But, the truth is that the profile of web traffic has been specially affected by the new contents and services provided by the application. Nowadays, from a network perspective, accessing a webpage may imply establishing multiple connections to different servers while the elements of the webpage (often coming from third parties) are downloaded and, in some cases, user information is collected. The result is a set of a variable number of connections of different durations and sizes to multiple server IP addresses. Moreover, as a sizeable amount of the content is dynamic, these connections may change if the webpage is accessed at a different time or by a different user.

In this paper we present a study of those sets of connections established by clients during the download of webpages. Although there is extensive work in web traffic characterization the approach has usually been very different to ours. Many proposals center their study on server operation, modelling the behaviour and habits of the users that access the server in order to provide them the best possible service [5], [6]. Others have taken a more user-centric perspective but have focused on application-level operation [7], [8] or the characteristics of the downloaded content [9], [10]. Finally, other works have characterized specific types of web applications, usually with the intention of being able to classify their traffic [11], [12].

In our case, we consider our research from the client's point of view in the sense that we study the connections between the client and multiple servers through Internet and we do not have any information about the relationships between those servers or the content they host. However, even if we work from a client-side perspective, we only consider data that can be captured directly from the network rather than being "inside" the client monitoring its operation. All in all what we offer is a thorough characterization of the set of connections (and by connections we are referring to bidirectional TCP flows) initiated by the client during the download of a webpage.

This TCP-level characterization is interesting because NetFlow-type records [13] are easy to collect and, specially when compared to full packet-level traces, store and process in any network. Having a good description of the connections involved in the download of a webpage can be very useful for multiple purposes. On one hand current users access multiple webpages in short periods of time, often concurrently thanks to tab-based browsers, so it is far from trivial to guess which webpages (or even how many different ones) a user visits. This characterization may allow the development of techniques that help identify each individual webpage download, offering

978-83-60810-58-3/$25.00

insight into the user's behaviour. On the other hand, nowadays multiple applications mask their traffic in order to pass it off as HTTP and avoid certain restrictions that network administrators may want to enforce. Characterizing normal webpage downloads could help in designing anomaly-based detection systems able to identify that kind of applications.

The study we present in this paper is focused on website landing pages (*i.e.* the page served when the user inputs the domain name of the website). The characteristics of landing pages can be very different to those of internal pages (*i.e* accesed via links from the landing page) of the same websites. However, we are studying a wide variety of landing pages from 1,000 websites of different popularity. We believe that this is a sample with enough diversity to be representative of the characteristics of most webpages.

The remaining of this paper is structured as follows: section II explains the methodology used to capture experimental data and gives an overview of said data; section III discusses the general characteristics of a webpage download from a TCP connection point of view; section IV presents some time-based metrics that describe when the connections that participate in the download are opened and closed; section V focuses on the accessed servers; and, finally, section VI concludes and presents future lines of work.

II. Data collection

In this section we describe the data set we are going to use for our analysis in the rest of the paper. The traffic captures that integrate it were made during the months from August to October 2013.

A. Website selection and measurement setup

In order to collect a representative sample of webpage loads, we have selected 1,000 sites from the top 100,000 websites of the Alexa global ranking [14]. We have chosen the 100 most popular websites, 300 websites selected randomly from the 100-1,000 most popular ones, another 300 from the 1,000-10,000 range and the last 300 from the 10,000-100,000 range. With this, we ensure that the most popular (and interesting) sites, like Google, Facebook, or Amazon are well represented in the sample while also collecting data from a wide variety of less popular sites from all around the world.

We have gathered our measures from a computer in the Public University of Navarra (Spain) network. This PC has a public IP address and runs Ubuntu Linux (version 13.04). We felt unnecessary to set more than one vantage point as the authors in [7] found few differences when collecting the traffic of the same websites from different locations around the world. In this PC we run an automated script which follows these steps for each webpage under study:

- Launch a network sniffer: Tcpdump [15].
- Open a web browser to the selected website. We have collected measurements for both Mozilla Firefox (version 22.0) and Google Chrome (version 29.0.1547) which are the most popular browsers for Linux systems and together are responsible for a big percentage of the global web traffic [16], [17]. Plug-ins such as Adobe Flash player were installed in order to ensure that websites render properly but, aside from that, we use clean installations of both browsers with no ad or pop-up blockers.
- Wait for two minutes. Although webpages usually load in a few seconds [3], we capture traffic while the browser is idle for longer in order to study data transfers that happen even after the webpage has been fully rendered (when, for example, refreshing dynamic content).
- Close the web browser and close Tcpdump (we leave a small guard interval before closing Tcpdump in order to capture the ending of the pending connections).

We have repeated this procedure gathering twenty captures of each webpage download in pcap format (ten for each browser). We also captured one additional 10 minute long load for each page and browser in order to study flow end times as we will explain in section IV.

B. Preprocessing

In order to obtain connection records from these packet traces, we use Argus [18]. Argus is an open-source audit tool that is able to generate connection reports with the same features (and more) than NetFlow/IPFIX. In particular, aside from the classic TCP/IP 5-tuple (IP addresses, ports and protocol) we consider: timestamps (start and end), total packets, total bytes, application-level bytes (upload and download) and TCP state at the end of the capture. As we said previously, we always consider bidirectional TCP connections. Additionally, we store the first 1,000 bytes of the upstream application data of every connection from which we will extract some HTTP header fields. As we are only interested in web traffic we select connections originated in our PC and with destination ports 80 and 443 (HTTP and HTTPS protocols). However, we also extract DNS information from the pcap traces: we consider all the different server IP addresses accessed during the load of the webpage and, by studying the DNS query responses captured, we obtain a list of related domain names and authoritative nameservers for each IP address. This DNS information will allow us to better understand the part each connection plays in the load of the webpage.

We parse the HTTP data captured for each connection so we are able to extract the name of the accessed server, the URI of the first element requested and the HTTP method. With this information we identify the *root connection* of the webpage load. The root connection uses the GET method and requests the server root ("/") of a host with the same name as the website name. We then label connections according to their *origin*: those connections whose server name is related to (*i.e.* contains) the site name are classified as *shared name connections* and, from the rest of connections, we distinguish between *same origin* and *other origin* connections by checking if the domain name of the related server comes from the same authoritative nameserver as the root connection's or not. If the root connection carries HTTPS traffic it is impossible to identify it by checking user data. In this case, the root

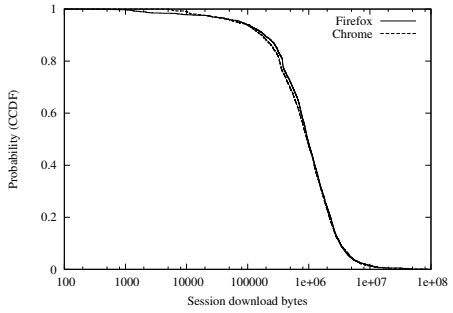


Fig. 1: CCDF of total downloaded bytes in a webpage load

connection will be the first flow opened to an IP address whose related domain name is the name of the website.

If we are unable to identify a root connection (i.e. there is no connection opened to an IP address related to the domain name of the website) or if we are able to identify it but it carries no application data we discard the capture as a failed webpage load. For the sake of simplicity in processing the data, we only consider websites that loaded successfully at every captured attempt. With this we reduce the list of considered websites from 1,000 to 912 resulting in a total of 18,240 flow records.

When comparing the total downloaded bytes for the same websites we observed discrepancies between the two web browsers. In particular, around 100KB of additional content were downloaded by Google Chrome in each capture. We discovered that this browser opens some HTTPS connections to Google servers for different purposes. Some of these connections are related to the webpage (for example, services like google translate or adsense) but others are automatic connections that are part of the browser operation and happen at fixed intervals from the start of the process. We decided to not consider these connections because, as we said previously, they are related to the browser behaviour and not to the particular websites. We also did not consider some connections opened by Firefox to Mozilla servers.

Now, figure 1 shows the empirical complementary cumulative distribution function (CCDF) for the total bytes of downloaded application data in every one of the captures of 120 seconds. This represents the total size of the different elements of each webpage. The distributions are very similar for both browsers as the effect of the browser in the elements downloaded from the webpage should be minimum.

III. General characterization

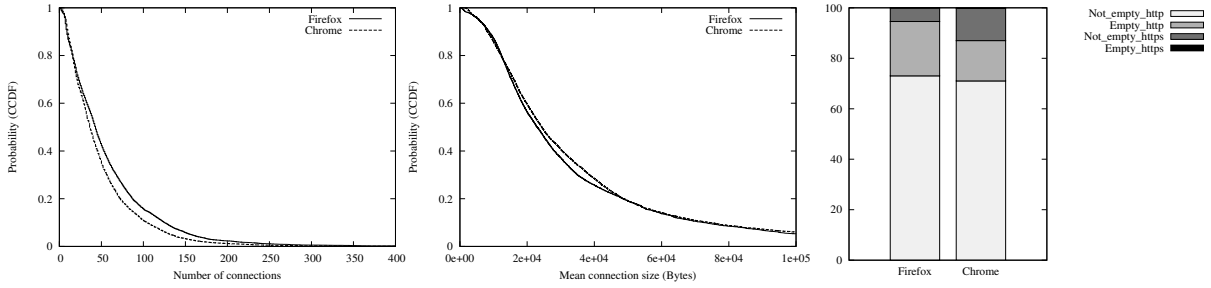
We start by providing some general connection-based metrics of webpage download. As it happened with figure 1 we have aggregated data from all our captures in order to plot the different graphs in this section. Because of this, each of the samples we use to calculate the empirical CCDFs corresponds to a different traffic capture (multiple downloads of the same webpages are present but there are the same number of them for every webpage so they are evenly represented).

Figure 2a shows the empirical complementary cumulative distribution function of the number of TCP connections initiated by the client during each webpage download. The distributions are, again quite similar: for both browsers the median is close to 40 connections while the 10 percentile sits around 10 connections (which means that for 90% of the studied webpages, at least 10 connections were used in the download). However, the tails of both distributions are long and, as we consider downloads with more connections, Firefox starts opening more of them (the 90 percentile is at 104 connections for Google Chrome and 125 for Firefox). As the total bytes downloaded by both browsers are very similar (fig. 1) this means that on average, Chrome connections are slightly bigger and more elements of the webpage are grouped in each of them. In any case, in figure 2b we can see that the difference in average connection size is small.

However, if we look at the individual connections (and, specially, the smallest ones) we do find some differences between the browsers' behaviour. In figure 2c we show the percentages of HTTPS and empty connections. These percentages have been calculated by aggregating all flows from all traffic captures of each web browser. As we can see, the percentage of normal HTTP connections is similar for both browsers however, the rest of connections are divided differently. Google Chrome has a higher percentage of HTTPS connections. By studying the servers this connections were established with, we realized that a lot of them were small connections related to services Google provides through Chrome to help the navigation process like, for example, Google Translate. However, the number of HTTPS connections initiated by Chrome to Twitter or Facebook servers is also higher.

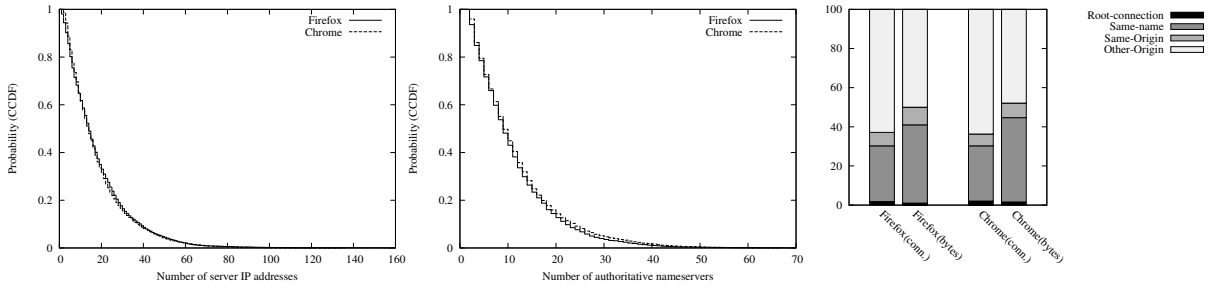
The other kind of connections we consider are empty connections. *Empty connections* are those that, having successfully completed their initial TCP three-way handshake, carry no application data. Most of this connections (around 95%) are also properly terminated although we also consider connections ended with a reset message or still open at the end of the capture. HTTPS empty connections are a very rare occurrence for both browsers but HTTP empty connections are a quite common occurrence specially for Firefox in whose traffic they represent around 20% of all connections. In most cases empty connections are a consequence of strategies employed by browsers with the objective of reducing webpage load times and, because of that, their number depends on the particular implementation of the program. Browsers may open multiple connections to a particular server before knowing how much content is going to be downloaded from it because it is faster to have connections prepared in case they can be used to download multiple elements simultaneously. This means that, in some cases, this connections are left unused. Additionally, browsers usually try to open a new connection if the server does not respond to a previous SYN packet within a time limit. This limit is low, again in order to reduce download times.

In figure 3 we focus on the different servers accessed during a webpage download. The differences between both browsers are minimal in this case as browser implementation cannot affect where the elements of a webpage are stored. Figure 3a shows the distribution for the number of different IP



(a) (b) (c)

Fig. 2: CCDF of the (a) number and (b) average size of connections. (c) Percentages of empty and HTTPS connections.



(a) (b) (c)

Fig. 3: CCDF of the (a) number of servers, (b) authoritative nameservers, and (c) percentage of connections by origin.

addresses accessed during each download. The distributions, with medians at 14 IP addresses but long tails, corroborate that modern websites download elements from multiple different servers. As we explained in section II-B we use authoritative nameservers in order to distinguish between different origins for web content. Figure 3b shows the distributions of the number of different authoritative nameservers seen on the DNS responses of the server names associated with each webpage download. Two authoritative nameservers are considered different if they have a different second-level domain name (third-level for some second-level domains like "co", *e.g.* "google.co.uk"). As we can see most webpages download content from servers of different origins (medians are 9 different origins).

Finally, in figure 3c we show the percentages of connections and bytes according to their origins. As in figure 2c we have considered every individual connection from the different traffic captures. The figure shows that, on average, more than 60% of the connections made during a webpage download are directed to third-party (other origin) servers. The most popular of them include, among others: analytics, social networking, image hosting, content distribution networks or video streaming. However, when considering downloaded bytes we can see that first-party content (root-connection, shared-name and same-origin servers) represents more than 50% of the total download suggesting that connections to first-party servers are bigger on average.

In table I we offer a summary of the different per-download metrics presented in this section (number of connections, mean connection size, number of accessed IP addresses, number of authoritative nameservers) providing the median and 10th and 90th percentiles for each of them.

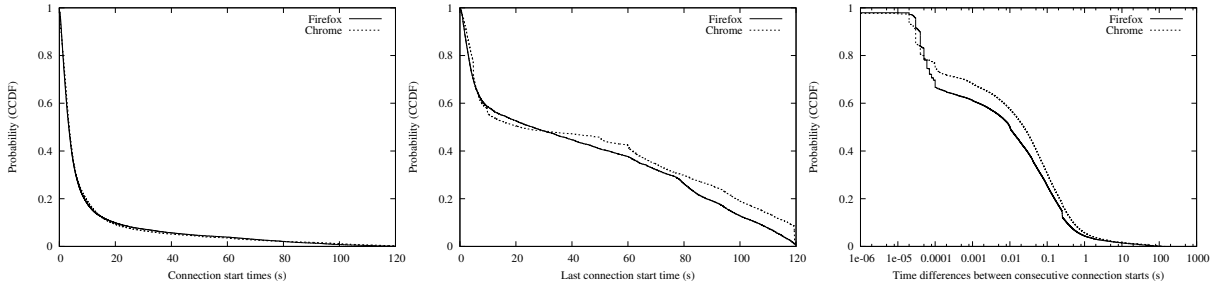
TABLE I: General characterization metrics

	Firefox			Chrome		
Metric	**P10**	**Median**	**P90**	**P10**	**Median**	**P90**
N. conn.	6	43	125	10	36	104
C. size (KB)	8.7	23.0	72.6	8.1	24.4	74.4
N. IPs	4	14	39	4	14	37
N. A.NS.	3	9	22	3	9	24

IV. Time metrics

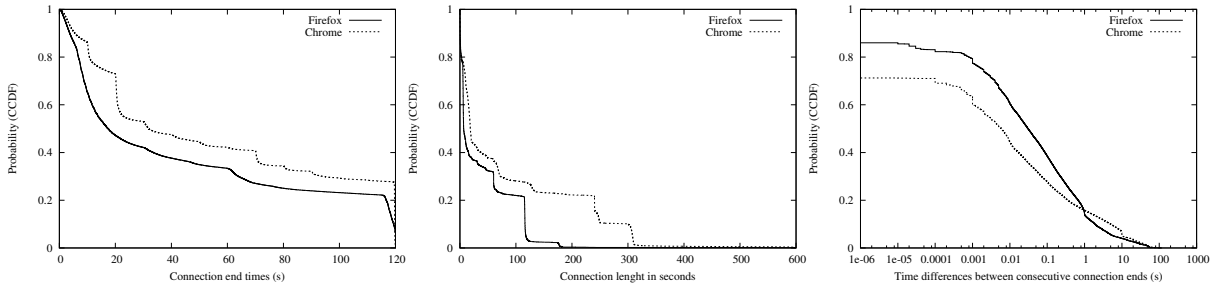
In this section we are going to discuss metrics related to the start and ending timestamps of the flows in the download of a webpage. Again, we consider the aggregated data of every download of every webpage as we want to characterize the behaviour of an average load rather than explore the differences between webpages. However, in this case, we are going to represent some parameters that are related to the individual connections rather than to the complete captures (as it was the case, for example, in figure 3a with the number of accessed servers per download). In all the figures in this section, the 0 seconds mark corresponds with the start timestamp of the first connection in each webpage download and the rest of connection timestamps in that download are calculated relative to it.

The first parameter that we are going to consider is connection start times which appears in figure 4a. In order to calculate



(a) (b) (c)

Fig. 4: CCDF of (a) connection start times, (b) last connection start time, and (c) time differences between connection starts.



(a) (b) (c)

Fig. 5: CCDFs of (a) connection end times, (b) connection lengths, and (c) time differences between connection ends.

these CCDFs we have considered the start timestamps of all the connections in every capture. As their timestamps are calculated relative to the first connection in each capture, we can compare them. We see that the majority of connections are opened during the first seconds of the download of a webpage (around 80% of connections in the first 10 seconds). This makes sense as this is an upper threshold for the time a webpage takes to load [3]. However, the tail of the distribution is long and a considerable amount of connections are opened later which suggests that even after the webpage is fully rendered some information is still exchanged. In figure 4b we represent the distributions of the start times of the last connection in each capture and we can confirm that even though most connections are opened in the first seconds, for the sizeable number of captures the last connections are opened much later. To shed some light about these late connections we used origins as defined in section II-B. We expected that the late connections could be specially related to third-party advertisment or analytics services. However, there is very little difference in the distributions of connection start times according to the different origins (aside from root connections happening always early in each capture). This suggests that the connections opened late do not only correspond to third-party content but also to dynamic content hosted in first-party servers.

In figure 4c we look at connection start times from a different perspective by representing the distribution of the time differences between consecutive connection starts. As expected, consecutive connections of the same webpage download are generally opened very close in time. Around 30% of consecutive connections are opened with less than a tenth of a millisecond between them (for smaller values, some precision/rounding artifacts appear) and only around 5% have their start times separated more than one second.

Lets now look at connection endings. Intuitively, we may think that connections are closed as soon as the elements they were opened to download are received by the client. However, this is not the case for a sizeable amount of them. Because servers and browsers implement persistent connections (all the HTTP connections observed were HTTP 1.1) some connections are kept open for longer in case they are needed for an additional download. As shown in figure 5a, both browsers keep more than 20% of the connections opened for all the duration of the capture and close them simultaneously as the browsers are closed. Nevertheless, studying the figure we saw that Firefox started ending some connections a few seconds before we closed the browser. We realized that Firefox implements a persistence timeout for HTTP connections that, by default, has a value of 115 seconds but can be tuned by the user via the configuration utility in about:config. The value of this timeout for Google Chrome is not documented.

In order to properly study the effects of these timeouts we captured one additional traffic trace of ten minutes for each webpage and browser. In figure 5b we show the distributions of connection length in seconds for these longer traces. For Firefox, the 115 seconds timeout is evident because most persistent connections have that length. The cause of the other step in the distribution (around 60 seconds) is more

difficult to pinpoint but should be related to a server timeout because it also appears in the Chrome distribution. For Google Chrome the default persistence timeout could be around 250-300 seconds. In any case, for both browsers around 60% of the connections are shorter than 20 seconds, either because they are not persistent or because of timeouts in the servers (Apache 2.0 has a default timeout of 15 seconds, for example).

Figure 5c is equivalent to 4c but this time we are representing time differences between connections that are closed consecutively. We can see that this time differences are usually very small. Because most connections are opened very close in time at the beginning of the download of the webpage, the effect of the persistence timeouts is very apparent when comparing their ending times. This, together with the fact that connections are ended immediately when the browsers are closed (note that in the figure around 20% of differences are 0), implies that the connections of the same webpage download usually end in almost simultaneous groups.

For table II we wanted to offer per-download metrics that describe the start and end timestamps of a webpage download. Again, we provide the median, 10th and 90th percentiles. The first metric we consider is the time of start of the last connection in the download (T. Last) as seen in figure 4b. However, this time may not, in some cases, represent the time interval during which most of the webpage is downloaded. In fact, if we consider the 10 and 90 percentiles we realize that we are basically covering the whole length of the capture. To give a better idea of the busiest time interval we consider the connections that carry the first 90% of the total data downloaded in each capture and provide the start time of the last of these connections (T. 90%). With this we eliminate the effect of small connections opened late in the download and give a better approximation of how close the connections of a webpage download are in time. For the time differences between flow starts and ends, in figures 4c and 5c we considered all flow pairs in every download but here we want to provide a per-download metric. We have calculated the median value for each capture and, in table II we show the median and percentiles of the distribution of said medians (T. Starts and T. Ends). As expected, the values of these two statistics are very low for almost every webpage download.

TABLE II: Time metrics (all values in seconds)

	Firefox			Chrome		
Metric	**P10**	**Median**	**P90**	**P10**	**Median**	**P90**
T. Last	1.62	25.96	105.82	2.34	21.29	117.10
T. 90%	0.61	2.67	12.51	0.76	3.36	17.85
T. Starts	0.00	0.01	0.12	0.00	0.02	0.13
T. Ends	0.00	0.04	0.32	0.00	0.01	0.22

V. SERVER METRICS

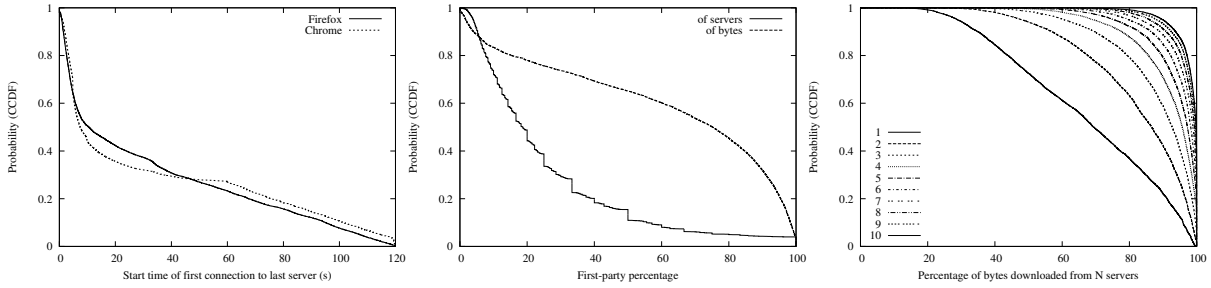
Of the classic 5-element tuple that traditionally describes an IP connection, the most interesting parameter when studying web traffic from the client point of view is the server IP address (protocol is always TCP, server port is 80 or 443 and client port is ephemeral and will not give us any information). Because of this, in this section we are going to center our study in the different servers accessed during the download of a webpage, considering that each of them corresponds to a different server IP address. Due to space constraints we only use Google Chrome data in most of the figures of this section. The results for Firefox are very similar because, aside from very specific browser services, the servers accessed during the download of a webpage should not vary depending on the used browser.

In section III we saw that multiple connections are opened to the same servers during each download. On the other hand, in figure 4b we realized that some of these connections happen very late in the captures. We wonder if these late connections are opened to servers that have already been accessed or to new ones. Figure 6a addresses this question by representing the CCDF of the start timestamp of the first connection to the last server that appears in each capture. The results are similar to the ones in figure 4b suggesting that a sizeable number of these late connections are opened to servers that have not been previously connected. However, as we said previously, these connections are usually small and the servers that host the main elements of the webpages are accessed in the first seconds of the download.

In section III we also gave information about the total number of different servers accessed in a download and the different authoritative nameservers related to them. However, it is clear that those servers play different roles in the webpage downloads depending on the elements they host or the services they provide so it should be interesting to study them individually. In figure 6b we show the distribution of the percentage of first-party servers for each webpage download (that is, the server the root connection is directed to, shared-name servers and same origin servers). We can see that, for most webpages, this percentage is low implying that most servers accessed during a download are third-party servers. However, the percentage of bytes downloaded from these first-party servers is much higher so, even if fewer first-party servers are accessed during a webpage download, they usually host a bigger part of the webpage content than the third-party ones.

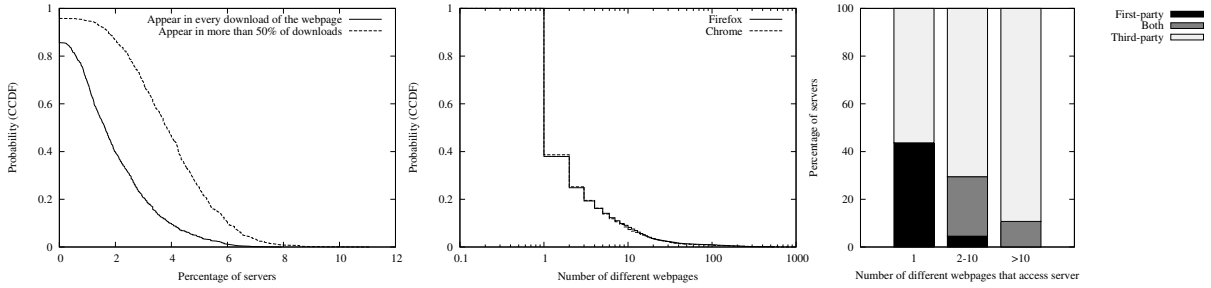
Another consequence of figure 6b is that, as we know, the content of a webpage is not equally distributed in the different servers accessed during the download. In figure 6c each CCDF represents the percentage of content downloaded from the "heaviest" server, the two heaviest servers and so on, of each capture. Even if many servers are accessed to download certain webpages, most of the content is hosted by few of them. For example, for 90% of the webpages, more than 80% of the downloaded content comes from only 5 different servers.

Until now, we have considered the servers in each webpage download independently. However, the content of many webpages is hosted in distribution networks or multiple hosts for load balancing purposes and because of this, connections to the same IP addresses are not always opened when accessing the same webpages. On the other hand, a lot of webpages use third-party services (*e.g.* analytics, image hosting, advertising, etc.) and, as a consequence, the same third party servers



(a) (b) (c)

Fig. 6: CCDFs of (a) time of first connection to last server, (b) first-party percentage of servers and bytes, and (c) percentage of bytes downloaded from heaviest servers.



(a) (b) (c)

Fig. 7: CCDFs of (a) percentage of servers that appear in all or half the captures of the same webpage, (b) number of different webpages that access each server, and (c) origins of shared servers.

are accessed when downloading different webpages. Figure 7 addresses these situations.

In figure 7a we consider all the IP addresses accessed in the ten captures we made for each website. We represent the CCDF of the percentage of servers that appear in every capture and that appear in, at least, half of the captures of each website. As we can see, both percentages are quite low. This means that most of the content in the webpages is dynamic or hosted dynamically and, because of that, the servers involved in the download change rapidly over time making very difficult to identify a particular IP address with a particular webpage.

Lets now compare all the servers accessed during the download of different webpages. Figure 7b represents the number of different webpages in whose downloads a particular IP address appears. For this figure we do not consider as different some of the webpages under study like, for example, amazon.com and amazon.co.uk, because they probably share an important part of their hosting infrastructure. A very high percentage (around 40%) of all the IP addresses appear in downloads of more than one webpage suggesting that a lot of services are shared between them. In figure 7c we have divided all the server IP addresses in three groups according to how many webpages download content from them. The servers of the first group only host content of one of the studied webpages, the ones in the second group host content of two to ten different webpages, and the servers in the third group host content of more than ten different webpages. For each group, we represent the percentage of first and third-party servers and servers that are both first and third-party depending on the webpage. As expected, most only first-party servers appear in downloads of just one webpage (the few of them that appear in 2-10 webpages are probably related to webpages that share a hosting platform like blogs). However, in a considerable amount of cases, servers that are considered first-party for a webpage appear in downloads of another one as third-party servers.

A consequence of all this is that even though IP addresses are an interesting parameter in order to group the connections of the same webpage download (as multiple connections are usually opened to the same servers) they should be used very carefully to relate different downloads of the same webpages. On one hand, the same content may be downloaded from different servers (or even the content itself may change in short periods of time). On the other, many servers are accessed by different webpages and even servers closely associated to a webpage by their name or their authoritative nameserver may host content for other webpages.

In table III we present some server metrics related to the variables we described in figure 6 providing, again, median values and percentiles. As it happened in the previous section, the time of the first connection to the last server (T. Last S.) is not very representative so we have calculated the time of appearance of the last server that, together with the ones that have already appeared is responsible for 90% of the total

download (T. 90 S.). This interval represents how long it takes for the servers that are responsible for the majority of the download to be contacted by the client. We also show the percentages of first-party servers and bytes (F.P. (S) and F.P. (B)) and the percentage of bytes downloaded from the heaviest 1, 5 and 10 servers.

TABLE III: Server metrics

Metric	Firefox P10	Firefox Median	Firefox P90	Chrome P10	Chrome Median	Chrome P90
T. Last S.	1.42	10.48	95.01	1.89	7.42	101.86
T. 90 S.	0.15	2.03	7.25	0.30	2.40	10.54
% F.P. (S)	5.26	20.0	78.57	5.26	18.75	55.56
% F.P. (B)	5.02	69.79	99.62	4.08	75.18	98.34
% 1 Serv.	30.00	61.14	95.63	34.97	69.42	95.97
% 5 Serv.	75.77	96.05	100	82.92	97.28	100
% 10 Serv.	90.43	99.72	100	93.75	99.81	100

VI. Conclusions and future lines of work

The increasing popularity of the web and the interesting complexity of its traffic have made it a popular topic of research for the scientific community. However, little work has been done in order to characterize web traffic at connection level, even though connection level data has the advantage of being easy to store and process in real time and can be collected even if the traffic is encrypted.

In this paper we have presented a thorough characterization of web traffic from the perspective of TCP connections. We have introduced various metrics that describe the set of connections involved in a webpage download focusing on their general characteristics, their distribution in time and the servers they reach. For each of these metrics we have shown its probability distribution and given some statistics to describe it. Taking into account the very limited nature of the information available in Netflow-type records, we have painted an accurate picture of the average webpage download.

We intend to apply the knowledge obtained in this work into designing a method able to identify webpage downloads in real traffic. We believe that by applying a combination of the metrics described in this paper we will be able to distinguish single webpage downloads in user traffic. A system based on this idea would be lightweight and able to process data in real time giving interesting information to network administrators about the behaviour of their users without accessing sensible information.

We also would like to explore the possibility of using these metrics in order to distinguish between websites of different categories (like social networks, news portals, etc.) or which offer different services (video streaming, games, etc.). As the normal range of the metrics is quite broad, the variability in them suggests that information about the characteristics of a webpage (indicative of the related website category or service provided) can be extracted from this kind of connection level data.

Other possible applications of this work are more related to network security as a thorough characterization as the one provided in this work can help with the tuning of anomaly-based detection systems. These systems that may be able to distinguish between normal web traffic and other applications that masquerade their traffic in order to avoid restrictions imposed by network administrators (or, in the case of malicious applications, in order to blend in and avoid detection).

References

[1] J. Charzinski, "Traffic properties, client side cachability and CDN usage of popular web sites," in *Measurement, Modelling, and Evaluation of Computing Systems and Dependability and Fault Tolerance*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2010, vol. 5987, pp. 136–150. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-12104-3_12

[2] H. Weinreich, H. Obendorf, E. Herder, and M. Mayer, "Off the beaten tracks: exploring three aspects of web navigation," in *Proceedings of the 15th international conference on World Wide Web*, ser. WWW '06. New York, NY, USA: ACM, 2006, pp. 133–142. [Online]. Available: http://doi.acm.org/10.1145/1135777.1135802

[3] S. Ihm and V. S. Pai, "Towards understanding modern web traffic," in *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, ser. IMC '11. New York, NY, USA: ACM, 2011, pp. 295–312. [Online]. Available: http://doi.acm.org/10.1145/2068816.2068845

[4] L. D. Catledge and J. E. Pitkow, "Characterizing browsing strategies in the world-wide web," *Computer Networks and ISDN Systems*, vol. 27, pp. 1065–1073, April 1995. [Online]. Available: http://dx.doi.org/10.1016/0169-7552(95)00043-7

[5] F. M. Facca and P. L. Lanzi, "Mining interesting knowledge from weblogs: A survey," *Data Knowl. Eng.*, vol. 53, no. 3, pp. 225–241, Jun. 2005. [Online]. Available: http://dx.doi.org/10.1016/j.datak.2004.08.001

[6] H. Liu and V. Kešelj, "Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests," *Data Knowl. Eng.*, vol. 61, no. 2, pp. 304–330, May 2007. [Online]. Available: http://dx.doi.org/10.1016/j.datak.2006.06.001

[7] M. Butkiewicz, H. V. Madhyastha, and V. Sekar, "Understanding website complexity: measurements, metrics, and implications," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, ser. IMC '11. New York, NY, USA: ACM, 2011, pp. 313–328. [Online]. Available: http://doi.acm.org/10.1145/2068816.2068846

[8] G. Xie, M. Iliofotou, T. Karagiannis, M. Faloutsos, and Y. Jin, "Resurf: Reconstructing web-surfing activity from network traffic," in *IFIP Networking Conference, 2013*, 2013, pp. 1–9.

[9] D. Fetterly, M. Manasse, M. Najork, and J. Wiener, "A large-scale study of the evolution of web pages," in *Proceedings of the 12th International Conference on World Wide Web*, ser. WWW '03. New York, NY, USA: ACM, 2003, pp. 669–678. [Online]. Available: http://doi.acm.org/10.1145/775152.775246

[10] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Characteristics of youtube network traffic at a campus network: measurements, models, and implications," *Computer Networks*, vol. 53, no. 4, pp. 501–514, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2008.09.022

[11] D. Schatzmann, W. Mühlbauer, T. Spyropoulos, and X. Dimitropoulos, "Digging into HTTPS: flow-based classification of webmail traffic," in *Proceedings of the 10th Conference on Internet Measurement (IMC '10)*. New York, NY, USA: ACM, 2010, pp. 322–327. [Online]. Available: http://doi.acm.org/10.1145/1879141.1879184

[12] F. Schneider, S. Agarwal, T. Alpcan, and A. Feldmann, "The new web: Characterizing AJAX traffic," in *Passive and Active Network Measurement*, ser. Lecture Notes in Computer Science. Springer, 2008, vol. 4979, pp. 31–40. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-79232-1_4

[13] E. B. Claise, "RFC 5101: Specification of the IPFIX protocol for the exchange of IP traffic flow information," Jan. 2008.

[14] "Alexa top 1,000,000 sites," Aug. 2013, http://www.alexa.com/.

[15] "Tcpdump & libpcap," Sep. 2013, http://www.tcpdump.org/.

[16] "Statcounter.com," http://gs.statcounter.com/.

[17] "w3counter.com," http://www.w3counter.com/.

[18] "Argus: Audit Record Generation and Usage System," http://www.qosient.com/argus/.

3rd International Conference on Wireless Sensor Networks

A FEW years ago, the applications of WSN were rather an interesting example than a powerful technology. Nowadays, this technology attracts still more and more scientific audience. Theoretical works from the past, where WSN principles were investigated, grew into attention-grabbing applications practically integrated by this time in a real life. It could be said, that countless application fields, from military to healthcare, are already covered by WSN. Together with this technology expansion, still new and new tasks and interesting problems are arising. Simultaneously, such application actions stimulate the progress of WSN theory that at the same time unlocks new application possibilities. The typical examples are developments within the "Internet-of-Things" field as well as advancements in eHealth domain with WBAN IEEE 802.15.6 standard progress.

Wireless sensor networks, as the spatially distributed networks consisted of a number of relatively simple, low-cost, low-power components interconnected mutually, provide quite wide application portfolio for different branches of economy. As the main examples could be mentioned military, industry, transport, agriculture and healthcare. However, in the near future, even stronger expansion of WSN application assortment is expected. In order to make this expansion possible, it is necessary to continually work on the solving of typical questions/problems related to the WSN development, e.g. standardization of communication protocols; the lack of energy-efficient power sources; the development of new ultra-low-power microelectronic components; etc.

An integration of WSN within the public data networks as well as within the domains where confidential and private data are processed (e.g. E-Health) brings along problems related to the ethical and legal questions too. Therefore, the terms as social safety or ethical safety should not be neglected.

The problematic of WSN is one of actual activities getting to the fore in the European Research Area since the issue of sensor networks was covered through "IoT" in FP7 program and strong continual extension is planned to be included also in Horizon 2020 program, especially in sections such Smart Transport; Health; Climate Action covered under Societal Challenges Pillar.

It is therefore essential to create an experience-sharing platform for scientific researchers and experts from research institutes, SMEs and companies who work in WSN domain where they can exchange some relevant skills and experiences as well as discuss upcoming trends and new ideas from this field. Moreover, the conference should also serve a function of a kind of networking platform facilitating interconnectivity between participants in case of a future collaboration.

Topics

Original contributions, not currently under review to another journal or conference, are solicited in relevant areas including, but not limited to, the following:

Development of sensor nodes and networks

- Sensor Circuits and Sensor devices – HW
- Applications and Programming of Sensor Network – SW
- Architectures, Protocols and Algorithms of Sensor Network
- Modeling and Simulation of WSN behavior
- Operating systems

Problems dealt in the process of WSN development

- Distributed data processing
- Communication/Standardization of communication protocols
- Time synchronization of sensor network components
- Distribution and auto-localization of sensor network components
- WSN life-time/energy requirements/energy harvesting
- Reliability, Services, QoS and Fault Tolerance in Sensor Networks
- Security and Monitoring of Sensor Networks
- Legal and ethical aspects related to the integration of sensor networks

Applications of WSN

- Military
- Health-care
- Environment monitoring
- Transportation & Infrastructure
- Precision agriculture
- Industry application
- Security systems and Surveillance
- Home automation
- Entertainment – integration of WSN into the social networks
- Other interesting applications

Event Chairs

Hodoň, Michal, University of Žilina, Slovakia
Kapitulík, Ján, University of Žilina, Slovakia
Micek, Juraj, University of Žilina, Slovakia
Ševcik, Peter, University of Žilina, Slovakia

Program Committee

Al-Anbuky, Adnan, Auckland University of Technology, New Zealand

Baranov, Alexander, Russian State University of Aviation Technology, Russia

Dadarlat, Vasile-Teodor, Univiversita Tehnica Cluj-Napoca, Romania

Diviš, Zdenek, VŠB-TU Ostrava, Czech Republic

Elmahdy, Hesham N., Cairo University, Egypt

Fouchal, Hacene, University of Reims Champagne-Ardenne, France

Giusti, Alessandro, CyRIC - Cyprus Research and Innovation Center

Grzenda, Maciej, Orange Labs Poland and Warsaw University of Technology, Poland

Gu, Yu, National Institute of Informatics, Japan

Hudik, Martin, University of Zilina

Husár, Peter, Technische Universität Ilmenau, Germany

Jin, Jiong, Swinburne University of Technology, Melbourne

Jurecka, Matus, University of Žilina, Slovakia

Kafetzoglou, Stella, National Technical University of Athens, Greece

Karastoyanov, Dimitar, Bulgarian Academy of Sciences, Bulgaria

Karpiš, Ondrej, University of Žilina, Slovakia

Kochláň, Michal, University of Žilina, Slovakia

Laqua, Daniel, Technische Universität Ilmenau, Germany

Li, Qinxue, Guangdong University of Petrochemical Technology

Monov, Vladimir V., Bulgarian Academy of Sciences, Bulgaria

Ohashi, Masayoshi, Advanced Telecomunications Research Institute International / Fukuoka University, Japan

Púchyová, Jana, University of Žilina, Slovakia

Ramadan, Rabie, Cairo University, Egypt

Scholz, Bernhard, The University of Sydney, Australia

Segal, Michael, Ben-Gurion University of the Negev, Israel

Shaaban, Eman, Ain-Shams university, Egypt

Shu, Lei, Guangdong University of Petrochemical Technology, China

Smirnov, Alexander, Linux-WSN, Linux Based Wireless Sensor Networks, Russia

Staub, Thomas, Data Fusion Research Center (DFRC) AG, Switzerland

Stojmenovic, Ivan, University of Ottawa, Canada

Teslyuk, Vasyl, Lviv Polytechnic National University, Ukraine

Wang, Zhonglei, Karlsruhe Intitute of Technology, Germany

Xiao, Yang, The University of Alabama, United States

Zuo, Liyun, Guangdong University of Petrochemical Technology

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 987–992
DOI: 10.15439/2014F85
ACSIS, Vol. 2

Energy Harvesting for Wireless Sensor Networks Review

Saba Akbari
"MATI"– Russian State Technological University Orshanskaya 3, 121552 Moscow, Russia
Email: akbarisaba@gmail.com

Abstract— With the widespread use of wireless sensors, the management of their energy resources has become a topic of research. Wireless sensors usually use batteries as their power supply but in some applications battery replacement can be cumbersome and require considerable amount of time which can affect the process being monitored. It is possible to harvest energy from the sources in nature for wireless sensors. In this article, a review of current alternative energy sources has been demonstrated in order to address the feasibility of their integration with wireless sensor networks.

I. Introduction

Fire and toxic gas leakage may have consequences resulting in pecuniary loss or fatality. Monitoring of hazardous gases is one of the areas where wireless sensors have been used [1]. Systems based on wires have some disadvantages as being dependent on power supply, high maintenance costs, sometimes electrical power supply cut off occurrence, and long deployment time. The study of forest fires has been also considered as an important issue and since it may take place in some areas of difficult access, the development of wireless sensor technology can be applied in this field as well [2],[3]. Wildfire which occurs in the countryside or wilderness area is a topic which is not yet completely resolved [4]. The satellite technology and aircrafts can track such incidents. However, there are some advantages and disadvantages associated with each method. The deployment costs for using satellites are high [1] and due to the orbital limitations not all satellites can monitor a fire event continuously [5]. Limitations of using aircrafts can arise in some areas which are prone to low altitude clouds and this can cause more obstacles in terms of visibility [5]. Wireless sensors have become an interesting topic to tackle this issue [6],[7]. But this technology also has its own challenges. The high power consumption of combustible gas detection systems is a restricting factor which needs to be considered [8],[9]. Sensors designed for this purpose and having on board batteries cannot have a long time life cycle [1] and moreover in some applications like structural health monitoring of critical infrastructures and buildings, it is difficult to replace or recharge batteries [10]. As mentioned in [11], some sensor networks like BAN systems require low energy levels. Table 1, 2 and 3 illustrate some off- the- shell components used in sensor node design [12]. As it can be seen from Table 1, the current consumption of the gas sensors is higher in comparison with other components used in the sensor module. The operation time of the autonomous wireless sensor device is currently limited by the batteries capacity which is about 3000 mAh, 8000 mAh and 15000 mAh for the AA, C and D batteries types, respectively. Taking into account the high power consumption of sensor nodes, it is necessary to optimize the use of battery. Therefore, demands on response time require that the measurement of gas concentration for example of combustible gases be done no more than every 120 s [13]. Nature provides us with variety of energy sources which can be harvested and implemented for wireless sensor systems [14]. Some various sources which can provide the necessary amount of power include solar, piezoelectricity, thermal, wind, and radio frequency. A compilation of various energy harvesting sources is given in Table 4 [15].

II. HARVESTING METHODS

A. Solar Energy

Typically, a sensor node which uses energy from nature, consists of several blocks such as energy harvesting module, the storage unit such as a supercapacitor, sensing element, microcontroller and a transmitter as shown in Figure 1.

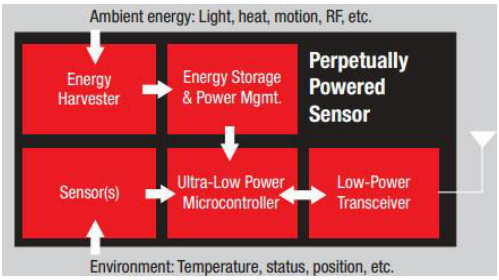


Fig. 1 Wireless sensor network with an energy harvesting module [16]

Photovoltaic is the technology that generates direct current from semiconductors when they are illuminated by photons. Most of the solar cells today are made of crystalline silicon (c-Si). Silicon has been the most widely used material in the production of photovoltaic modules. The mono- crystal, multi-crystal, micro-crystal and amorphous are the most widely used forms of silicon in the production of photovoltaic cells [17].

TABLE I.
POWER CONSUMPTION OF SOME OFF – THE - SHELF GAS SENSORS[12]

TYPE	COMPANY	POWER (mW)
DTK-2 Catalytic Sensor	NTC-IGD, Russia	120 mW
SGS-21XX Semiconductor Sensor	Delta, Russia	200 mW
TGS2610 Semiconductor Sensor	FIGARO, Japan	280 mW
NAP-66A Catalytic Sensor	Nemoto, Japan	360 mW
MQ-4 Semiconductor Sensor	Hanwei Electronics, China	750 mW

Nanotechnology based semiconductors including nanowires, nanocrystals, nanorods and nanodots have been opening a new perspective for scientists to design solar cells [18].

TABLE II.
CURRENT CONSUMPTION OF SOME OFF – THE - SHELF TRANSCEIVERS[12]

TYPE	COMPANY	CURRENT CONSUMPTION (mA)
CC2500 Transciever	Texas Instruments	Tx: 21.2 mA @ 0 dBm Rx: 13.3 mA
CC2430 Trasnceiver	Texas Instruments	Tx: 27 mA, Rx: 25 mA
ETRX35x Transceiver	Telegesis	Tx: 31 mA @ +3 dBm, Rx: 25 mA @ 12 MHz clock speed
TR1000 Transceiver	RF Monolithics	Tx: 12 mA @ 0 dBm, Rx: 3.1 mA
JN5148-001-M00/03 Transceiver Sensor	Jennic	Tx: 15 mA @ +2.5 dBm, Rx: 17.5 mA

Solar energy can generate enough amount of power necessary for wireless sensors [19]. A solar panel is used as an energy harvesting source for the wireless sensor in [20].

Solar energy harvesting provides direct DC voltage and therefore additional circuit rectifications are not required. This type of energy scavenging is free of emissions since it does not produce contaminants or bypass products that are harmful to the environment [21].

One of the limitations of solar energy is its dependency on solar radiation, which can be degraded in areas where enough sunlight is not available. Solar energy is characterized by having a varying nature [22]. The other issue associated with solar cells and for which some research is ongoing is the cooling of solar cells and recovering heat [23].

TABLE III.
POWER CONSUMPTION OF SOME OFF – THE - SHELF COMPONENTS USED IN SENSOR NODE[12]

TYPE	COMPANY	CURRENT CONSUMPTION (mA)
MSP430F247 Microcontroller	Texas Instruments	Active mode: 321 uA @ 3 V / 1 MHz Low power mode: 1 uA @ 3V / 32768 Hz
ATmega168P Microcontroller	Atmel	Active mode: 1.8 mA @ 3 V / 4 MHz Power-save mode: 0.9 uA @ 3 V / 32 kHz
ATxmega32A4 Microcontroller	Atmel	Active mode: 1.1 mA @ 3 V / 2 MHz Power-save mode: 0.7 uA @ 3 V / 32 kHz
ADuC824 Microcontroller	Analog Devices	Active mode: 3 mA @ 3 V / 1.5 MHz Power-down mode: 20 uA @ 3 V / 32 kHz

B. Piezoelectricity

Piezoelectricity stems from the Greek word “piezo” for pressure and the word “electric” for electricity. The main advantage of piezoelectric materials as shown in Figure 2 is the high amount of voltage they can provide. Some materials which have piezoelectric effect are quartz, soft and hard lead ziconate titane piezoceramics (PZT-5H and PZT5A), barium titanate (BaTiO3) and polyvinylidene fluride (PVDF)[24].

In the piezoelectric effect the usable output voltage can be obtained directly from the material and there is no need for applying multistage post processing for generating the desired amount of voltage [25]. Piezoelectric materials require dynamic forces in order to retain the output voltage and a notable drawback of piezoelectric sensors is their inability to respond to static loads [26].

C. Radio Energy

The possibility of harvesting RF energy, from ambient, enables wireless charging of a sensor node [27].

Having a transmitter set, one can harvest energy from radio waves and the advantage of this alternative is the fact that the scavenging mechanism can be flexible and it is

TABLE IV
SOME ENERGY HARVESTING SOURCES [15],[16]

Energy Source	Conditions	Performance
Solar	Outdoors	7500 μW/ cm^2
Solar	Indoors	100 μW/ cm^2
Vibration	1m/s^2	100 μW/ cm^3
RF	WiFi	0.001 μW/ cm^2
RF	GSM	0.1 μW/ cm^2
Thermal	$\Delta T = 5\,^{\circ}C$	60μW/ cm^2

possible to control the amount of transferred energy by making it continuous i.e. on regular intervals or the amount of radiated energy can be adjusted according to the requirements of the relevant application [28].

It is necessary to note that, according to the Friis equation, the amount of power collected at the receiver side is not equal to the exact value sent by transmitter [29].

D. Thermal Energy

Thermal energy produces electricity when there is a temperature difference (Seebeck effect).

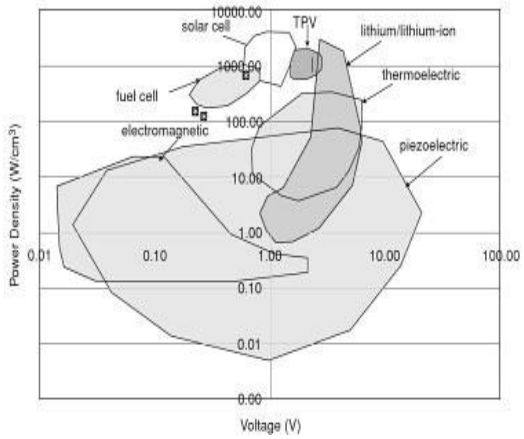


Fig. 2. The voltage generated by the piezoelectric material itself is greater than the value produced by some other harvesting sources [25]

Most thermoelectric generators consist of n type and p type an array of p- and n- type semiconductors in series. Heat from a hot source reaches the hot shoe and is conducted through to the semiconductors. The hot shoe has a high thermal conductivity and as well as a high electrical conductivity. Electrons are released from n type semiconductor and establish a current through p type semiconductor. Cold shoes transfer the current to the electric load [30].

Commercially available thermoelectric generators require a temperature difference of 10 – 200 ºC [31].

This type of energy is able to provide supply DC power continuously but temperature differentials can be difficult to generate in enclosed environments [32].

D. Wind Energy

Among the various forms of existing renewable energy sources, wind energy is one of the remarkable sources in the macro scale for generating electricity. Countries like Germany, Spain and Denmark have provided a significant amount of their required electricity by using this type of energy.

Having evaluated the given possibilities in high power electronics, its application can be also surveyed for miniaturized devices.

In the study presented in [33], different wind turbine blade configurations (two, three and six blade propellers) within a wind speed interval of 3.5 to 7 m/s were tested. The propeller with six blades output the maximum power. The authors propose to add a solar panel to the harvesting system in case the amount of energy provided by wind energy will not suffice.

During the analysis of alternative sources for wireless sensor networks, an experiment to survey the feasibility of using wind energy as a complementary system for wireless sensor systems was performed. In this experiment, two types of blades were tested. The measured wind speed was 3.7 m/s and 4.5 m/s. In the first case, the length of each blade was 4.35 cm and in the second one, it corresponded to 8.85 cm.

Figure 3.a shows power generated by the wind turbine as a function of load. The peak power for the propeller with smaller blades at the speed of 4.5 m/s is 7.2 mW.

At 3.7 m/s, the peak value of power for the first type of blade is 5.5 mW and for the second type of blade, the maximum amount of power at the speed of 4.5 m/s is approximately 5.5 mW (Figure 3.a). As the rotor of the DC motor starts rotating, the AC current flows in the circuit which consists of motor windings and the load.

As a result of this, self induction takes place. This effect can be seen in Figure 3 a and b for the first type of propeller when the load resistance is less than 500 Ohm and in the meantime the experimental points and the approximation do not converge. The experiment indicates that a microturbine based on a DC motor can be represented as a voltage source and an internal resistance according to the Thevenin theorem. Two factors which have caused less amount of power in our experiment in comparison with [33] are the wind speed and motor. It can be also observed from Figure 3.a that the larger propeller provides maximum output at smaller values of load (~60 ohm) and the smaller one reaches its power peak values at higher loads (~250 ohm), i.e. depending on the load value, the necessary blade can be chosen.

This model is valid only in cases when the induced current is greater than the self induction current. Wind is a free energy that nature provides us and there is no limit in using it. It produces no polluting emissions of any kind [34]. Wind energy in some environments cannot be considered as a primary source for electricity generation due to its intermittent behavior which causes instability to the power system [34].

III. APPLICATIONS

Alternative sources can be used along with batteries as a complementary system or in case of existence of enough energy from the environment, they can be employed as primary power sources. Pipeline monitoring in industrial complexes is a remarkable example where wireless sensors can be deployed [35]. It is quite plausible to employ energy harvesting techniques as primary or secondary power

supplies for the sensors in this scenario. In case of battery replacement, the harvesting technologies increase the possibility of continuous operation of the system and consequently the process will not be stopped.

Bridge health monitoring is another example discussed in [36]. In the Jindo bridge project, eight sensors out of 70 are equipped with solar panels. The charging process of the solar system is working well except one node which is not receiving enough sunlight due to its location. It is suggested that in the next deployment, either a more sensitive solar panel can be used or another type of energy harvesting may be considered [37].

IV. ENERGY STORAGE

At the moment, batteries are still the most common way to provide energy for low electronic devices. The specifications of Li – ion and thin film batteries are shown in Table 5.

The majority of wireless sensor networks are powered by batteries. As mentioned at the beginning of the article, the batteries used in wireless sensor networks have a finite lifetime which requires replacement or recharging.

Primary cells which cannot be recharged may cause a huge amount of work associated with replacement if used in a large network [38] like in the case of pipeline monitoring. Secondary batteries are rechargeable but they have a lower energy density than primary batteries [38].

The combination of an energy harvesting scheme with a rechargeable battery or with another storage system such as a thin film rechargeable battery or a supercapacitor can be implemented for wireless sensor networks. Figure 4 shows a comparison of several storage schemes. It is assumed that the volume of energy storage scheme is 1 cm^3. If the energy consumption is 100μW then the operation time of the primary battery lasts just for only a few months. The combination of an energy harvester with a rechargeable battery ensures a long term operation of the system.

Developing this strategy maybe necessary since sometimes the output power from the energy harvesting due to their intermittent nature may not be enough to compensate for the current consumption at receive and transmit periods. This scenario is illustrated in Figure 5.

V. CONCLUSION

An overview of the alternative energy sources was presented in this review article and the experiment using wind energy demonstrated the potentiality of this source to power low electronic devices. It is necessary to note that systems for harvesting and storing energies shall be designed based on the combination of two or more sources of alternative energy. For instance, for those electronic systems performing outdoor, solar and wind energies can be considered as sources for powering the devices or in the case of low power electronic devices located indoor, radio waves and thermal energy can be designated.

The hybrid energy sources increase the probability of having uninterruptible power supply because each source can compensate some fluctuation of energy of its counterpart transmitted to the powered device.

In addition to the conventional applications of energy harvesting, this technology can be used in the concept of “Smart Homes” as well as the management and powering of real objects in the Internet of Things [40].

In order to optimize the harvesting process using this type of energy, the following points can be considered in the future research:

1. Implement a micro wind turbine model. This defines a system where one can enter an input parameter such as wind speed and blade radius and based on that the output power can be obtained.

2. Develop hybrid models in order to feed the sensor simultaneously from more than one source. As a result, more energy can be stored for the node and this leads to storing more energy as well as providing a more realistic battery – less operation of the overall system

3. In order to ensure a continuous operation of the sensor,

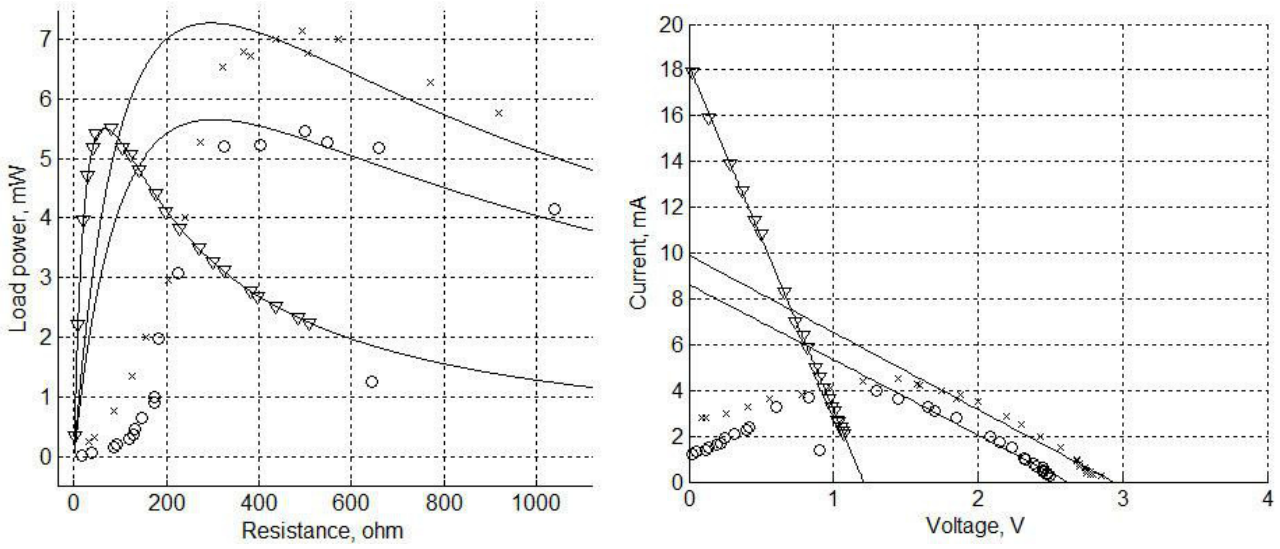


Fig.3. a) Power versus resistance for two types of propellers b) Current versus voltage for the same propellers x - Propeller with blade size of 4.35 cm at 4.5 m/s o - Propeller with blade size of 4.35 cm at 3.7 m/s ▽ - Propeller with blade size of 8.85 cm at 4.5 m/s - Linear approximation

a combination of an energy storage scheme and the harvesting system may output a steady operation of the sensor node.

4. In terms of applications, a system composed of solar cell and micro wind turbine can be designated for pipeline monitoring since replacing batteries in these areas can take a long time and stop the process. In some areas which have enough sunlight and wind, this model can be applied to provide a continuous operation of sensors.

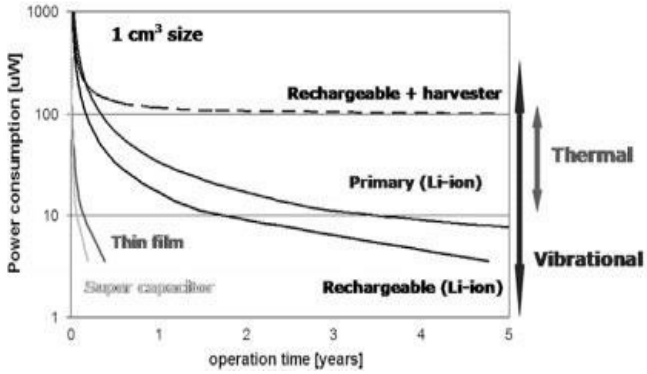


Fig. 4. A comparison of the operation time versus power consumption of various storage systems [39]

5. Thermal energy harvesting has also gained popularity for powering microelctronic devices. This technology can be also used in pipeline monitoring as a source to power sensors in which case it is feasible to use the temperature difference between the pipes and surrounding environment.

The possibility of harvesting energy from nature is fascinating but the cost issues of deploying such a technology shall be considered as well.

A scientific approach to implementing this feature for low electronic applications has to evaluate it and decide if energy scavenging proves better results in both technical and economical aspects in comparison with other technologies.

The research which have been carrying out in different microelectronics fields such as transmitter, microcontroller, storage schemes along with the advancements in energy harvesting technology will demonstrate more compelling results in the future.

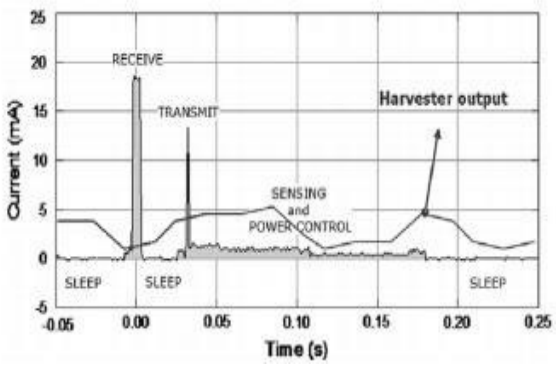


Fig. 5. In some cases the amount of harvested energy is not enough to provide the necessary power for the transceiver module [39]

ACKNOWLEDGEMENTS

The author would like to express the deep gratitude to Prof. Alexander Baranov from "MATI"– Russian State Technological University for the help in the manuscript preparation and useful discussion. This work was supported in the frame work of Russian Federal Program Grant No. 14.577.21.0022.

References

[1] A. Somov, A. Baranov, A. Savkin, D. Spirjakin, A. Spirjakin, R. Passerone Development of wireless sensor network for combustible gas monitoring Sensors and Actuators A: Physical, vol. 171, pp. 398 – 405, 2011. DOI: 10.1016/j.sna.2011.07.016.

[2] V.G. Gasull1, D.F. Larios1, J. Barbancho, C. León1 and M.S. Obaidat "A Wildfire Prediction Based on Fuzzy Inference System for Wireless Sensor Networks", in E-Business and Telecommunications, vol. 314. Mohammad S. Obaidat, José L. Sevillano, Joaquim Filipe , Ed. Berlin Heidelberg: Springer, 2012, pp 43-59, DOI: 10.1007/978-3-642-35755-8_4.

[3] A. Somov, D. Spirjakin, M. Ivanov, I. Khromushin, R. Passerone, A. Baranov, and A. Savkin "Combustible Gases and Early Fire Detection: an Autonomous System for Wireless Sensor Networks: an Autonomous System for Wireless Sensor Networks", Proceedings of the e-Energy 2010 - 1st Int'l Conf. on Energy-Efficient Computing and Networking. pp. 85 -93 . DOI: 10.1145/1791314.1791327.

[4] A. Volokitina, M. Sofronov, and T. Sofronova, "Topical scientific and practical issues of wildland fire problem," Mitigation and Adaptation Strategies for Global Change journal (Springer Netherlands), vol. 13, no. 7, pp. 661-674, August 2008. DOI: 10.1007/s11027-007-9120-7.

[5] A. Enis Cetin, D. Akers, I. Aydin, N. Dogan, O. Günay, B. Ugur Toreyin "Using Surveillance Systems for Wildfire Detection", www.firefighternation.com/article/wildland-urban-interface/ using-surveillance-systems-wildfire-detection [Jun. 5, 2013].

[6] I. Yoon, D. K. Noh, D. Lee, R. Teguh, T. Teguh and H. Shin, "Reliable Wildfire Monitoring with Sparsely Deployed Wireless

TABLE V.
CHARACTERISTICS OF BATTERIES AND SUPERCAPACITORS [39],[41],[42],[43],[44],[45],[46]

	Battery				Supercapacitor
	Li - ion	**Li – Po**	**Nickel Cadmium**	**Thin film**	
Operating voltage (V)	3 – 3.70	3.7	1.25	3.70	1.25
Energy density (W h/l)	435	300-450	~110 - 150	<50	6
Specific energy (W h/kg)	211	150-200	40 - 60	<1	1.5
Self-discharge rate (%/month)	5-10 at 25 'C	1 -2 at 25'C	20 – 30 at 25' C	0.1 – 1 at 20'C <2%* at 25'C	100
Cycle life (cycles)	2000	>1200	500 - 2000	>1000	>10.000
Temperature range (C)	-20/50	-40 to 60	-20 to 60	-20/+70	-40/+65

* According to Ref. [46], a battery with a 2μm-thick cathode, can output the mentioned value.

Sensor Networks, " in Proc. IEEEAINA, pp. 460-466, 2012, DOI: 10.1109/AINA.2012.107.

[7] Y. Li, Z. Wang, and Y. Song, "Wireless sensors network design for wildfire monitoring," in Proc. of 6th IEEE World Congress on Intelligent Control and Automation, Jun. 2006, pp. 109 – 113, DOI: 10.1109/WCICA.2006.1712372.

[8] S. So and G. Wysocki, "Ultra efficient laser spectroscopic trace-gas sensors for sensor networks and portable chemical analysis," Innovation Forum 2009, Princeton school of engineering and applied science, May, 2009.

[9] Andrey Somov, Alexander Baranov, Denis Spirjakin, Andrey Spirjakin, Vladimir Sleptsov, Roberto Passerone, "Deployment and evaluation of a wireless sensor network for methane leak detection". Sensors and Actuators A: Physical Physical- vol. 202, pp. 217–225, November 2013, DOI: 10.1016/j.sna.2012.11.047.

[10] Seah, W.K.G., Zhi Ang Eu, Tan, H. "Wireless sensor networks powered by ambient energy harvesting (WSN-HEAP)-Survey and challenges," in Wireless VITAE 2009, May 2009, pp. 1-5, DOI: 10.1109/WIRELESSVITAE.2009.5172411.

[11] J. Puchyova, M. Kochlan, M. Hodon "Development of Special Smartphone-Based Body Area Network: Energy Requirements" in Proc. Federated Conference on Computer Science and Information Systems (FedCSIS), 2013, pp. 895 – 900.

[12] A. Somov, A. Baranov, A. Savkin, M. Ivanov, L. Calliari, R. Passerone, E. Karpov, and A. Suchkov, "Energy-Aware Gas Sensing Using Wireless Sensor Networks", in Wireless Sensor Networks Series: Lecture Notes in Computer Science, vol. 7158. Gian Pietro Picco, Wendi Heinzelman, Ed. Berlin Heidelberg: Springer, 2012, pp.245-260, DOI: 10.1007/978-3-642-28169-3_16.

[13] British Standard Institution. Electrical apparatus for the detection of combustible gases in domestic premises. Test methods and performance requirements. 2000.

[14] J. Paulo and P.D. Gaspar "Review and Future Trend of Energy Harvesting Methods for Portable Medical Devices" Proceedings of the World Congress on Engineering, pp. 909-914, 2010.

[15] C. Mathna, T.O Donnell, R.V. Martinez – Catala, J. Rohan, and B.O Flynn, "Energy scavenging for long – term deployable wireless sensor networks", Elsevier: Talanta, vol. 75, no.3, pp. 613 – 623, 2008, DOI: 10.1016/j.talanta.2007.12.021.

[16] Murugavel Raju, Mark Grazier, Texas Instrument, http://www.ti.com.cn/cn/lit/wp/slyy018a/slyy018a.pdf [Apr. 2010].

[17] P. Jayarama Reddy. Solar Power Generation: Technology, New Concepts & Policy. Leiden, The Netherlands: CRC Press/Balkema, 2012, pp. 2-3.

[18] A. R. Jha "Solar Cell Technology and Applications",Boca Raton, FL: CRC Press, 2009 p.158.

[19] D. Dondi, A. Bertacchini, L. Larcher, P. Pavan, D. Brunelli, and L. Benini, "A solar energy harvesting circuit for low power applications," in Proc. IEEE International Conf. on Sustainable Energy Technologies (ICSET), Nov. 2008, pp. 945-949, DOI: 10.1109/ICSET.2008.4747143.

[20] M. Kochlan, P. Sevcik, "Supercapacitor power unit for an event-driven wireless sensor node" in Proc. Federated Conference on Computer Science and Information Systems (FedCSIS), 2012, pp. 791 - 796

[21] Dongsheng Ma, Rajdeep Bondade, "Reconfigurable Switched-Capacitor Power Converters: Principles and Designs for Self-Powered Microsystems" New York: Springer, 2011, p. 3, DOI: 10.1007/978-1-4614-4187-8.

[22] Zekai Sen "Solar Energy Fundamentals and Modeling Techniques: Atmosphere, Environment, Climate Change and Renewable Energy", 2008, Springer, p. 36, DOI: 10.1007/978-1-84800-134-3.

[23] Z. Abdin, M.A.Alim, R.Saidur, M.R.Islam, W.Rashmi, S.Mekhilef, A.Wadi "Solar energy harvesting with the application of nanotechnology", Elsevier: Renewable and Sustainable Energy Reviews, vol. 26, 2013, p. 843, DOI: 10.1016/j.rser.2013.06.023.

[24] Stephen Beeby, Neil M. White "Energy Harvesting for Autonomous Systems", Norwood, MA: ,2010, pp 99-100.

[25] Alper Erturk, Daniel J. Inman "Piezoelectric Energy Harvesting", UK: John Wiley & Sons, 2011, pp.2-3.

[26] Waldemar Karwowski, William S. Marras "The Occupational Ergonomics Handbook", Boca Raton, FL: CRC Press, 1998, p. 570.

[27] Joshua R. Smith, "Range Scaling of Wirelessly Powered Sensor Systems" in Wirelessly Powered Sensor Networks and Computational RFID, Joshua R. Smith, Ed. New York: Springer, 2013, p.3, DOI: 10.1007/978-1-4419-6166-2.

[28] Harry Ostaffe, Jason Tollefson "Harvested RF Powers Remote Sensors,"[Dec.12, 2014] http://www.digikey.com/en-US/articles/techzone/2011/dec/harvested-rf-powers-remote-sensors.

[29] R. Vias, H. Nishimoto, M. Tentzeris, Y. Kawahara, T. Asami, "A Battery-Less, Energy Harvesting Device for Long Range Scavenging of Wireless Power from Terrestrial TV Broadcasts," IEEE 2012 IMS Digest, Montreal, Canada, June 2012, DOI: 10.1109/MWSYM.2012.6259708.

[30] Kenneth Kroos, Merle Potter, "Thermodynamics for Engineers", Stamford CT: Cengage Learning, 2014, p.472.

[31] V. Çağrı Güngör, Gerhard P. Hancke, " Industrial Wireless Sensor Networks: Applications, Protocols, and Standards", Boca Raton, FL: CRC Press, 2013, p.127.

[32] Tony Armstrong "Aircraft Structures Take Advantage of Energy Harvesting Implementations", http://www.eetimes.com/document.asp?doc_id=1278767, [May. 11, 2011].

[33] J.W. Park, H.Jo Jung, , H. Jo and B. F. Spencer, Jr. "Feasibility Study of Micro-Wind Turbines for Powering Wireless Sensors on a Cable-Stayed Bridge" MDPI: Energies, vol. 5, 2012, pp. 3450-3464. DOI:10.3390/en5093450

[34] Hermann-Josef Wagner, Jyotirmay Mathur "Introduction to Wind Energy Systems: Basics, Technology and Operation", Berlin Heidelberg: Springer, 2013, p.3, DOI: 10.1007/978-3-642-32976-0

[35] Huaping Yu, Mei Guo, "An effective oil and gas pipeline monitoring systems based on wireless sensor networks", International Conference on Information Security and Intelligence Control (ISIC),2012, pp. 178 – 181.

[36] M. Reyer, S. Hurlebaus, John Mander, Osman E. Ozbulut, "Design of a Wireless Sensor Network for Structural Health Monitoring of Bridges" Wireless Sensor Networks and Ecological Monitoring Smart Sensors, Measurement and Instrumentation" in Wireless Sensor Networks and Ecological Monitoring, v.3, Subhas C Mukhopadhyay, Joe Air Jiang, Ed. Berlin Heidelberg: Springer, 2013, pp 195-216, DOI: 10.1007/978-3-642-36365-8_8.

[37] Jiaonnong Cao and Xuefeng Liu, "Structural Health Monitoring Using Wireless Sensor Networks", in Mobile and Pervasive Computing in Construction Chimay J. Anumba, Xiangyu Wang, ,Ed. UK, 2012, John Wiley & Sons, p.224, DOI: 10.1002/9781118422281.ch11.

[38] Edgar H. Callaway, Jr., "Wireless Sensor Networks: Architectures and Protocols", Boca Raton, FL: CRC Press, 2004, p.140.

[39] R.J.M. Vullers, R. van Schaijka, I. Doms, C. Van Hoof, R. Mertens, "Micropower energy harvesting", Elsevier: Solid-State Electronics,vol. 53, 2009. pp. 684–693, DOI: 10.1016/j.sse.2008.12.011.

[40] D. Kelaidonis, A. Somov, V. Foteinos, G. Poulios, V. Stavroulaki, P. Vlacheas, P. Demestichas, A. Baranov, A. Biswas, and R. Giaffreda, "Virtualization and cognitive management of real world objects in the internet of things" in IEEE International Conference on Green Computing and Communications (GreenCom), 2012, pp. 187-194. DOI: 10.1109/GreenCom.2012.37.

[41] Roland Büchi. Radio Control with 2.4 GHz. Norderstedt: Books on Demand, 2014, p. 26.

[42] Pei Zheng, Lionel Ni. Smart Phone and Next Generation Mobile Computing. San Francisco CA: Morgan Kaufmann, 2006. p. 86,

[43] Petar J. Grbovic. Ultra Capacitors in Power Conversion Systems: Analysis, Modeling and Design in Theory and Practice. West Sussex, UK: Wiley – IEEE Press. 2014. p. 17.

[44] Frank R. Spellman Handbook of Water and Wastewater Treatment Plant Operations. Boca Raton, FL: CRC Press. 2013. p. 348.

[45] A. Manthiram. "Materials Aspects: An Overview" in Lithium Batteries: Science and Technology. G.A. Nazri, G. Pistoia, Ed. Springer US, 2003, p. 4, DOI: 10.1007/978-0-387-92675-9.

[46] N.J. Dudney "Solid-State Thin-Flm Rechargeable Batteries". Materials Science and Engineering B, vol. 116, pp. 245 – 249, Feb. 2005, DOI:10.1016/j.mseb.2004.05.045.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 993–1000
DOI: 10.15439/2014F443
ACSIS, Vol. 2

Lifetime and Reliability Evaluation Models based on the Nearest Closer Protocol in Wireless Sensor Networks

Ning Cao Russell Higgs Gregory M.P. O'Hare Rui Wu
CLARITY: The Centre for Sensor Web Technologies
University College Dublin, Belfield, Dublin 4, Ireland
{ning.cao, russell.higgs, gregory.ohare}@ucd.ie, rui.wu.1@ucdconnect.ie

***Abstract*—This paper aims to introduce some key parameters for the tracking application in wireless sensor networks. In this paper, the Nearest Closer protocol has been implemented in J-sim simulation platform, and consequently some useful trade-off analysis results among the density, reliability and lifetime have been obtained. Based on these results, two evaluation models are proposed.**

***Index Terms*—J-Sim; Nearest Closer; Evaluation Model**

I Introduction

A WIRELESS sensor network consists of a number of sensors deployed either randomly or in a pre-determined state in a given two or three dimensional space, thus forming a network of sensors. The sensors are designed to measure one or more physical quantities in the space, such as temperature or location. The sensors need to transmit this collected data to the end-user, who often will be outside of the space being measured, which could well be a dangerous environment. As the sensors concerned are wireless they are typically powered by a battery with a finite lifetime and power output, it may be impossible or impracticable to recharge or replace such batteries. Thus in a real wireless sensor network a number of parameters naturally need to be considered such as energy consumption, network lifetime and network throughput. The network may be assigned a routing protocol. There exist a lot of routing protocols, so to choose the suitable routing protocol has become a significant problem.

The routing protocols can be divided into three main types: flat protocols, location-based protocols and hierarchical protocols. Single-hop, Nearest Closer and LEACH are typical and basic routing protocols for these three types respectively.

We have integrated the Single-hop and LEACH [1] protocols into the simulation tool J-Sim and provided mathematical models for each protocol. This paper will focus on the multi-hop routing protocol.

The multi-hop protocol we realized in J-Sim is called Nearest Closer protocol. To implement this protocol each node has to know: its own position, the position of its neighbours within its transmission range and the position of the sink node. The main idea in the Nearest Closer protocol is that the distance between the receiver sensor node and the sink node is shorter than the distance between the transmitter sensor and the sink node. In addition, the transmitter sensor will transmit to its nearest neighbour that is closer to the sink node.

In this paper, the relationships among density, lifetime, and reliability will be investigated for the Nearest Closer protocol by simulating the tracking application. Based on the results of these parameters, two intelligent evaluation models will be proposed. This means that wireless sensor network users can predict lifetime and reliability directly. Thus sensor nodes can be deployed in such a network without further simulations.

The rest of this paper is structured as follows: Section 2 describes some details of J-Sim. Section 3 defines the evaluation parameters. Section 4 focuses on the experimental set-up information. Section 5 describes the results which have been obtained so far. Section 6 proposes the Lifetime model. Section 7 proposes the Reliability model. Section 8 concludes this paper.

II J-Sim

J-Sim [2] (formerly known as JavaSim) is an open-source, component-based compositional network simulation environment. The system is based on the IEEE 802.11 [3-4] implementation provided with J-Sim. IEEE 802.11 is the first wireless LAN (WLAN) standard proposed in 1997. J-Sim is implemented on top of the autonomous component architecture (ACA), components are the basic elements in this architecture and through these J-Sim implements the data transmission process. J-Sim provides a script interface that allows its integration with Tcl, and has been developed entirely in the Java platform. Java is a general purpose object-oriented computing language that is specifically designed to have as few implementation dependencies as possible.

This work selected J-Sim as its simulation tool for the following reasons:

The authors of J-Sim have performed detailed performance comparisons in simulating several typical WSN scenarios in J-Sim and ns-2. The simulation results indicate J-Sim and ns-2 incur comparable execution time, but the memory allocated to carry out simulation in J-Sim is at least two orders of magnitude lower than that in ns-2. As a result, while ns-2 often suffers from out-of-memory exceptions and was unable to carry out large-scale WSN simulations, the proposed WSN framework in J-Sim exhibits good scalability.

978-83-60810-58-3/$25.00

J-Sim models are easily reusable, so users can combine the components in the framework freely. J-Sim also provides a Graphical User Interface (GUI), which makes it easy to operate the simulation.

J-Sim is a Java-based platform. The Java-based sensors could be integrated with Java-based simulation tools in the future.

III. Evaluation Parameters

In this section, the evaluation parameters of Reliability Lifetime and Density will be defined for use in the following sections. This paper will focus on the parameters of Energy, Density, Lifetime and Reliability. Based on the results of experiments measuring these parameters, an intelligent evaluation model will eventually be constructed.

This work has completed power control over the radio components on the J-Sim simulation platform, which makes simulation of power consumption possible.

There is no need to assume that all sensors in a WSN measure the same phenomena; however, in this paper it is assumed that all sensors can transmit data about the phenomena they detect and can receive and retransmit data from other sensors about any of the phenomena detected in the system.

A. Reliability

In this work, experiments are conducted using the concept of reliability, defined by:

Reliability = the number of packets received by the sink node / the number of packets sent to the sink node

The advantages of using reliability are two-fold. Firstly, this definition can be used in the field and laboratory. In the field provided the sink node (end-user) knows how many cluster heads or sensors there are and the average number of packets of sensed data per sensor per time unit, then reliability can be estimated at any time during the lifetime of the network using the number of time units elapsed. Any difference between reliability measured in the field and laboratory could be used to detect how many packets of data are lost between sensors and cluster heads or the sink node (this would include counting sensors that have failed or have otherwise been lost to the system). Secondly, this simple definition of reliability can be used in a network where there are sensors of various types monitoring different things to produce a single simple measure of accuracy. In this situation our definition gives a systemic measure of accuracy for the network as a whole independent of the applications. This definition also makes it easy to analyze communication among the sensor nodes; in particular it makes the estimation of data collision in the wireless sensor network possible.

A frequently used definition for tracking accuracy is to set the tracking error equal to the Euclidean distance between the estimated and actual locations of the target. The chief advantage of our definition for the tracking application, that this paper is going to simulate, is that it is independent of the target position.

B. Lifetime

Network lifetime has become the key characteristic for evaluating sensor networks in an application-specific way. In particular the availability of nodes, and connectivity have been included in discussions on network lifetime. In fact, even quality of service measures can be reduced to lifetime considerations. Network lifetime is the time span from the deployment of the sensors to the instant when the network is considered non-functional. When a network should be considered non-functional is application specific. It could be, for example, the instant when the first sensor dies, a percentage of sensors die. Conserving sensor energy and increasing tracking accuracy are the two main goals for the research of target tracking applications in wireless sensor networks. Simulation is a common way to compare these two parameters. Before the real deployment of sensor nodes, users always perform some simulation tests for the tracking application. In the simulations, finding the trade-off point between energy consumption and tracking accuracy is one of the key questions to be addressed. In addition, evaluation analysis will be performed here among all the related parameters.

Sensors need to send packets to other sensors or sink node in a WSN. On the other hand, the sink node will receive packets from sensors. Thus, the definition for network lifetime we have taken in our experiments is the time when the last packet is received by the sink node. Last packet here means the sink node cannot receive any packet from sensors after receiving this packet.

C. Density

The number of sensors deployed in a fixed area will be taken as the Density parameter in this paper. Obviously as the number of sensors increases, so does the average number of sensors per square metre.

IV. Experimental Set-up

The simulated area for the experiments in the following sections is a 10 metre × 10 metre square with randomly deployed nodes. The sink node for this application is located in the middle of this square. One of the primary reasons for selecting this setup is to allow the results to be generalized to large areas by concatenation of networks similar to this. For example, a 50 metre × 50 metre square region could be configured using 25 instances of the setup used here in a 10 × 10 grid formation. All the points (in the figures) in the following section are the average value from at least 5 separate experiments.

In the following section, reliability and lifetime will be evaluated using density as a parameter. As the number of nodes increases in a fixed space, the efficiency of the sensor network may become better or worse. So in some simulations the researchers may care about the number of sensors they are using in the network. It is obvious that

some of the nodes will be out of power in a real wireless sensor network. Thus attention will move to the number of sensors that are alive.

V. RESULTS

A series of experiments were carried out with the number of sensors starting at 10, and increasing in increments of 10 to 300 sensors. The transmission radius for each sensor was fixed at 15 metres.

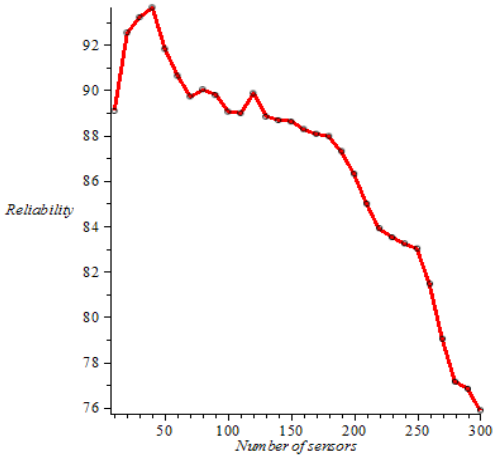


Figure 1. Density, Reliability relationship for NC.

In Fig. 1, the relationship between the number of sensors and Reliability is very clear. Reliability for this application increased as the number of sensors increased to 40, when it reached its highest value. It then essentially decreased as the number of sensors increased from 40 to 300, when it reached its lowest value. This may be explained by observing that as the number of nodes increases in the fixed space more sensors will join the data transmission process and consequently communication among the sensors will become more and more complex. Consequently dropped data due to data collision and latency cannot be ignored. We conclude that it is reasonable to expect reliability to decrease with density. In addition, the variance for Reliability is 23.3978 and the standard deviation for Reliability is 4.83713.

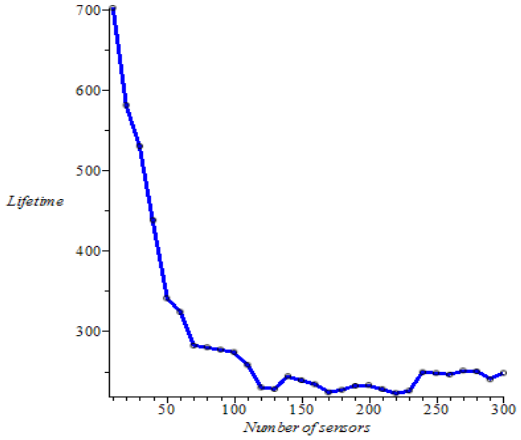


Figure 2. Density, Lifetime relationship for NC.

In Fig. 2, Lifetime reached its lowest value when the number of sensors equalled 220, whereas the highest value of Lifetime occurred when there were 10 sensors. When the number of sensors equalled 40, Reliability was 93.64%, its highest level, but with this number Lifetime is fairly short. This illustrates that it is possible for users to choose an optimum density value for this application depending on the Reliability and Lifetime required. In addition, the variance for Lifetime is 13602.64483 and the standard deviation for Lifetime is 116.63038.

VI. LIFETIME MODEL

Hop count is one way to measure the energy requirement of a routing task, thus using a constant metric per hop. However, if nodes can adjust their transmission power (knowing the location of their neighbours) the constant metric can be replaced by a power metric $u(d) = e + d^{\alpha}$ (or some variation of this) for some constants α and e that depend on the distance d between nodes. The value of e, which includes energy loss due to start up, collisions, retransmissions, and acknowledgements, is relatively significant, and protocols using any kind of periodic hello messages are extremely energy inefficient.

The basic idea for the NC protocol model to be constructed in this section is that the energy consumed by an average sensor in sending all its data to its nearest neighbour is of the form $w(s)E(d^2)$, where $E(d^2)$ is the expected value of the square of the distance between sensors and $w(s)$ is the average number (weight) of packets of data to be sent to its nearest neighbour. This figure also represents the average amount of energy to send one packet of data to the sink node.

Now in the NC protocol one possible parameter that could affect Lifetime is the number of paths or trees formed by the sensors. This is not the case as demonstrated by the following experiments, which are of independent interest. Using J-Sim it is difficult to extract the paths that sensors transmit along to the sink node. For this reason a program was written in the algebraic software package Magma [5] to not only compute the individual paths, but to marry them together to form the initial distinct trees.

Of course during the Lifetime of the network various sensors will die, normally from the centre outwards, and new trees will be formed. The mean number of trees corresponding to different numbers of sensors is given in Table 1 below. This mean number of trees was found by running the program 1,000 times for each given number of sensors. It could be seen from this table that the number of trees is almost independent of the number of sensors (excluding very small values) and is of the order of 2.55 trees.

TABLE I
MEAN NUMBER OF TREES

No. of sensors	Trees
10	2.537
20	2.527
30	2.561
40	2.559
50	2.531
60	2.550
70	2.542
80	2.534
90	2.553
100	2.549
110	2.568
120	2.578
130	2.555
140	2.543
150	2.574
160	2.550
170	2.557
180	2.552
190	2.575
200	2.582
210	2.574
220	2.600
230	2.571
240	2.573
250	2.569
260	2.551
270	2.575
280	2.543
290	2.563
300	2.562

The conclusion from Table 1 is that the mean number of trees is independent of the number of sensors and therefore can be discounted as having any effect on Reliability or Lifetime. The reason that on average two or three trees occur can be satisfactorily explained by considering the diagram below.

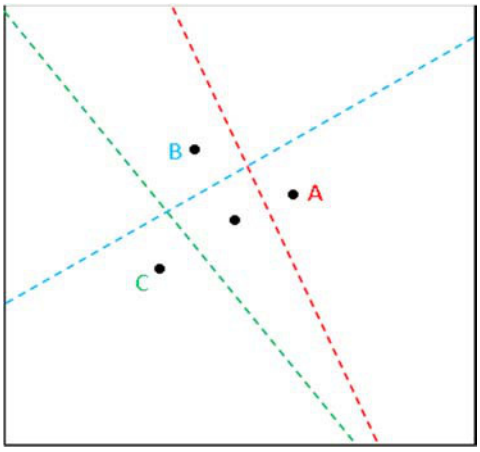


Figure 3. Typical Tree Zones

In Fig. 3 the points A, B and C represent the three nearest sensors to the sink node in the centre of the square amongst the randomly distributed sensors in the region. The perpendicular bisectors of the line segments joining A, B and C to the sink node have then been constructed. The next furthest sensor will with high probability be in one of the three zones containing A, B or C and its corresponding perpendicular bisector. Thus, in this diagram that three trees will probably result with A, B and C as the roots. Then, in 'zone' A it may be the case that further out in the tree that sensors cross over into zone B or C. It is fairly easy to draw diagrams where two or four trees will result, but only a very lop-sided distribution will result in just one tree if a reasonable number of sensors are being distributed.

For the NC protocol the relation 'is the closest neighbour to' is not symmetric, that is B can be the closest neighbour to A without A being the closest neighbour to B. Thus for this protocol it is necessary to find the closest neighbour to a sensor A, but with the stipulation that the neighbour is closer to the sink node than A. It is also the case that the sink node could be the closest neighbour to A, but the sink node is a fixed point and therefore does not fit in with the assumption that this work is dealing with a random distribution of sensors.

The expected distance d between a sensor node and its closest neighbour that is closer to the sink node (including the sink node itself) is given in the second column of the table and the expected distance of the square of the distance is given in the third column. Finally in the fourth column the average weight of the sensors is given, where each branch in the various trees count as weight one. This is then a measure of the number of packets of data an average sensor has to transmit to its nearest neighbour and it is also the average number of hops from the sensors to the sink node. The results of these simulations are tabulated in Table 2.

TABLE II
EXPECTED DISTANCES AND WEIGHTS, VARIABLE SENSORS

Sensors	$E(d)$	$E(d^2)$	$w(s)$
10	2.1945	6.0533	2.3526
20	1.5985	3.2761	3.3421
30	1.3223	2.2444	4.0821
40	1.1512	1.7023	4.7152
50	1.0264	1.3535	5.2218
60	0.9402	1.1350	5.7941
70	0.8672	0.9679	6.2249
80	0.8104	0.8450	6.6531
90	0.7657	0.7545	7.0685
100	0.7256	0.6759	7.4110
110	0.6926	0.6155	7.7859
120	0.6618	0.5634	8.1589
130	0.6360	0.5200	8.4701
140	0.6124	0.4817	8.7956
150	0.5910	0.4485	9.0999
160	0.5724	0.4210	9.4502
170	0.5553	0.3957	9.7218
180	0.5394	0.3734	9.9967
190	0.5240	0.3525	10.3059
200	0.5108	0.3346	10.5684
210	0.4983	0.3183	10.7823
220	0.4874	0.3046	11.0709
230	0.4759	0.2904	11.2847
240	0.4658	0.2785	11.5108
250	0.4562	0.2670	11.8269
260	0.4466	0.2559	12.0056
270	0.4393	0.2475	12.2536
280	0.4308	0.2378	12.4573
290	0.4231	0.2296	12.7054
300	0.4160	0.2216	12.9399

It should be noted that for those sensors whose nearest neighbour is the sink node that the distribution of the distance and distance squared to the sink node is slightly different to those between sensors, and the averages are always slightly higher than those given in Table 2, with the exception of the 10 sensor case when there is a significant difference.

Two results emanate from this table. The first is that $w(s)E(d)$ is nearly constant as it only ranges from 5.16 to 5.45 and is a measure of the average distance from a sensor to the sink node measured along the hop path. The second is that $w(s)E(d^2)$ decreases as the number of sensors increases, this shows that the energy expended by a sensor in receiving and sending received packets to its nearest neighbour decreases as the number of sensors increases. It also represents a measure of the average amount of energy needed to transmit one packet of data to the sink node. Two theories are proposed to explain why the lifetime decreases as the number of sensors increases.

A. *Early Termination*

To explain early termination it is useful to consider the situation when two trees occur and the nearest (root) sensors to the sink node. These two sensors have to receive and transmit all the sensor data to the sink node and therefore have long transmission times with one time slot per packet. A transmission break, which in our definition of lifetime is interpreted as the end of lifetime, could result when both of the root sensors do not possess sufficient energy to transmit their data to the sink node. In a sense this is like the situation for the LEACH protocol with the root sensors acting as fixed cluster heads and the lifetime being controlled by the lifetimes of those sensors. A transmission break could also result if there is a catastrophic data collision in which all the packets of data from the root sensors collide with each other at the sink node and from which the system cannot recover. It is highly probable in a large system that some data collisions will occur at the sink node.

B. *Data build up*

As explained above the tree root sensors have long transmission times to transmit their data in a number of slots. Each sensor can transmit and receive data, but cannot do both simultaneously, also in the slotted ALOHA [6-8] MAC design transmissions in a tree are sequenced. So the sensors immediately prior to a root sensor will receive more data than previously when the root sensor transmits, which will subsequently be transmitted to the root sensor and then the pattern will be repeated. The situation will become more complicated higher up the tree, but generally the amount of data to be transmitted by a sensor will increase with time in a non-linear manner. In particular the root sensors will have more and more packets of data to transmit as data builds up in the tree. So for a fixed number of sensors the amount of energy required for a root sensor to transmit its packets of data to the sink node will increase with time. Delays in the slotted ALOHA protocol will lead to data build up and data collisions will result in retransmissions.

If a varying number of sensors is considered then the amount of energy needed to send a frame of packets may increase with the number of sensors despite the fact that $w(s)E(d^2)$ decreases with the number of sensors. The energy required to transmit such a frame of data (in reality such a frame may not be transmitted in consecutive slots) will increase with the number of sensors if the frame length (the number of packets) exceeds the ratio between $E(d^2)$ for the corresponding number of sensors; this latter ratio is always less than 1.85 between successive values in Table 2 and approaches 1 as the number of sensors

increase, so that it is reasonable to expect that the corresponding ratio of frame lengths will exceed this. This effect is balanced by the fact that in a smaller system sensor will transmit and receive data more often than in a larger system. It is also of note that most energy is used in data transmission rather than reception, however, energy used for data reception could be significant when compared to energy use in a small system.

The conclusion is that $w(s)$ may not be the correct weight to be attached to an average sensor, but instead represents a lower bound for the weight.

The two theories can be related in that with data build up the root sensors may end up transmitting a large amount of data to the sink node using a lot of energy in the process. This could ultimately lead to root sensors having insufficient energy to transmit the frame or a catastrophic data collision, in either case this will lead to early termination of the system. In reality both theories proposed here contribute to the lifetime figures obtained with the latter theory difficult to quantify in a model because of the complexity of the trees.

Anyway lifetime in general can be modelled by only considering the roots of the normally two or three trees. The case of just one tree occurring is of interest for a small number of sensors, for in this scenario the system should run until the only root sensor dies naturally, at that time the system might interpret the death as a transmission break or the next sensor(s) out will take over the role of the root(s). The basic model constructed here is of the form

$$\frac{E(T)K}{nE(d^2)}$$

where n is the number of sensors, K is the total initial energy of a sensor and $E(T)$ is the expected number of trees. Here $n/E(T)$ represents the average weight per tree and since we know that $E(T)$ is nearly constant, the model should roughly resemble the curve $1/n$. Our model in particular assumes that the sensors are evenly distributed amongst the trees and disregards data build up, it should therefore represent an upper bound for the lifetime once the system is running at full capacity. This would imply that lifetime decreases monotonically as the number of sensors increases; however, this basic model assumes that the system runs at full capacity from the start and ignores idle slots in the slotted ALOHA design. Thus there is a base value below which lifetime will not fall, as the system will behave like a smaller system during its initial phase and there will also always be a number of idle slots, this base value will be taken to be the smallest lifetime value of 223 that occurs when $n = 220$. So the revised model is

$$(1 - \lambda(n))223 + \lambda(n)\frac{E(T)K}{nE(d^2)}$$

where $\lambda(n)$ is a proportionality measure with $\lambda(220) = 0$. Now from the lifetime data $\lambda(n)$ will be very close to 0 for all $n \geq 120$. Setting $\lambda(10) = 1$ yields the value $K = 16750$. Computation gives the following values for $\lambda(n)$:

TABLE III
EVALUATION OF THE PROPORTIONALITY MEASURE

n	10	20	30	40	50
$\lambda(n)$	1	0.846	0.741	0.529	0.292
n	60	70	80	90	100
$\lambda(n)$	0.250	0.146	0.141	0.133	0.125
n	110	120	130	140	150
$\lambda(n)$	0.085	0.017	0.012	0.051	0.038
n	160	170	180	190	200
$\lambda(n)$	0.027	0.002	0.009	0.021	0.024
n	210	220	230	240	250
$\lambda(n)$	0.012	0	0.007	0.062	0.059
n	260	270	280	290	300
$\lambda(n)$	0.055	0.064	0.065	0.040	0.059

The parameter $\lambda(x)$ can be piecewise approximated by

$$1.1438571428571 - 0.01501071428514x$$

for $10 \leq x \leq 70$, and

$$0.3954 - 0.0029514285714286x$$

for $80 \leq x \leq 130$ by fitting straight lines through the given values. An alternative that will yield better local approximations is to interpolate between the known values of λ. Thus if $n \leq x < n + 10$, then

$$\lambda(x) \cong \left(1 - \frac{x - n}{10}\right)\lambda(n) + \frac{x - n}{10}\lambda(n + 10).$$

The general equation obtained for $10 \leq x \leq 133$ is

$$(1 - \lambda(x))223 + \lambda(x)\frac{42712.5}{xE(d^2)}$$

and the value of $E(d^2)$ for x sensors can again either be interpolated between the known values in Table 2 or by using the fitted formula

$$\frac{4918141600.24}{6653x^2 + 68674970x + 125057800}$$

for $10 \leq x \leq 300$. Using the latter formula (which is a good fit) yields at worst an implicit equation for the lifetime for $10 \leq x \leq 130$.

Finally we are stating that lifetime is essentially constant for $x \geq 140$ with a value between 223 and 250 depending on the topology of the sensors.

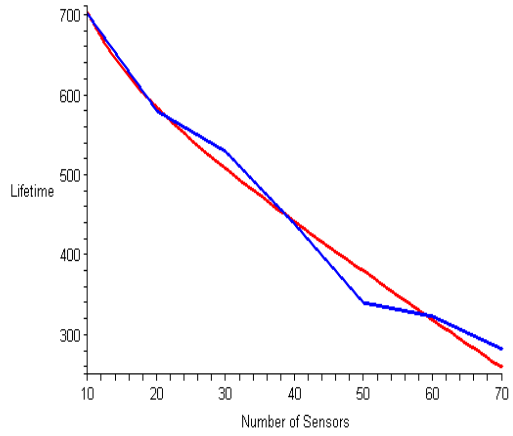


Figure 4 (a). Lifetime model (red), Lifetime (blue).

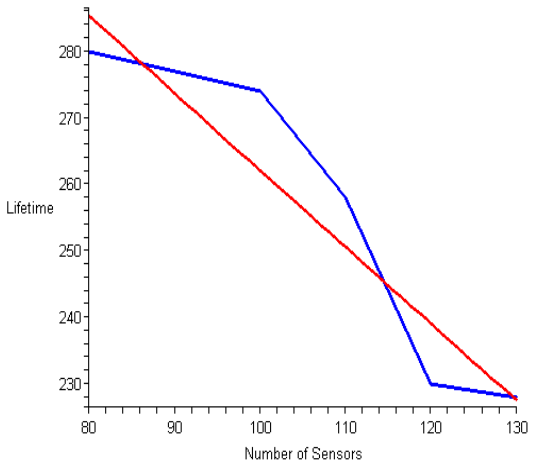


Figure 4 (b). Lifetime model (red), Lifetime (blue).

In both Fig 4 (a) and (b) the lifetime model is nearly a straight line and is clearly a better fit for ≤ 70 sensors.

VII. Reliability Model

In the following Table 1 Reliability (as a percentage) divided by Lifetime is listed for the 15 metre transmission radius case, which gives (100 times) Reliability per unit Lifetime $R(n)/L(n)$:

TABLE IV
Evaluation of Reliability per Unit Lifetime

n	10	20	30	40	50
$R(n)/L(n)$	0.1269	0.1593	0.1759	0.2138	0.2692
n	60	70	80	90	100
$R(n)/L(n)$	0.2797	0.3182	0.3215	0.3241	0.3250
n	110	120	130	140	150
$R(n)/L(n)$	0.3449	0.3911	0.3896	0.3634	0.3708
n	160	170	180	190	200
$R(n)/L(n)$	0.3773	0.3932	0.3875	0.3763	0.3704
n	210	220	230	240	250
$R(n)/L(n)$	0.3728	0.3763	0.3695	0.3343	0.3348
n	260	270	280	290	300
$R(n)/L(n)$	0.3311	0.3149	0.3086	0.3202	0.3060

In the first of these regions the function $R(n)/L(n)$ may be approximated by

$$0.12609 + 0.002226n$$

or equivalently since we have determined $L(n)$ in the previous section

$$R(n) = 0.12609 * L(n) + 0.002226 * n * L(n)$$

for $0 \leq n \leq 120$.

This indicates that 12.5% of the Lifetime figure represents a lower bound for Reliability, but that Reliability per unit Lifetime will increase by about 2% of the Lifetime as the number of sensors is increased by 10.

In the next region the rate $R(n)/L(n)$ reaches its highest level and stays constant at about 0.3779, so that Reliability is about 38% of the Lifetime figure for $130 \leq n \leq 210$.

In the last region the rate decreases, this can be satisfactorily explained by the fact that Lifetime may be considered to be constant in this region and thus increasing the number of sensors just increases the complexity of communication to the sink node and the rate can be approximated by

$$0.549282 - 0.000832666n$$

for $220 \leq n \leq 300$. So in this region 55% of the Lifetime is an upper bound for Reliability, but this decreases by about 0.8% for each extra 10 sensors.

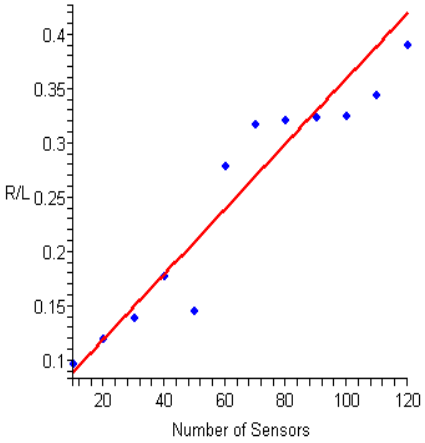


Figure 5 (a). R/L model (red), R/L data points (blue).

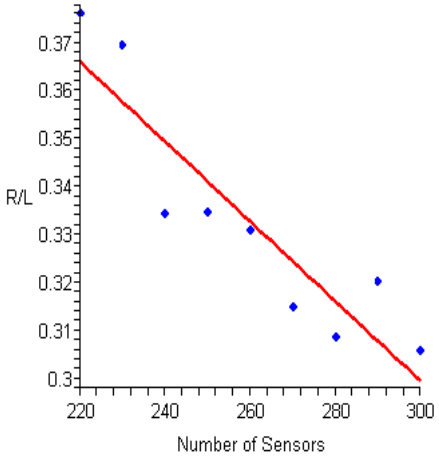


Figure 5 (b). R/L model (red), R/L data points (blue)

The model for R/L in Fig. 5 compared to the actual data point is fairly good with one exception of the point with 50 sensors, which either is genuinely exceptional or it may represent a data error.

VIII. Conclusions

In this paper, from a series of experiments using the J-Sim simulation tool several conclusions for NC were reached: (i) Reliability for this application increased as the number of sensors increased to 40, when it reached its highest value. It then essentially decreased as the number of sensors increased from 40 to 300. (ii) Lifetime decreased as the number of sensors increased to 130 and then it fluctuated slightly as the number of sensors increased to 300. Lifetime reached its lowest value when the number of sensors equalled 220, whereas the highest value of Lifetime occurred when there were 10 sensors.

Based on these simulation results, two evaluation models among the parameters of lifetime, reliability, and density have been proposed. Thus, wireless sensor network users can predict the lifetime and reliability directly. This means that sensor nodes can be deployed in such a network without further simulations.

Acknowledgment

This Research is supported by Science Foundation Ireland under grant 07/CE/1147.

References

[1] N. Cao, R. Higgs and G.M.P. O'Hare. "An Intelligent Evaluation Model Based on the LEACH Protocol in Wireless Sensor Networks", CyberC 2012, pp. 381-388, Sanya, China, 2012. http://dx.doi.org/10.1109/cyberc.2012.70

[2] A. Sobeih, W. Chen, J.C. Hou, L. Kung, N. Li, H. Lim, H. Tyan, and H. Zhang. "J-Sim: a simulation environment for wireless sensor networks", Proceedings of the 38th Annual Symposium on Simulation, pp. 175–187, IEEE Computer Society, Washington DC, 2005. http://dx.doi.org/10.1109/anss.2005.27

[3] G. Bianchi. "Performance analysis of the IEEE 802.11 distributed coordination function", IEEE Journal on Selected Areas in Communications, 18 (3), pp. 535–547, 2000. http://dx.doi.org/10.1109/49.840210

[4] S. Garg, M. Kappes. "An experimental study of throughput for UDP and VoIP traffic in IEEE 802.11b networks", Wireless Communications and Networking, 3, pp. 1748–1753, 2003. http://dx.doi.org/ 10.1109/wcnc.2003.1200651

[5] W. Bosma, J. Cannon, and C. Playoust, "The Magma algebra system", Journal of Symbolic Computation, 24 (3), pp. 235–265, 1997. DOI: 10.1006/jsco.1996.0125

[6] Y.C. Jenq. "On the stability of slotted ALOHA systems", IEEE transactions on Communications, 28 (11), pp. 1936–1939, 1980. http://dx.doi.org/10.1109/tcom.1980.1094610

[7] A.B. Carleial, M.E. Hellman, "Bistable behavior of ALOHA type systems", IEEE Transactions on Communications, 23 (4), pp. 401–410, 1975. http://dx.doi.org/10.1109/tcom.1975.1092823

[8] L.G. Roberts, "ALOHA packet system with and without slots and capture", Computer Communication Review, 5, pp. 28–42, ACM, 1975. http://dx.doi.org/10.1145/1024916.1024920

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1001–1007
DOI: 10.15439/2014F270
ACSIS, Vol. 2

Universal Synchronization Algorithm for Wireless Sensor Networks—"FUSA algorithm"

Michal Chovanec
University of Žilina
Faculty of Management Science and Informatics
Department of Technical Cybernetics
Univerzitná 8215/1, 010 26 Žilina, Slovakia
Email: michal.chovanec@fri.uniza.sk

Jana Púchyová, Martin Húdik, Michal Kochláň
University of Žilina
Faculty of Management Science and Informatics
Department of Technical Cybernetics
Univerzitná 8215/1, 010 26 Žilina, Slovakia
Email: {jana.puchyova, martin.hudik, michal.kochlan}@fri.uniza.sk

***Abstract*—Synchronization, in general, is the process of event coordination so that a system performs in unison. In wireless sensor networks (WSN), the network topology often changes dynamically in time. This brings technical hitches and challenges into time synchronization of the sensor nodes in the WSN. In such networks, data routing scheme is either data fusion, fusion of decisions or hybrid fusion. No matter, what data routing scheme is utilized, the time synchronization among the sensors is highly desirable. This paper presents an advanced synchronization algorithm for WSN. Hierarchical as well as decentralized network topology can be used with this algorithm. The designed algorithm is versatile, scalable and easy for implementation. At first, application area of the proposed synchronization algorithm is described. The detailed explanation of the synchronization algorithm at the node level and the network level is supported by the simulation results along with the implementation remarks.**

I. Introduction

HAVING wireless sensor networks (WSN), regarding their nature, they belong to the category of systems with a great measure of parallelism [1]. In order to effectively utilize the parallelism nature, to ensure real-time communication ability and to focus nodes' computational power to application-oriented algorithms, it is necessary to synchronize the sensor nodes. The time synchronization is crucial for any distributed system. In particular, WSN make extensive use of synchronized time in many contexts [1] (e.g. for data fusion, fusion of decisions, hybrid fusions, time division multiple access (TDMA) schedules, synchronized sleep periods, etc.).

This paper describes a new synchronization algorithm based on the fireflies synchronization process. The described algorithm is versatile regardless the network topology. This means, hierarchical networks with master nodes that control the synchronization process as well as fully distributed homogeneous-sensor-type networks are usable for the proposed synchronization algorithm. This new algorithm has been given the name *"Firefly-based Universal Synchronization Algorithm (FUSA)"*.

A. Application area

WSN represents an application area of a great potential. The increasing number of WSN deployment, recent advances in micro-electro mechanical systems (MEMS) and new energy harvesting possibilities make WSN very topical issue [2]. The WSN application potential as well as implementation potential of the proposed synchronization algorithm FUSA includes the following application areas:

- Health-care;
- Transportation;
- Military applications;
- Wearable electronics;
- Industrial applications;
- Intelligent automation/buildings;
- Monitoring of objects and environment (detection of floods, illegal logging, fires, etc.);
- and many other applications.

Health-care WSN applications are modern and interesting application area of WSN [3]. These applications mostly include wearable electronics such as intelligent watches, drug pumps, motion sensors monitoring elder people and those with diseases and disabilities so that they can be kept under the track of their vitality status without major limitations [4]. This modern application area and increasing popularity of health-care applications allow us meeting also WSN that monitor human body vital signs, track patients, monitor hospital environment and many more [5], [6], [7].

Transportation represents a significant portion (at least one third) of the national gross income in the developed countries [8]. Therefore, the need for efficient control of the traffic flows and intelligent monitoring of the traffic is in place. WSN help accomplish aims of the intelligent transportation systems such as traffic mirrors and intelligent traffic crossroads [9], traffic flow classification and quantification, intelligent car park systems and many more.

WSN concept originates in the military application field. Sensor network applications like sniper localization or battle-field monitoring are typical [10].

Wearable electronics include already mentioned health-care applications. However, large application field include sports equipments as well.

Industrial control systems that implement wireless technologies and sensors with actuators represent an advantage over the traditional distributed control systems [11], [12].

This work was supported by Tatra banka Foundation Slovakia

Mostly various diagnosing systems and production line control systems create the industrial application field of WSN.

The energy consumption of the whole society is a crucial problem in terms of the long-term sustainability of the environment. A big portion of the problems related to the environment connects to the energy consumption and energy requirements of the buildings [13], [14]. Intelligent control of the whole building ecosystems represents quite new and very interesting application field not only for the WSN [15], [16].

Property surveillance, object monitoring, environment monitoring (floods detection, fire detection, illegal logging detection, etc.) including protected areas monitoring is another significant application field of WSN [17], [18]. From functional point of view it is very important not to have only algorithms for monitoring - image recognition, voice recognition [19], but also the proper algorithm for increasing the function potential of the whole network [20].

Having in mind WSN for monitoring open area, it is necessary to consider the main characteristics of WSN [8]:

- The network is spread in outdoor terrain, where the energy resource is limited;
- The sensor and actuator nodes have limited energy storage, memory capacity and computational power;
- The network throughput and RF communication bandwidth are limited;
- The interference, path-loss and diffraction phenomena applies in RF communication;
- The environmental conditions as well as network attributes vary in time.

Synchronization algorithm demands are application specific. Typically, applications monitoring environment have the following characteristics [20]:

- Energy efficiency - time needed for synchronization, communication window length and active power modes should be are minimized;
- Scalability - usable for different number of nodes;
- Precision - the nodes are able to send the data in proper time;
- Synchronization time - the amount of time needed for synchronization should be as short as possible.

This paper assumes that the application field of the proposed synchronization algorithm is an application for monitoring forests in order to prevent the illegal wood logging situations. All simulations and application remarks are based on this assumption. However, this fact is not in contrary with the versatility of the proposed synchronization algorithm. The mentioned application field is for the illustrative and interpretation purposes.

B. Related work

Since the WSN applications are specific, not every synchronization algorithm is suitable for WSN purposes. The following paragraphs mention synchronization algorithms whose nature allows WSN utilization.

Cristian's Algorithm [21] and *Berkeley Algorithm* [22] are considered as essential synchronization algorithms and we will not discuss them. Computer networks very often use *Network Time Protocol (NTP)* [23] for time synchronization. However, the standard computer networks do not suffer from limited energy constraints.

Well known and very often used algorithms in WSN are *Reference Broadcast Synchronization (RBS)* [24] and *Timing Synchronization Protocol for Sensor Network (TPSN)* [25]. In RBS, the master node called the beacon node is used for synchronization. The synchronization of the whole network is performed from the beacon node that sends the reference broadcast towards one-hop-distant nodes from beacon node. Large networks with many sensor and/or actuator nodes are usually divided into smaller virtual networks called clusters. TPSN synchronization algorithm works with synchronization master as well. This master node is elected by all nodes. As soon as the mater node is elected, the spanning tree of the network is created. The children nodes are being synchronized by their parent node. In case any change in the network topology happens (e.g. a node becomes unavailable) a new master has to be elected again.

A reference point and the construction of the network tree is also used in *Tree Structured Referencing Time Synchronization (TSRT)* [26] and *Lightweight Tree-based Synchronization (LTS)* [27].

Other class of synchronization algorithms that use master node for synchronization or the group of master nodes contains *Time Diffusion Synchronization (TDP)* [28] and *ETSP* [29], which use both TPSN and RBS methodology. They switch between the TPSN and RBS based on the threshold value. The hierarchical structure of WSN is also used in [30], where the big accuracy of synchronized clock can be achieved, but only in simulation environment. Some of the algorithms use conditional probability estimation as well [31], [32], [33].

II. Synchronization Algorithm - FUSA

The synchronization algorithm proposed in this paper can be used in hierarchical as well as in network with fully distributed cooperation and coordination.

Since the communication subsystem in active mode consumes the significant, and in many cases the most, energy of all sensor node subsystems, minimizing active RF communication minimizes the energy consumption of the whole node too. The presented algorithm is based on the fireflies synchronization process - [34] (main idea), [35] (simple practical implementation), [36] (theory).

Each node periodically transmits synchronization packet (any data can be used). Let the basic time period be T. By using crystal-based clock, it is possible to set the period for each node precisely. However, each node starts at random time instant. This phenomena results in different timing phase start. Despite having crystal-based clock, small deviations in every clock source create deviations in time phase, thus getting all nodes out of the synchronization. This presents a problem and therefore synchronization algorithms are being used to suppress the unwanted effects.

Let's denote the phase for N nodes by $\phi_n \in < 0, 1 >$. Then, the maximum phase error in the network can be defined as:

$$\phi_{max}(t) = max_n(\phi_n(t)) - min_n(\phi_n(t)) \quad (1)$$

This definition can be represented as network synchronization quality.

The network is fully synchronized when $\phi_{max}(t) = 0$. After this point, all nodes are allowed to switch to a sleep mode and they wake up only for a short period of time. In discrete time domain, the best way how to divide the time period T is to divide it into D count of the same parts (frames) T_d, i.e. $T = D \cdot T_d$. For the reason of simple practical implementation, in simulations, the dividing factor D has been equal to $D = 128$. The nodes have transmitted the data in $1s$ time period, which implies $128Hz$ interrupt timer frequency. Each node is allowed to transmit data only in its time frame, while the rest of the time frames (assigned to other nodes) this node is expected to respect the radio silence (usually turns into the sleep mode).

Each node has a time counter which is incremented periodically in the timer interrupt routine. When it reaches the maximum, then the timer starts decrementing the counter value and at the same time the synchronization packet is transmitted. When the counter reaches the minimum value, the timer starts incrementing. The described process generates triangle wave output, which can be transformed into a phase represented as the sawtooth wave. The triangle (sawtooth) wave is important, as presented in [34], the function cannot be homogeneous (fe. time = time%128 will never work).

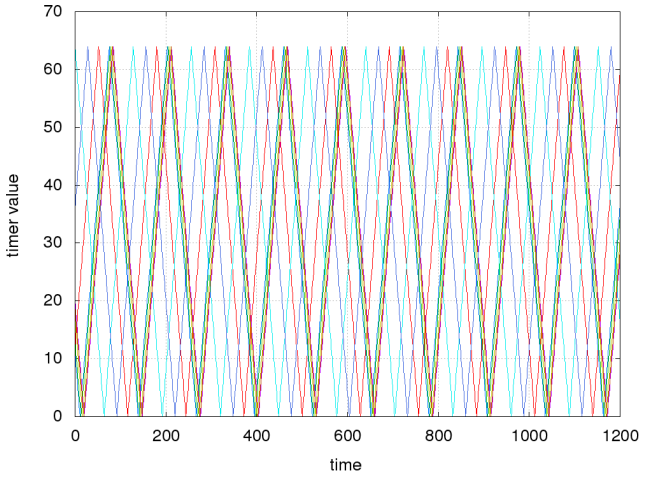


Fig. 1. Phases of nodes without synchronization

Simulations were performed with 64 nodes, in a $8x8$ grid network topology, where each node can see only four neighboring nodes. This grid network topology is called an *anuloid grid*. Fig. 1 and Fig. 2 demonstrate the network without synchronization - random initial phases. Fig. 1 represents phase values of first 8 nodes. Synchronization quality can be evaluated using (1) and is show in Fig. 2. It is obvious that the maximum phase error oscillates around the maximum phase value (128) and the average phase is in the center (64).

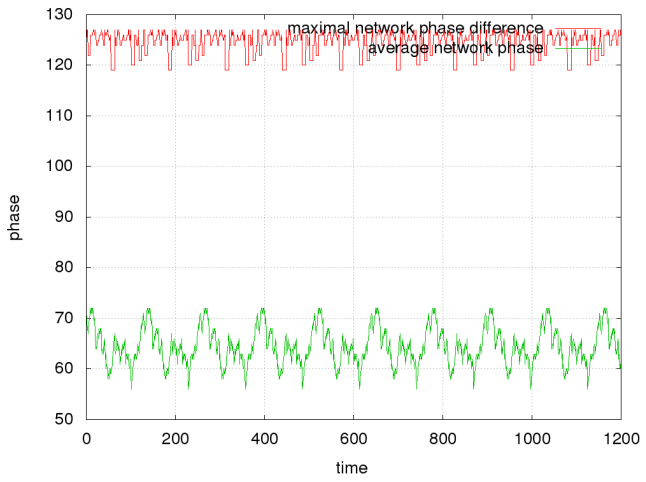


Fig. 2. Maximum and average phase of the network without synchronization

The source code (Algorithm 1) demonstrates the synchronization process in Ruby programming language.

Algorithm 1 The node synchronization

```
def tick(fired)
    @fired_tmp = false

    if fired
        @timer = TIMER_MAX
    end

    if (@state == 0)
        if (@timer >= TIMER_MAX)
            @timer-= 1
            @state = 1
            @fired_tmp = true
        else
            @timer+= 1
        end
    else
        if (@timer <= 0)
            @timer+= 1
                @state = 0
            else
                @timer-= 1
            end
    end
end
```

The method *tick* is called periodically, in discrete time, on each node. The input parameter *fired* has two values:

$$fired = \begin{cases} true\ when & node[j][i+1].fired\ or \\ & node[j][i-1].fired\ or \\ & node[j+1][i].fired\ or \\ & node[j-1][i].fired \\ false & else \end{cases}$$

When any of the neighboring nodes fires (timer is on top), this node sets the timer counter to $TIMER\_MAX$ value. Depending on the state, in the next step, the timer is either decremented, or incremented and compared to the top value. If the counter reaches the maximum value, the node fires, and the state is switched.

Algorithm 2 Network synchronization

```
for j in 0..@net.size-1
 for i in 0..@net[j].size-1
  if (@net[(j+1)%@net.size][i].get_fired) or
     (@net[(j-1)%@net.size][i].get_fired) or
     (@net[j][(i+1)%@net[j].size].get_fired) or
     (@net[j][(i-1)%@net[j].size].get_fired)

   fired = true
  else
   fired = false
  end

  @net[j][i].tick(fired)
 end
end
```

The network synchronization must work in discrete time so *fired* flag is stored in $@fired\_tmp$ first. After all nodes call the method *tick*, the $fired$ flags are updated.

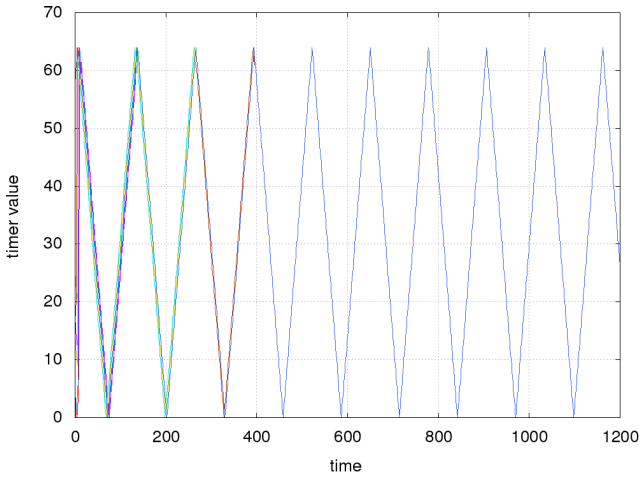


Fig. 3. Phases of nodes with synchronization

Fig. 3 and Fig. 4 demonstrate the network synchronization process. Fig. 3 illustrates the phase of the nodes when synchronization process applies On Fig. 4 we can see fluent decreasing of the phase error, until it reaches 0.

The following figures Fig. 5, Fig. 6, Fig. 7 and Fig. 8 represent the phase synchronization process and demonstrate the *synchronization wave* in the network in different iterations of the synchronization algorithm. As it can be seen, the waves with the increasing number of synchronization algorithm iterations slightly disappear.

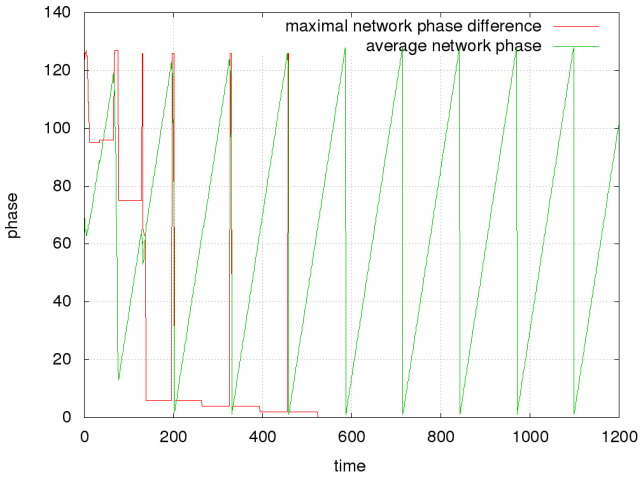


Fig. 4. Maximum and average phase of the network with synchronization

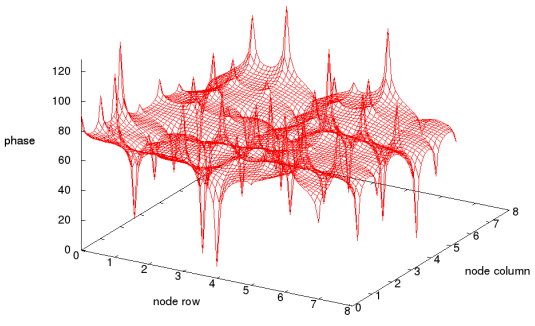


Fig. 5. Network initial phases, iteration 0

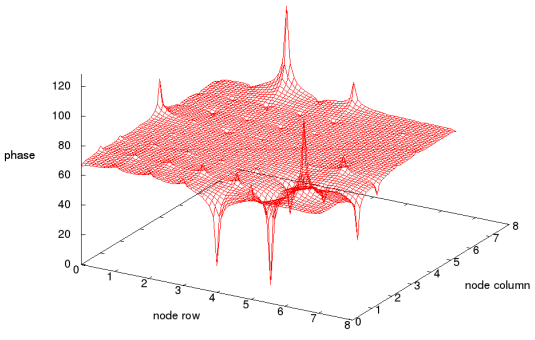


Fig. 6. Network synchronization, iteration 10

By comparing Fig. 6 and Fig. 7 we can see the *synchronization wave*. When the network is fully synchronized, all

phases are the same and we get a straight plane as illustrated in Fig.8. During the simulations performed on the self-developed simulator, we found out that the proposed synchronization algorithm is fully functional despite of node failure. This has also no effect on the overall WSN operation.

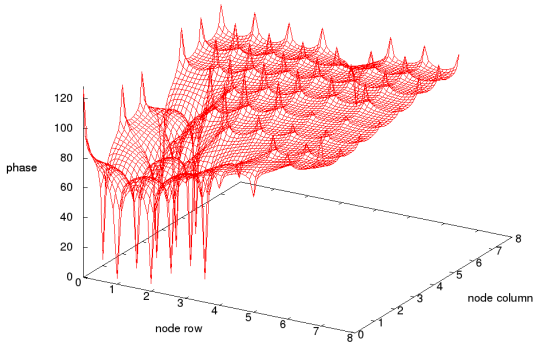


Fig. 7. Network synchronization, iteration 70

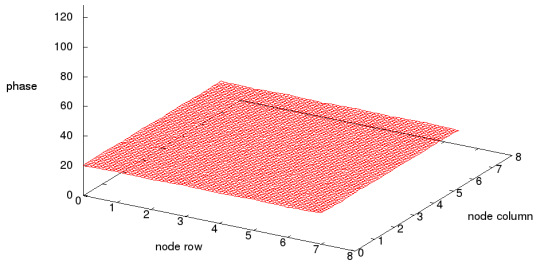


Fig. 8. Network synchronization, iteration 1180

III. Node Implementation

The experimental evaluation of the proposed synchronization algorithm has been performed on wireless sensor nodes based on Texas Instruments MSP430 family microcontrollers (MSP430F2232, 8kB FLASH, 512b RAM) (Fig. 9). The communication subsystem (RF) part is based on Texas Instruments transceiver CC1101 - a low power transceiver operating at $868MHz$ ISM band. Other sensor node subsystems built-in on-board are $3.3V$ low-drop regulator, three-axis magnetometer, two LEDs and an UART peripheral for debugging. Further details about the nodes can be found in [9]. Red board in Fig. 9 was used as a power supply and for debugging purposes.

After general input/output peripheral (GPIO) initialization and RF module initialization, the $timer0a$ interrupt is set to $256Hz$ periodic invoke (derived from external $32.768kHz$ crystal-based clock source). This initialization is described in the Algorithm 3.

Algorithm 3 Low power timer initializiation

```
/* f = ACLK (32768Hz) / TACCR0*/
TACCR0 = 128;

/* select ACLK/1, up mode, and clear the TAR */
TACTL = TASSEL_1 + MC_1 + TACLR;

/* enable timer interrupt*/
TACCTL0 = CCIE;
```

The Algorithm 1 described in the previous paragraphs is implemented in three program subroutines as follows:

- Synchronization with received packet (in GPIO pin interrupt routine);
- Periodical packet transmission and timer control (in timer interrupt routine);
- Higher-level network functions (in the main loop routine).

In GPIO initialization phase, RF receive pin is configured as an interrupt pin. After this steps the microcontroller (MCU) is allowed to enter the LPM3 sleep mode, which it can wake up from with $256Hz$ period and/or on the packet receive interrupt. The received packet processing is described in the following algorithm - Algorithm 4.

Algorithm 4 Packet receive interrupt service rutine

```
interrupt (PORT2_VECTOR) Port_2 (void)
{
  if ((__state__ == 2) &&
    (TI_CC_GDO0_PxIFG & TI_CC_GDO0_PIN))
  {
    u8 len = PACKET_LENGTH-1;
    if (RFReceivePacket((u8*)rxBuffer,&len))
    {
      __state__ = 0;

      i16 add = TIMER_MAX - __phase__;
      __phase__+= add;

      if (__phase__ > TIMER_MAX)
        __phase__ = TIMER_MAX;

      if (__phase__ < TIMER_MIN)
        __phase__ = TIMER_MIN;
    }
  }

  TI_CC_GDO0_PxIFG &= ~TI_CC_GDO0_PIN;
  LPM3_EXIT;
}
```

This routine (Algorithm 4) is executed only when the program is in $\_\_state\_\_ == 2$. That means packet receiving is enabled, which maximizes the sleep time. When a packet from other node is received, this interrupt routine is executed.

The phase synchronization is done on the following source code lines:

$$i16\ add = TIMER\_MAX - \_\_phase\_\_;$$
$$\_\_phase\_\_ += add;$$

In the timer interrupt routine the timer phase control is processed. This routine is exemplified in Algorithm 5.

Algorithm 5 Timer interrupt service rutine

```
interrupt(TIMER0_A0_VECTOR)
TIMER0_A0_ISR(void)
{
  __phase__+=__sh_output__/(i16)TIMER_MAX;

  if (__phase__ > TIMER_MAX)
    __phase__ = TIMER_MAX;

  if (__phase__ < TIMER_MIN)
    __phase__ = TIMER_MIN;

  if ((__phase__ + DEVICE_ADDR) == TIMER_MAX)
    __state__ = 1;


  if (__sh_output__ == TIMER_MAX)
  {
    if (__phase__ >= TIMER_MAX)
      __sh_output__ = TIMER_MIN;
  }
  else
  {
    if (__phase__ <= TIMER_MIN)
      __sh_output__ = TIMER_MAX;
  }

  LPM3_EXIT;
}
```

There is important to add a small phase difference, to avoid transmission collision, which is done as follows:

$$if((\_phase\_ + DEVICE\_ADDR) == TIMER\_MAX)$$
$$\_\_state\_\_ = 1;$$

When program enters into $\_\_state\_\_ = 1$ the received packet can be released to the main loop.

IV. Conclusion and Future Work

For the proper function of each WSN, the node synchronization is very important part of proposed solution. In this paper the algorithm for the synchronization was proposed, which is very easy for implementation. This algorithm is not only for hierarchical networks, it is universal and scalable. Functionality of synchronization algorithm was tested on mesh network up to 1024 nodes. It is good to note that in this paper only part for synchronization process is written. In future the authors would like to extend the algorithm and use algorithm together with monitoring techniques. That means that the communication window should be divided into the smaller slots for sending the data.

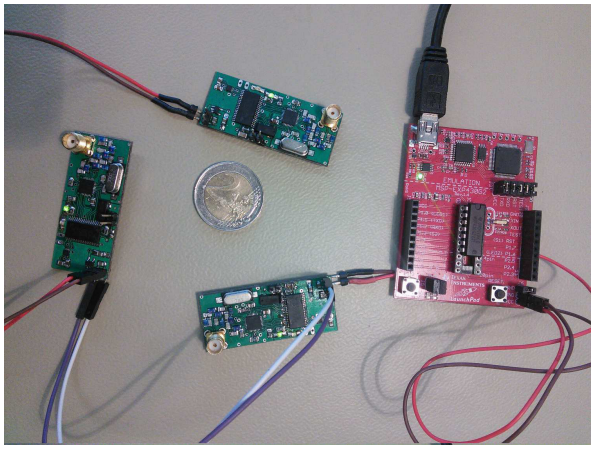


Fig. 9. Testing nodes photo

Acknowledgment

This publication is the result of the project implementation *"WSN pre monitoring a ochranu územia pohoria Malá Fatra"* with grant no. 2013et032 in grant program *"E-Talent 2013"* supported by Tatra banka Foundation Slovakia.

References

[1] J. Elson, and K. Römer, "Wireless sensor networks: A new regime for time synchronization", *in ACM SIGCOMM Computer Communication Review*, Vol. 33, No. 1, 2003, pp. 149–154.

[2] M. Kochláň, J. Miček, and M. Hyben, "Wireless sensor network energy harvesting: radio frequency harvesting case study", *in Intelligent transportation systems 2013 (ITS 2013)*, August 26-30, 2013, ISSN 1339-4118, ISBN 978-80-554-0763-0, pp. 93–97.

[3] M. Kochláň, M. Hodoň, and J. Púchyová, "Vital functions monitoring via Sensor Body Area Network with smartphone network coordinator", *in MEMSTECH 2013: perspective technologies and methods in MEMS design*, April16-20, 2013, Polyana, Ukraine, Lviv Polytechnic Publishing House, ISBN 978-617-607-424-3, pp. 143–147.

[4] J. Púchyová, M. Kochláň, and M. Hodoň, "Development of special smartphone-based Body Area Network: Energy Requirements", *in Federated conference on computer science and information systems (FedCSIS)*, September 8-11, 2013, Kraków, Poland, ISBN 978-1-4673-4471-5, pp. 915–920.

[5] D. Laqua, and P. Husar, "Intelligent Power Management enables Autonomous Power Supply of Sensor Systems for Modern Prostheses", *in Biomedizinische Technik/Biomedical Engineering (Impact Factor: 1.16)*, September, 2012, pp. 247–250. DOI:10.1515/bmt-2012-4055.

[6] J. Miček, O. Karpiš and P. Ševčík, "Body area network: analysis and application areas", *in International journal of engineering research and development (IJERD)*, ISSN 2278-800X, Vol. 6, No. 8, 2013, pp. 22–26.

[7] A. Hofmann, D. Laqua, and P. Husar, "Piezoelectric Based Energy Management System for Powering Intelligent Implants and Prostheses", Biomedizinische Technik/Biomedical Engineering (Impact Factor: 1.16). September, 2012, pp. 263–266. DOI:10.1515/bmt-2012-4265.

[8] M. Kochláň, and J. Miček, "Indoor Propagation of 2.4GHz Radio Signal: Propagation Models and Experimental Results", *in Digital Technologies 2014 (DT2014)*, July, 9-11, 2014, [in print].

[9] M. Hodoň, M. Chovanec, and M. Hyben, "Intelligent traffic-safety mirror", *in Studia informatica universalis*, ISSN 1621-7545, 2013, Vol. 11, No. 1, pp. 87–101.
[10] O. Karpiš, and J. Miček, "Sniper localization using WSN", *in ICMT'11: International conference on military technologies*, Brno, Czech Republic, May 10-11, 2011, ISBN 978-80-7231-787-5, pp. 1063–1068.
[11] V. C. Gungor, and G. P. Hancke, "Industrial Wireles Sensor Networks: Challenges, Design Principles and Technical Approaches", *in IEEE Transactions in Industrial Electronics*, Vol. 56, No.10, 2009.
[12] A. Flamminia, P. Ferraria, D. Mariolia, E. Sisinnia, and A. Taronib, "Wired and wireless sensor networks for industrial applications", *in 2nd IEEE International Workshop on Advances in Sensors and Interfaces*, Vol. 40, No. 9, September 2009, pp. 1322–1336.
[13] U. S. Department of Energy, "2011 Building Energy Data Book", 2012.
[14] European Commission, "Report from the Commission to the European Parliament and the Council", Brussels 18.4.2013.
[15] P. Ševčík, and O. Kovář, "Alternative energy sources for WSN node power supply", *in ITS 2013: Intelligent transportation systems 2013*, August 26-30, 2013, ISSN 1339-4118, ISBN 978-80-554-0763-0, pp. 146–149.
[16] J. Miček, and M. Kochláň, "Energy-efficient communication systems of wireless sensor networks", *in Studia informatica universalis*, ISSN 1621-7545, 2013, Vol. 11, No. 1, pp. 69–86.
[17] J. Miček, and J. Kapitulík, "WSN sensor node for protected area monitoring," *in Federated Conference on Computer Science and Information Systems*, 2012, pp. 803–807.
[18] O. Karpiš, J. Juríček, and J. Miček, "Application of wireless sensor networks for road monitoring," *in 10th IFAC workshop on programmable devices and embedded systems*, Vol. 3, 2013, pp. 611–617.
[19] M. Hyben and M. Hodoň, "Low-cost command-recognition device," *in Przegląd teleinformatyczny*, 2013, ISSN 2300-5149, Vol. 1, No. 3, pp. 19–28.
[20] J. Papán, M. Jurečka, and J. Púchyová, "WSN for forest monitoring to prevent illegal logging", *in FedCSIS: proceedings of the Federated conference on computer science and information systems*, September 9-12, 2012, Wroclaw, Poland, ISBN 978-83-60810-51-4.
[21] F. Cristian, "Probabilistic clock synchronization," *Distributed Computing*, vol. 3, 1989, pp. 146–158.
[22] R. Gusella and S. Zatti, "The accuracy of the clock synchronization achieved by TEMPO in Berkeley UNIX 4.3BSD," *in IEEE Transactions on Software Engineering*, 1989, pp. 847–853.
[23] D. L. Mills, "Internet time synchronization: the Network Time Protocol," *in IEEE Trans. Communications COM-39, 10*, 1991, pp. 1482–1493.
[24] E. Jeremy, G. Lewis and E. Deborah, "Fine-grained network time synchronization using reference broadcasts," *in Fifth Symposium on Operating Systems Design and Implementation OSDI*, 2002.
[25] S. Ganeriwal, K. Ram and M. B. Srivastava, "Timing-sync protocol for sensor networks," *in First ACM Conference on Embedded Networked Sensor Systems* 2003.
[26] A. Rahamatkar and A. Agarwal, "A Reference Based, Tree Structured Time Synchronization Approach and its Analysis in WSN," *International Journal of Ad hoc, Sensor & Ubiquitous Computing*, Vol. 2, 2011, pp. 20–31.
[27] J. V. Greunen and J. Rabaey, "Lightweight Time Synchronization for Sensor Networks," *in Proceedings of the 2nd ACM International Conference on Wireless Sensor Networks and Applications (WSNA)*, San Diego, CA, September 2003.
[28] W. Su and I. F. Akyildiz, "Time-Diffusion Synchronization Protocols for Sensor Networks," *in IEEE/ACM Transactions on Networking*, Vol. 13, 2005, pp. 384–397.
[29] K. Shahzad, A. Ali and N.D. Gohar, "ETSP: An Energy-Efficient Time Synchronization Protocol for Wireless Sensor Networks," *in 22nd International Conference on Advanced Information Networking and Applications - Workshops*, 2008, pp. 971–976.
[30] R. Albu, Y. Labit, G. Thierry and B. Pascal, "An Energy-efficient Clock Synchronization Protocol for Wireless Sensor Networks," *Wireless Days*, 2010, pp. 1–5.
[31] A. P. Sage and G. W. Husa, "Algorithms for Sequential Adaptive Estimation of Prior Statistics," *IEEE Symposium on Adaptive Processes Decision and Control*, 1969, pp. 760–769.
[32] C. Bo, D. Enqing, L. Xiaoyang, Z. Dejing and W. Jiaren, "A time synchronization algorithm based on bimodal clock frequency estimation," *in 18th Asia-Pacific Conference on Communications (APCC)*, 2012, pp. 75–78.
[33] J. Kim, J. Lee, E. Serpedin and K. Qaraqe, "A robust clock synchronization algorithm for wireless sensor networks," *in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 3512–3515.
[34] R. E. Mirollo and S. H. Strogatz, "Synchronization of Pulse-Coupled Biological Oscillators,"*SIAM Journal on Applied Mathematics*, vol. 50, 1990, pp. 1645–1662.
[35] A.Tyrrell, G. Auer and C. Bettstetter, "Firefly Synchronization in Ad Hoc Networks,"*in 3rd MiNEMA Workshop*, Lueven, Belgium, 2006.
[36] A. Tyrrell, G. Auer and C. Bettstetter,"Fireflies as Role Models for Synchronization in Ad Hoc Networks,"*Bio-Inspired Models of Network, Information and Computing Systems*, 2006, pp. 1–7.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1009–1015
DOI: 10.15439/2014F20
ACSIS, Vol. 2

A hybrid indoor localization solution using a generic architectural framework for sparse distributed wireless sensor networks

Tom Van Haute*, Jen Rossey, Pieter Becue, Eli De Poorter, Ingrid Moerman, Piet Demeester
Ghent University - iMinds, Department of Information Technology (INTEC), Belgium
* tom.vanhaute@intec.ugent.be, Gaston Crommenlaan 8, Bus 201, 9050 Ghent, Belgium, +32 9 331 49 46

***Abstract*—Indoor localization and navigation using wireless sensor networks is still a big challenge if expensive sensor nodes are not involved. Previous research has shown that in a sparse distributed sensor network the error distance is way too high. Even room accuracy can not be guaranteed.**

In this paper, an easy-to-use generic positioning framework is proposed, which allows users to plug in a single or multiple positioning algorithms. We illustrate the usability of the framework by discussing a new hybrid positioning solution. The combination of a weighted (range-based) and proximity (range-free) algorithm is made. Both solutions separately have an average error distance of 13.5m and 2.5m respectively. The latter result is quite accurate due to the fact that our testbeds are not sparse distributed. Our hybrid algorithm has an average error distance of 2.66m only using a selected set of nodes, simulating a sparse distributed sensor network. All our experiments have been executed in the iMinds testbed: namely at "de Zuiderpoort". These algorithms are also deployed in two real-life environments: "De Vooruit" and "De Vijvers".

I. Introduction

COMBINING wireless sensor network nodes with the upcoming trend of smartphones creates a totally new range of possibilities. Normally, wireless sensor networks are used to monitor a certain environment and measure e.g. the temperature and humidity. But also tracking of persons and equipment can be done by sensor networks. GPS [1] is the traditional way of tracking people or vehicles outdoor, however this does not work properly indoor because line of sight (LOS) is required to receive the GPS signals.

Sensor nodes inside buildings can fix this issue, however other factors have to be taken into account: interference, infrastructure, the amount of sensor nodes that is required, energy consumption,... It will always be a trade-off between cost and accuracy. Further, a myriad of positioning algorithms have been developed in the last few years. A standalone solution generally does not offer sufficient accuracy in different environments (indoor/outdoor, different type of buildings,...). In this paper however, we will try to find a solution with an acceptable accuracy when only a sparse distributed sensor network is available. Our algorithm described in this work is a combination of two already existing algorithms. Each belonging to a different subdivision, namely range-free and range-based. Both solutions show too many defects in thinner environments. Combining them results in a noticeable improvement. In this way, room accuracy can be guaranteed.

The rest of the paper is organized as follows. In section 2, the already existing algorithms are clarified. Section 3 describes our generic architecture framework. The hybrid algorithm build in this framework whereby the two previous are combined is discussed in section 4. Section 5 handles about our different testbeds. The experiments with their results are summarized in section 6. Finally, some conclusions can be made. These are, together with the future work, clarified in section 7.

II. Related work

In this section, we conclude the work that is essential to comprehend our hybrid solution. Localization algorithms can be subdivided in two different categories. The first category is called the "range-based"-algorithms. In order to calculate a position pertaining to multiple fixed nodes, a distance measurement is essential. Then, on the base of this distance, a position can be determined by means of trilateration. Trilateration is a method to find the intersection of three circles whose center and radius are known. There are many different ways to measure the distance. The most familiar techniques are Received Signal Strength Indication (RSSI), Time of Arrival (ToA), Time Difference of Arrival (TDoA) and Angle of Arrival (AoA).

The second category, "range-free"-algorithms, does not require a distance measurement to calculate the position of a sensor node. They are based on the information of the connection. If two sensor nodes can connect to each other, than the maximum distance between them is the maximum transmission range. Thus the position of the mobile node can be estimated with this information. This is a very simple and cheap technique. Moreover the accuracy will depend on the density of the wireless sensor network. Centroid, triangle elimination and proximity are common range-free algorithms.

The hybrid solution uses both techniques. A combination of a range-based and a range-free algorithm is made. In the following two sections, both algorithms will be explained more in detail.

A. Range-based: weighted

The first one is a range-based solution described by Tareq Ali Alhmiedat et al. in [2]. The proposed algorithm is based on weighted RSSI values. The main idea of RSSI is that the transmission power P_T directly affect the received power P_R

978-83-60810-58-3/$25.00

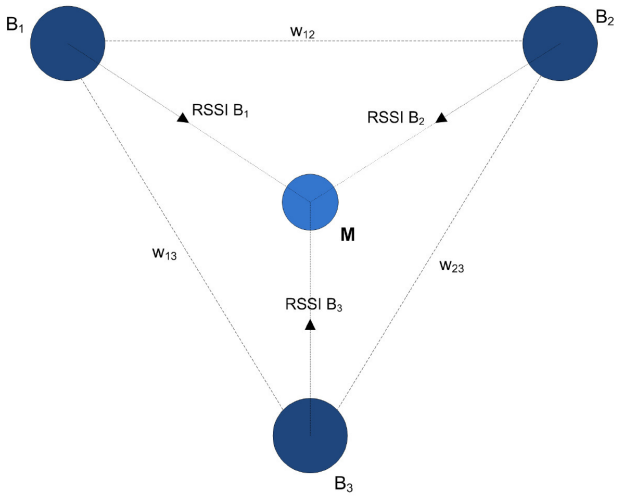


Fig. 1: Weighted algorithm: schema

of a signal. Via the Friis transmission equation, also defined in [2], the linear relationship becomes clear:

$$P_R = P_T * G_T * G_R \left(\frac{\lambda}{4\pi d}\right)^2 \tag{1}$$

where G_T, G_R are gain of transmitter and receiver respectively. λ is the wavelength of the signal and d is the distance between sender and receiver. The received signal strength indicator (RSSI) can be defined as the ratio of the received power to the reference power P_{Ref}.

$$RSSI = 10 * \log \frac{P_R}{P_{Ref}} \tag{2}$$

Each RSSI value can be matched with a certain distance. The proposed algorithm in [2] not only uses the RSSI-values to measure the distance between a fixed and mobile node, but also the distance between the fixed nodes mutually is measured. These values function as weight factors for the distance calculation between the fixed and mobile node. These weight factors are shown in Figure 1 as w_{12}, w_{13} and w_{23}. The distance from M to, for example, B_1 can be calculated as follows:

$$Distance(M, B_1) = \frac{RSSI\, B_1 * w_{12} + RSSI\, B_1 * w_{13}}{2} \tag{3}$$

Their results prove that these weight factors add value to the accuracy. A drawback of the RSSI technique is that these measurements are very sensitive to the environment and potential changes in it. The relationship between the distance and RSSI depends on the room. For example, in a long corridor, the fixed nodes their signals will have a greater range because they reverberate through the long walls. In this way, completely different results can be obtained.

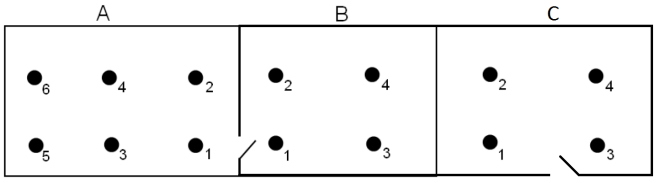


Fig. 2: Three neighboring offices

B. Range-free: proximity

In contrast to the previous category, range-free algorithms do not take RSSI-values into account. If a mobile sensor node has a range of 10 meters, than a fixed node can only receive his messages if the mobile node is maximum 10 meters away. This is the only information that is used to calculate the position of a mobile node. This technique is used by J. Wyffels et al. in [3]. A proximity-based algorithm is used to localize the patients and the nurses in a healthcare environment. Important here is that the transmission power is well configured. If the power is too low, the mobile node could be out of range between two fixed nodes. And also vice versa if the power is too high, too many fixed nodes will receive the beacon and a wrong estimation could be made.

The latter problem can be solved by using a centroid algorithm [4]. This is only useful if there is a set of fixed nodes with an overlapping coverage area. The beacon of the mobile node is received by multiple fixed nodes. In order to determine the position, the centroid of all the receiving fixed nodes is calculated (Eq. 4).

$$[x_M, y_M] = \left[\frac{\sum_{n=0}^{k-1} x_n}{k}, \frac{\sum_{n=0}^{k-1} y_n}{k}\right] \tag{4}$$

Normally would this algorithm give a 100% guarantee that room-accuracy is ensured. However, experiments have shown that this is not always the case. If the walls are small enough and/or not made of concrete, signals can go trough and a fixed node in a different room can catch up the beacon. In order to prevent a wrong location estimation, some extra logic can be implemented in the algorithm.

To implement the extra logic, some extra information is necessary as well. Suppose we have the exact coordinates of all the walls, doors and nodes inside a building. Knowing that every beacon has an index number, the direct path could be checked between the two fixed nodes who received the consecutive beacons. If the mobile node goes from one room to another, without using a door, then the last beacon can be dismissed. For example (Fig. 2) when node A_2 receives a beacon and the next beacon is received by node B_2. It is impossible to move directly from A_2 to B_2 without passing nodes A_1 and B_1. So the message that was received by beacon B_2 will be rejected.

With this optimization room-accuracy can be guaranteed. Still, this solution has the drawback that a lot of fixed sensor nodes are necessary to retrieve good results. If the network is sparse distributed, then the algorithm would not work properly.

III. Positioning framework

The framework is developed in Java and consists of three parts: the positioning server, the web server and the client application (Fig. 3).

The positioning server has two functional blocks. The interconnection gateway is responsible for the retrieval of positioning information gathered by the network infrastructure or mobile unit that is being located. The interconnection gateway further incorporates an abstraction layer which hides the underlying technology (ZigBee, Wi-Fi, Bluetooth, ...) from the positioning server. In Figure 3, two different approaches for positioning in wireless sensor networks are shown. On the left side, a mobile device broadcasts positioning beacons and the sink node of the WSN forwards the beacons to the interconnection gateway. On the right side, the infrastructure nodes broadcast beacons and the mobile unit collects and forwards the beacons to the interconnection gateway. The interconnection gateway further passes the positioning information to the position calculator, which consists of pluggable positioning algorithms. Multiple positioning algorithms can be active at the same time. A reasoner is used to select the algorithm giving the most accurate position or to intelligently combine the results of multiple algorithms into a more accurate (hybrid) position. Map info can also be taken into account when calculating the position.

The web server can poll the positioning server for the user's position. And the client application can either run on a smartphone or a central monitoring station. The client communicates with the web server through e.g. Wi-Fi or Ethernet.

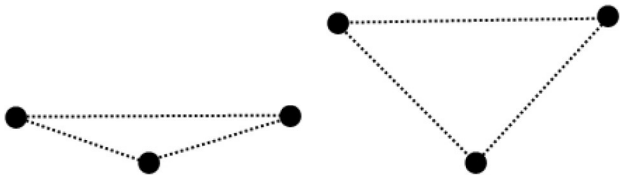
Some advantages of the framework:

- Existing smartphone applications can use position information by implementing a simple interface allowing the application to request a user's position from the web server.
- Conversion of relative coordinates to GPS notation is possible. This implies that client applications developed to work outdoor (GPS), can easily use this framework.
- The user of the client application can pinpoint his correct location on the floor plan (for testing purposes). The application then calculates the difference between the estimated and the real position, thus allowing the user to evaluate the algorithm.

IV. Hybrid algorithm

Having this framework described above, designing a hybrid solution is very efficient. The reasoner allows the position calculator to combine the results of different algorithms and other available information. In the hybrid solution the reasoner has two choices: if the mobile node is in range of a fixed node we use the result of the proximity algorithm. If no fixed node can hear the proximity node, the reasoner will decide to use the weighted RSSI algorithm, where the mobile unit has a wider range.

The biggest problem of the stand-alone weighted algorithm, is the selection of the nodes. An ideal situation would be that the proximity node would be surrounded by the closest fixed nodes. As discussed in subsection II-A, selecting the closest nodes is not possible if only the RSSI values are available. The proximity algorithm can give extra information whereby finding the closest nodes can be realized. Hence, the node selection can be optimized using the latest information of the proximity algorithm. In that way, the first node of the triangle is determined. In order to have a good coverage of the area, the two other nodes must be well selected. If the angles of the triangle are too sharp (Fig. 4a) than the weighted algorithm will not function properly. In certain situations, the two last nodes will have to be reselected until a good triangle (Fig. 4b) is founded.



(a) Bad triangle (b) Good triangle

Fig. 4: Node selection

Data: Three circles of each fixed node
Result: Position of the mobile node
if *three circles do not intersect* **then**
  while *smallest circle does not intersect with the second smallest circle* **do**
    increase the smallest circle
  end
end
`// Now at least two circles intersect`
Calculate the intersection of the two smallest circles
position mobile node = intersection of the two smallest circles closest with the biggest circle

Algorithm 1: Adapted trilateration

Once the three fixed nodes are selected, a distance measurement is the next step in the procedure. This is done the same way as the stand-alone weighted algorithm (Eq. 3) except for one thing. The RSSI values are slightly adapted because results of previous experiments have shown that the calculated distance was almost always too big. This adaption is estimated experimentally. After the distance calculation, the three circles can be created and trilateration can be applied. In perfect circumstances, the three circles will intersect in exactly one point. However, in practice this is never the case. Due to the environment and interference, the three circles will never intersect in one single point. Therefore, an adapted trilateration technique is shown in Algorithm 1.

Finally, if the reasoner has access to other input, such as information about walls, rooms, doors, we can use this to influence our position estimate.

V. Environment descriptions

This positioning framework including the hybrid solutions has been tested in two wireless testbeds and also in two dif-

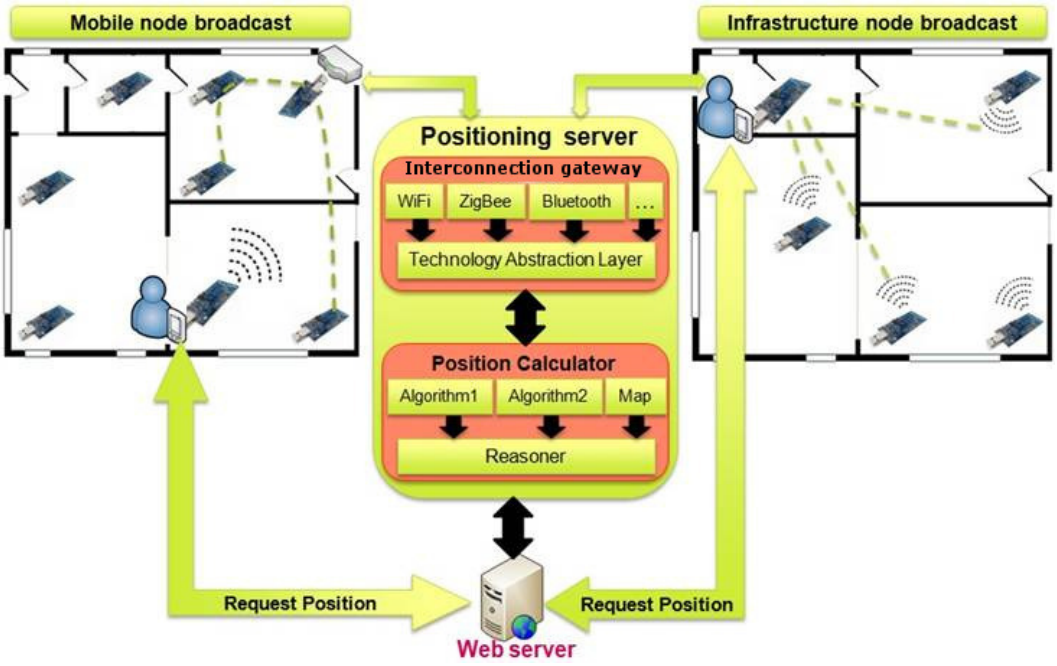


Fig. 3: Framework architecture

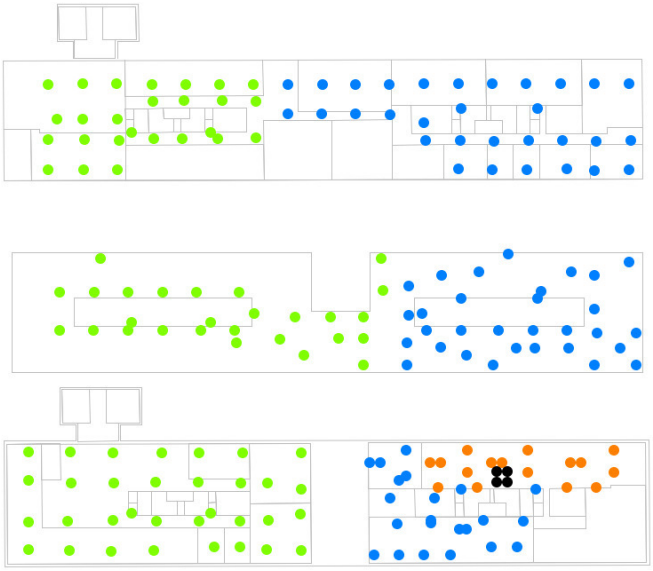


Fig. 5: w-iLab.t at the "Zuiderpoort". 200 nodes on three floors (18m x 90m). The different colors indicate the nodes can be divided in groups for executing tests.

ferent real-life situations. Each environment will be explained further in detail in the next sections.

A. *w.iLab-t at the "Zuiderpoort"*

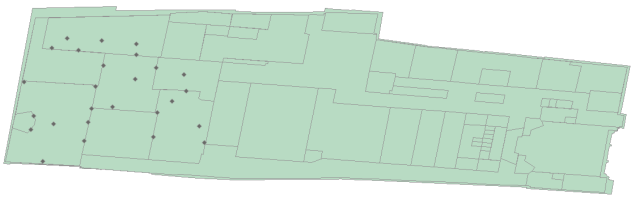
The w.iLab-t is an extensive facility that is introduced in detail by S. Bouckaert et al. in [5]. The infrastructure is distributed on three floors of the iMinds office in Ghent, Belgium (Fig. 5). The network consists of 200 nodes. Each node has (i) a Tmote Sky IEEE 802.15.4 mote, (ii) two Compex WLM54SAG 200mW AR5006XS 802.11a/b/g 54/108 Mbps miniPCI wireless cards and (iii) an environment emulator. The latter one is self-made and used for simulations: environment (e.g. temperature change), battery drop, user input, etc. These nodes are centrally managed for control and monitoring purposes and remote access by using an Intel x86 architecture (PC Engines Alix Boards).

B. *"De Vooruit"*

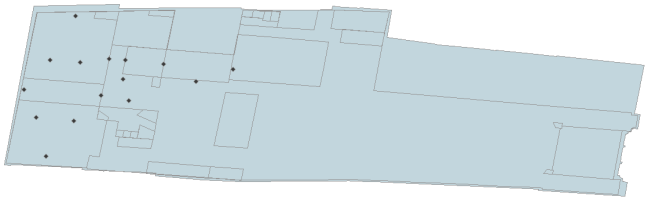
De Vooruit is an ancient building close to the historical center of Ghent [7]. In the past, this building was a place for the working class where they could eat, drink and enjoy culture at democratic prices. Since 1982 De Vooruit is recognized as a monument and nowadays it is still used to organize lectures, debates, concerts, parties, ... This location was a perfect use case to test the indoor localization solutions. Due to the fact that the building was recognized as a monument, it was not allowed to use a cabled network. In this situation, wireless sensor networks were the only solution to handle this problem. 50 nodes, distributed over four different floors (Fig. 6), were used to locate the mobile nodes worn by the visitors. In this use case, Sentilla JCreate nodes in combination with battery packs were used.

C. *"De Vijvers"*

As a second use case, the positioning was tested in a home for the elderly. The goal here was to track people with dementia that are not allowed to leave the home. When a person goes in a restricted zone, an alarm was sounded. The position of the person could then be seen on a map in the reception. In this building (Fig. 7), 25 Sentilla JCreate



(a) Fourth floor of De Vooruit



(b) Sixth floor of De Vooruit

Fig. 6: Floorplan of the fourth and sixth floor of De Vooruit, the diamonds represent the fixed nodes installed in the building.

Fig. 7: The southern part of "De Vijvers" where 25 nodes were attached in the central and eastern part of the building. Their positions are marked with red dots.

nodes were attached to obtain the required accuracy for this application.

VI. Results

In this section, we present the results of all the interesting measurements. First, the two algorithms are tested separately, followed by the results of the hybrid solution. All these measurements are done at De Zuiderpoort (Section V-A) on the third floor.

A. Range-based: weighted

The results from [2] showed that the weighted RSSI-values give a more accurate position than the normal RSSI-values. For

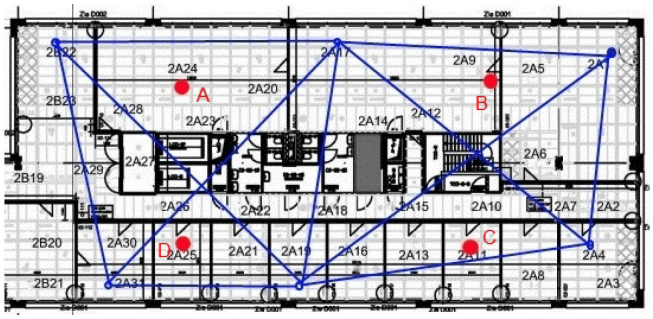


Fig. 8: Dividing the third floor in big triangles for the second test of the weighted algorithm.

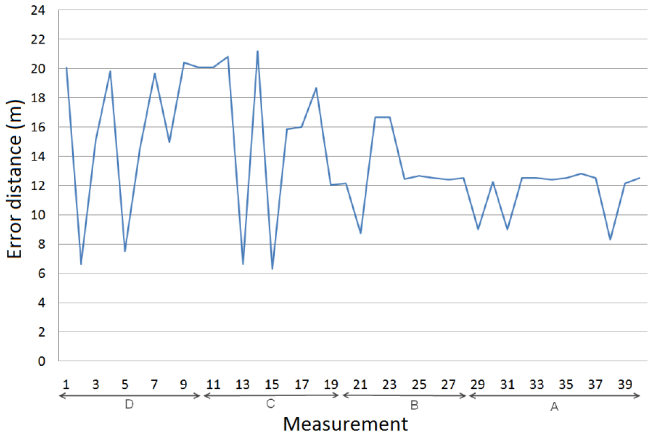


Fig. 9: Graphical overview of the results of the weighted algorithm using a sparse distribution of the fixed nodes

those reasons, only the results of the weighted algorithm will be shown.

Several tests have been executed. First, all available nodes on the third floor were used. This, however, gave very poor results so they are not published in this work. Some measurements had an error distance of more than 20 meters. An explanation for these large error distances is multipath fading of the nodes in the corridor. The setup of the second test is shown in Figure 8. The third floor is divided in big triangles (marked with blue lines) to calculate the position of A, B, C and D (marked with a red dot). The results of these measurements can be found in Figure 9. For each location, ten measurements are executed. The smallest error distance is $6.3m$, the biggest is $21.2m$ with an average error distance of $13.8m$. These results are not acceptable because room accuracy cannot be guaranteed. The large error distance is due to the fact that a high transmission power was necessary to communicate trough the concrete walls in the center of the building. The concrete walls has a strong influence on the RSSI measurements. For those reasons, a third test was implemented that avoids the concrete walls.

The triangles of the third test can be found in Figure 10. In this way, the signals do not need to go through the concrete walls so a lower tx-power can be used. The results of this

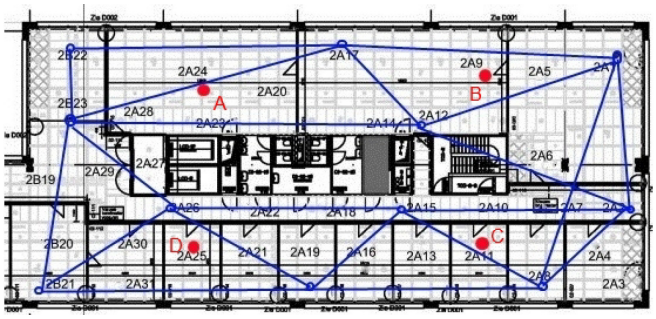


Fig. 10: Dividing the third floor in small triangles for the third test of the weighted algorithm.

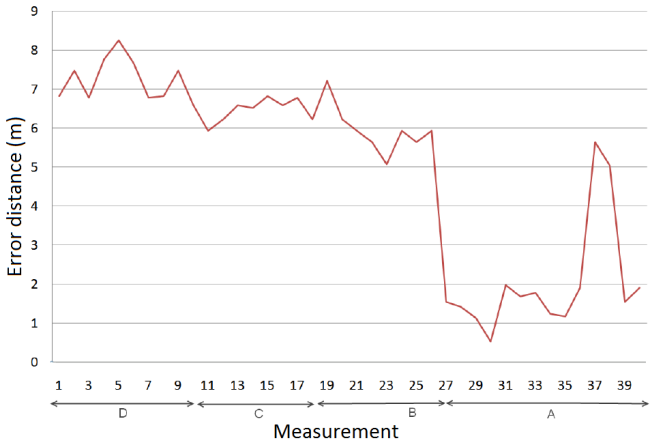


Fig. 11: Graphical overview of the results of the weighted algorithm using a dens distribution of the fixed nodes

setup are represented in Figure 11. Again, in this test, ten measurements at each location are recorded. The smallest error distance here is only $0.5m$, the largest one $8.2m$. The average error distance is $4.8m$. These results are much better than with the large triangles, but still, an error distance of $8m$ is unacceptable.

In these results, it became clear that this single algorithm was not capable to deliver the room accuracy.

B. Range-free: proximity

The results of the proximity based algorithm are completely dependent of the used infrastructure. The density of the fixed nodes determines the accuracy of the localization. Our algorithm is tested in the w-iLab.t at De Zuiderpoort (Section V-A) where the fixed nodes have an intermediate distance of 5 meters. This means that the maximum error distance is about $2.5m$. In the best case, the mobile node is located right under the fixed node, meaning that the error distance is $0m$.

C. Hybrid algorithm

The hybrid algorithm is designed to work properly in sparse distributed sensor network. However, the w-iLab.t is not sparse distributed. For this reason, it was hard to retrieve results using the whole testbed, the proximity beacons were always

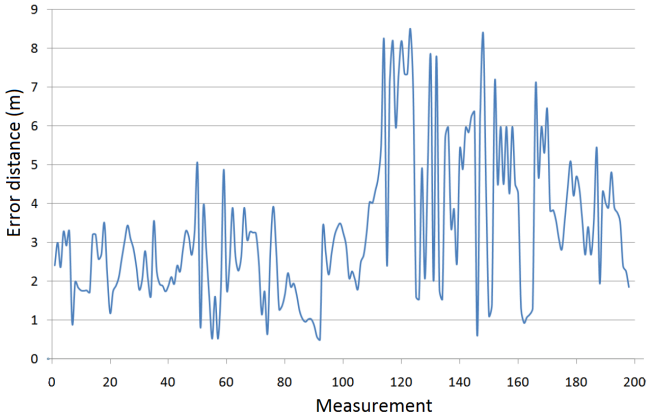


Fig. 13: Graphical overview of the results of the hybrid algorithm using a sparse distribution of the fixed nodes

reachable whereby the hybrid algorithm almost never switched from the proximity to the weighted algorithm. This produced the same results as in Subsection VI-B. In order to test this algorithm its full functionality, some artificial tests are done. From the moment a fixed node received a proximity beacon, the transmitting of proximity beacons by the mobile node will be stopped. Hereby, the weighted algorithm has to come active to calculate the final location of the node.

The mobile node is placed at different locations on the third floor in De Zuiderpoort building, these are marked in Figure 12 with the blue spots. The results of these measurements can be found in Figure 13, these are the worst possible results because the proximity is often disabled in order to activate the weighted part of the hybrid algorithm. 200 measurements were made across the different locations. The minimum error distance was $0.49m$ and maximum $8.5m$. The average of all the measurements together was $3.28m$. The worst results are due to the fact that some fixed nodes are placed in ventilation ducts. These are hard to reach for the signals of the mobile node. The RSSI-values of these messages are extremely low causing a greater error on the distance calculations from the mobile node to the fixed node in the ventilation duct. This affected the results significantly, when we drop all the results of the fixed nodes in the ventilation ducts, the new average error distance is $2.66m$.

Hence, this algorithm has also some drawbacks. Each algorithm uses a different transmission power. It is very important that the proximity algorithm his transmission range can be limited to the half of the distance between the fixed nodes. The idea is that only one fixed node can receive the beacons at a time. But with the weighted algorithm, enough nodes need to receive the beacons from the mobile node in order to make triangulation work properly. The tx power of a Tmote Sky can be programmed dynamically, but in our case, extra attenuators were necessary to reduce the transmission range. To fix this issue in our situation, two mobile nodes were used.

D. Summary

A summary of all the experimental results can be found in Table I, all the minimum, maximum and average error

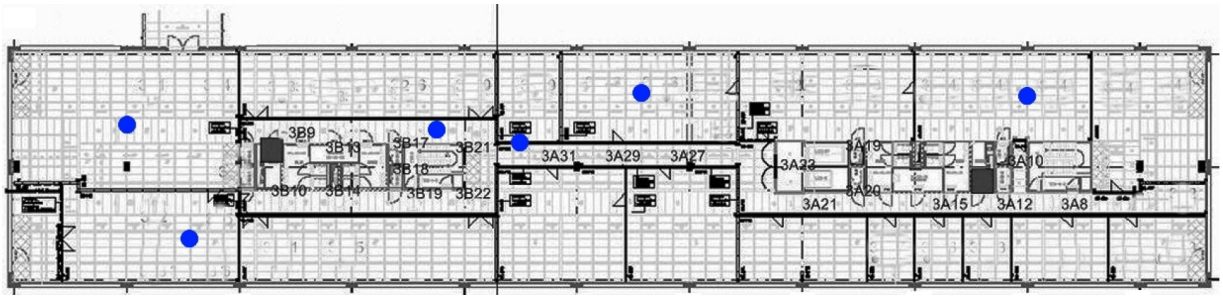


Fig. 12: Positions of the mobile node for testing the hybrid localization solution

TABLE I: Summary of the experimental results

Localization Solution	Error Distance (in meter)		
	Minimum	Maximum	Average
Weighted algorithm *(big triangles)*	6.3	21.2	13.8
Weighted algorithm *(small triangles)*	0.5	8.2	4.8
Proximity algorithm	0	2.5	-
Hybrid algorithm *(all nodes)*	0.49	8.5	3.28
Hybrid algorithm *(filtered nodes)*	0.49	8.5	2.66

distances are collected in one organized table. It becomes clear that the hybrid solution has an improvement (if you compare the average error distances).

VII. Conclusion and Future Work

This paper presents a hybrid indoor localization solution designed to achieve room accuracy using sparse distributed sensor networks. Hereby a positioning framework is developed to accomplish a hybrid solution, based on two existing solutions.

The positioning framework consists of two functional blocks: the interconnection gateway and the position calculator. The interconnection gateway gathers all the necessary data from the fixed and mobile node in order to calculate the position. This data can come from any kind of hardware device/technology. The position calculator contains all the different localization algorithms that calculate the position of the mobile node using the data from the interconnection gateway. This calculator includes also a reasoner that decides which algorithm calculates the most accurate position at a certain moment, it can also combine multiple algorithms to improve the accuracy even more.

The hybrid solution is based on a range-based and range-free algorithm. The former is a category of techniques that requires distances measurements in order to calculate the position of a mobile node. These distance measurements can be done in different ways. In this paper, the range-based "weighted" algorithm is proposed. It uses RSSI measurements to calculate the distance between the nodes. The higher the value, the shorter is the distance between the nodes. Innovative here is that RSSI measurements are also used to calculate the distance between the fixed nodes mutually. Using this extra information, a *weighted* distance calculation can be done using triangulation.

The range-free solution "proximity" does not require these distance calculations, the localization is only based on the information of the connection. This means that the reception range of a fixed node is as well as the maximum error distance. However, an extra optimization is possible, if multiple fixed nodes receive a beacon, then the centroid of all the fixed nodes can be calculated and be assumed as the point closest to the mobile device.

Both algorithms show issues in sparse distributed sensor networks. The accuracy of the weighted algorithm is far from acceptable because it is not easy to determine the correct triangle for the calculation and the proximity solution is completely depended on the density of the fixed nodes. Therefore, combining both algorithms can resolve the biggest issues of both solutions. First a proximity beacon is received by a fixed node, this is the first corner of the triangle. Then the other two corners are determined in order to get a good triangle.

In the results, it became clear that the improvement of the hybrid solution is significantly. The average error distance dropped from $13.8m/4.8m$ to $3.28m/2.66m$. Still, some future work can be done. First, the issue with the transmission power must be tackled. Further, comparative tests using WiFi or other technologies are in progress.

Acknowledgement

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 317989 (STREP EVARILOS).

References

[1] Bulusu, N.; Heidemann, J.; Estrin, D. 'GPS-less Low Cost Outdoor Localization For Very Small Devices'

[2] Alhmiedat, T.A. and Yang, S-H. (2008) 'A ZigBee-based mobile tracking system through wireless sensor networks', *Int. J. Advanced Mechatronic Systems*, Vol. 1, No. 1, pp.63-70.

[3] Wyffels, J.; Goemaere, J.-P.; Verhoeve, P.; Crombez, P.; Nauwelaers, B. and De Strycker, L. (2012) 'A novel indoor localization system for healthcare environments', pp.1-6.

[4] Qiu, T.; Zhou, Y.; Xia, F.; Jin, N.; Feng, L. (2012) 'A localization strategy based on n-times trilateral centroid with weight', *Int. J. Communication Systems*, pp.1160-1177.

[5] Bouckaert, S.; Vandenberghe, W.; Jooris, B.; Moerman, I.; Demeester, P. 'The w-iLab.t testbed'.

[6] Bouckaert, S.; Becue, P.; Vermeulen, B.; Jooris, B.; Moerman, I.; Demeester, P. 'Federating wired and wireless test facilities through Emulab and OMF: the iLab.t use case'.

[7] De Vooruit, the monument. Online on 29 March 2013 (in Dutch) - http://vooruit.be/nl/information/detail/36/Monument

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1017–1022
DOI: 10.15439/2014F370
ACSIS, Vol. 2

Wireless Sensor Network – Value Added Subsystem of ITS Communication Platform

Ján Kapitulík
University of Žilina
Univerzitná 8215/1, 010 26 Žilina, Slovakia
Email: Jan.Kapitulik@fri.uniza.sk

Juraj Miček
University of Žilina
Univerzitná 8215/1, 010 26 Žilina, Slovakia
Email: Juraj.Micek@fri.uniza.sk

Matúš Jurečka
University of Žilina
Univerzitná 8215/1, 010 26 Žilina, Slovakia
Email: Matus.Jurecka@fri.uniza.sk

Michal Hodoň
University of Žilina
Univerzitná 8215/1, 010 26 Žilina, Slovakia
Email: Michal.Hodon@fri.uniza.sk

***Abstract*—The article is dedicated to the analysis of design of WSN road transportation system for developing applications related to parking management, traffic flows and emergency vehicles monitoring, weather and environmental conditions monitoring as well. The analysis is done for all units of sensor node. Many technical aspects of the design are discussed in the paper.**

I. Introduction

COMPARING to railway, civil aviation and marine transportation modes, road one is characteristic by the most of accidents and passenger fatalities, passenger and freight transport as well as CO_2 emission production. This is reason why safety, effective transport and environmental questions are core topics of research and development activities in the field of the Intelligent Road Transportation Systems (IRTS), [1].

In the ninetieth of last century, public and industry authorities decided about deploying electronics, information and communications technologies to improve negative situations by developing new V2X technology (V2V- vehicle-to-vehicle, V2I – vehicle-to-infrastructure sensor/dedicated short-range communications technology), [2] [3]. The technology is expected to guarantee reliable driving car coordinating movement and timing, particularly through detection of the vehicle´s position relative to other vehicles, intersections and infrastructures behind 250 m as well as self-driving car in case of critical situations, smooth traffic flows based on V2V and V2I communications to inform driver about recommended driving speed, in order to reduce unnecessary acceleration or slowdown. Unfortunately, V2X research and development works has still not been finished by now. It is expected that practical deployment of the technology could become reality in the second half of next decade. In present days it is possible to follow intensive discussion about future of V2X technology. Many other sensor, communications technologies have been developed during last fifteen years. No one expected so fast growing of especially mobile communications. Using personal digital devices is standard today. Simply said, situation has changed during last 25 years.

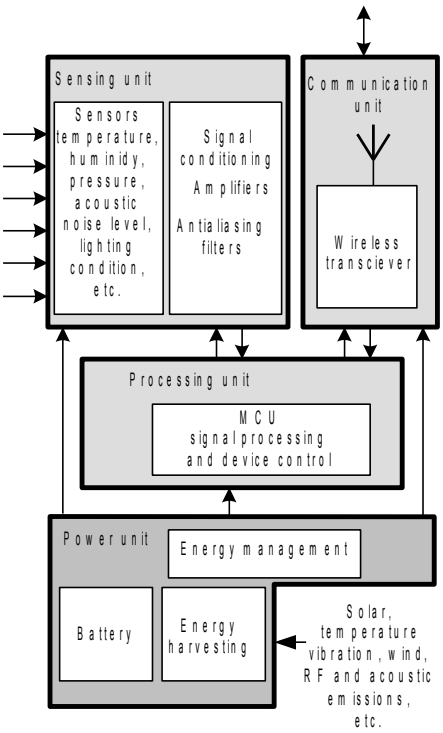


Fig 1.Block structure of wireless sensor node

This is reason why research and industry authorities in the USA and EU split into two groups concerning future development of V2X technology. While the goals of both of them are the still the same: decreasing traffic accidents, traffic congestion and improving fuel economy, European authorities focused on developing Advanced Driver Assistance Systems (ADAS) supported by mobile communications 4G

LTE technologies and services for upcoming events within 250 meters. The USA authorities prefers developing V2X technology for communication distance 1000m. It is expected that both of above mentioned technologies will be developing simultaneously to operate independently concerning V2V services and cooperating with each other in case of V2I ones.

Referring to above mentioned information, it would be suitable to answer question if WSN could be integrated in sensor/communication platform of future intelligent transportation systems. Since road transportation mode is dominant one in ITS, next part of article will be focused on analysis of specific features of WSN, particularly core subsystems of its sensor node (figure 1), supporting development of road traffic services keeping in mind value added complementary character of WSN.

II. Sensor Node – Sensing Unit

Mote is sensing of selected signals values by proper sensors and transforming of measured signals to ones which are suitable for additional processing (most often electrical voltage).

Respecting two facts:

- Sensor node is stationary in WSN
- Vehicle acts as mobile mote of ad-hoc wireless network

sensing of chosen signals must cover larg area. In such case WSN keeps value added complementary character to V2X and ADAS technologies. Selection of sensors is closely related to precision of measurement and developing applications. Thinking about road traffic, it is clear that WSN will be optimally used for monitoring purposes to support control, maintenance and planning processes. Core applications could be defined as follows:

- Parking monitoring and management
- Traffic flows monitoring
- Emergency vehicles monitoring
- Weather conditions monitoring
- Environmental conditions monitoring.

Parking monitoring and management

Vehicles detection at parking places is main task of the application. Sensors choice depends on indoor or outdoor detection. Magnetometer is proper sensor for both of cases. Measurement accuracy is not depended on weather conditions using the sensor. Numerous other ones are available for indoor detection: infrared, ultrasonic, acoustic, cameras, etc. Collected data are processed and information about free spaces can be presented on navigation tables (signs) in the street or distributed via communication channels to personal devices. Entry gate to parking lots is controlled on the basis of occupancy of them.

Traffic flows monitoring

mainly supports:

- *Vehicle detection* – at stop lanes of intersections
- *Vehicles counting* – counting number of vehicles waiting in queues of intersections, valued information for traffic lights control
- *Vehicle classification* – vehicle type identification for planning applications
- *Traffic flow intensity measurement* – in vehicles per time period, necessary information for effective road surface maintenance, planning of road network extension, traffic light control, see figure 2 and figure 3
- Vehicle speed measurement – supports safety and effective driving.

Referring to precision measurements, magnetometer sensor is proper for traffic flows monitoring.

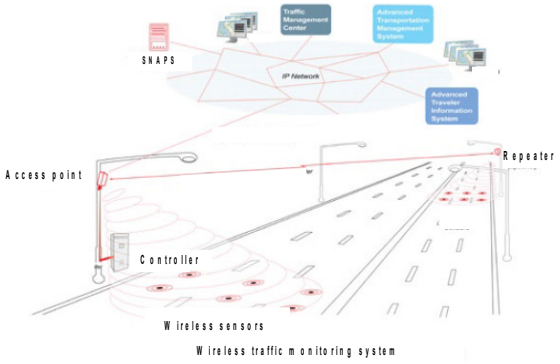


Fig2. Wireless traffic flows monitoring system

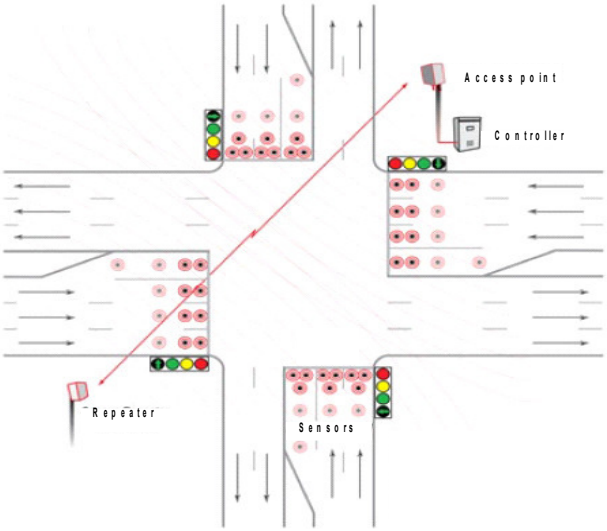


Fig3. Traffic lights control system

Emergency vehicles monitoring

Due to using sirens, emergency vehicles are effectively detected and classified by acoustic sensors. Monitoring of them allows safety crossing intersections.

Weather conditions monitoring

allows drivers to adapt style of driving in time to avoid accidents.

Environmental conditions monitoring

collects information about air pollution: CO_2 emissions, dust concentration, acoustic noise related to traffic in the area. Information is valuable for:

- Developing applications in compliance with citizens health protection
- Planning of reconstructions of historical buildings plasters.

III. Sensor Node – Processing Unit

Basic functions of processing unit are: digital processing of measured signal values, control of sensor node units, security of transmitted data, potentially another additive functions required by particular applications.

Keeping in mind value added complementary character of WSN to V2X and ADAS technologies (high speed transmission), processing unit very often uses reformatting, compression, classification technics and algorithms to transmit low amount of data. This approach supports low data rate transmission over communication channel (will be discussed in communication unit section). Referring to mentioned monitoring applications, number of transmitted bits vary from 1 bit to maximally 16 bits. It opens possibilities for developing of variable and reliable communication strategies.

IV. Sensor Node – Power Unit

Energizing of sensor node by electrical energy is main function of the unit. Time period of network operation without maintenance is important WSN feature.

Energy sources could be split as follows:

- *Mobile* – energizing of sensor node from primary batteries, rechargeable batteries, system exploitation for collecting energy from surrounding environment – energy harvesting, combination of rechargeable batteries and energy harvesting. This approach is standard in WSN applications.
- *Static* – energizing of mote is from standard electrical network. The option could be used in case of nodes responsible for communications services. They represents communication backbone of WSN. Such solution outgoing from the fact that communication unit consumes the most of energy of the battery. This option of motes energizing would be used, if necessary, in case of requirements for: increasing communication distance among nodes (to compensate path losses on frequency channels), guarantee of reliable communication and life time of motes operation as well.

Selection of energy source is strongly depended on requirements of application which WSN is designed for. Primary battery is expected to be preferable solution for road transportation applications (operational life time could be even 10 years).

V. Sensor Node – Communication Unit

Any sensor node must be able to communicate with adjacent ones, potentially with base station via wireless communication channel. Communication unit is main consumer of energy of the battery.

Analysis of WSN features, assuring successful communications in road transportation, area will be done in several steps.

Frequency bands allocation for WSN road transportation applications

In 5.8.2008, EU Committee decided to allocate frequency band from 5875 to 5905 MHz for ITS applications, which is going to be used on non-exclusive basis. Referring to 5.9 GHz center frequency, channel bandwidth is 10 MHz and default data rate 6 Mbps. Detailed information about V2X communication frequency band as well as maximum limit of mean spectral power density (EIRP) and European channel allocation is presented in [4].

It is clear that for 2-byte useful information transmission (low data rate is expected for WSN applications) above mentioned frequency channel parameters are not defined properly. That is why lower frequency bands are subjects of interest for successful WSN communication in road traffic applications.

Selecting lower frequencies for communication among communication units of motes has positive influence in general:

- *Channel bandwidth is more narrow* – saving frequencies for other applications, increasing number of channels for defined frequency range
- *Communication distance is increased* – competitive argument for WSN, development of cooperative services
- *Path Loss is decreased* – less energy losses higher quality and reliability of communication
- *Lower influence of obstacles on signal strength*
- *Improved radio receive sensitivity* – increased communication distance, weaker signal could be successfully received
- *Better resistance against weather conditions*- lower Bit Error Rate

Lower frequencies require larger antennas to achieve the same gain and improve signal robustness against interference in general. On the basis of above mentioned information it is possible to state that lower frequency band is attractive solution for design of WSN road transportation services, especially, when low data rates are required. It strengthens competitiveness of designed solutions.

For assuring successful wireless communications, it is suitable to mention practical path loss rules of thumb, [5]:

- To ensure basic fade margin in a perfect line of sight application, never exceed 50% of the manufacturer´s rated line of sight distance. This in itself yields a theoretical 6dB fade margin – still short of the required 10dB.
- Decrease data rate more aggressively if you have obstacles between the two antennas, but not near the antennas.
- Decrease data rate to 10% of the manufacture´s line of sight ratings if you have multiple obsta-

cles, obstacles located near the antennas, or the antennas are located indoors.

Defining frequency band for wireless communications in the ITS field must take into account one very important actuality. 5.9 GHz frequency band for V2X technology is used in non-exclusive basis. Drivers do not pay monthly recurring charges for using services requiring the band. This is reason why WSN road transportation applications must typically operate in "license free" frequency bands, also referred to as ISM (Industrial, Scientific and Medical). The most common frequencies encountered are:

- 2.4 GHz – nearly global coverage
- 915 MHz – North America, South America
- 868 MHz – Europe.

For any given distance, a 2.4 GHz installation will have roughly 8.5 dB of additional path loss when compared to 900 MHz. However, lower frequencies require larger antennas to achieve the same gain. Antenna type must be selected during in a proper way and earlier physical installation on site. Antennas increase the effective power by focusing the radiated energy in the desired direction. This fact could be evaluated during time period of WSN application design.

Table 1 presents result of simulations described in [6]. Mathematical backround is covered in the article as well. The paper is easily accessible via Internet in IEEE Digital Library, respectively in SCOPUS one.

Communication distance of V2X technology is required up to be 1000m. From the table is clear that no theoretical model satisfied expectations. Testing of V2X communication reliability was done at road infrastructure in California two years ago. Unfortunately, reliability of wireless communication was on the level of 85%. Still not satisfactory. It is discussed that the technology could be practically operating even in 2027. This is reason why European authorities prefer developing of ADAS system supported by mobile 4G LTE communication technologies.

Lack of reliable communication up to 1000m has negative influence on cooperative services development. This is chance for WSN road applications.

Selection of RF technology for WSN based road transportation applications

Table 2 presents chosen RF transmission systems applicable in the field of transportation.

Referring to battery life item, only ZigBee® technology is applicable in WSN networks in present days. It is designed for short range low power operation. The radio is relatively low data rate (up to 250 Kbps); the packets are short (< 128 bytes) and low energy. For example, sending a few bytes of sensor data, with routing, cryptography and other headers takes less than 1ms and burns less than 30µJ of energy, including receiving a secure link-layer acknowledgment. Sensors can forward radio packets from peers, extending the range of the network far beyond the range of single radio and providing the network with immunity to any single radio link failure.

TABLE I.
COMMUNICATION DISTANCE LIMITS AT MAXIMAL TRANSMIT POWER

Models	**Communication distance limit [m]**	
	G_R = 5 dBi	**G_R = 8 dBi**
Free-space model	1278	1805
ETSI model	405	523
ECC model ***Urban***	279	328
ECC model ***Suburban***	471	565
ECC model ***Rural***	933	1033

Every ZigBee standard and specification is the powerful IEEE 802.15.4 physical radio standard. It delivers raw data throughput rates of 250Kbs at 2.4GHz (16 channels), 40Kbs at 915MHz (10 channels) and 20Kbs at 868MHz (1 channel). Further information about Zigbee technology is possible to find in [7].

Design of WSN road transportation monitoring/control systems must meet a four core performance targets:

- *The first, the system must meet a minimum reliability goal.* For industrial applications, the target is typically to receive at least 99.9% of the generated data. Referring to RF communications, 99.5% link availability is defined as standard. 99.9% availability is considered as high one. This target is not met in current V2X technology.

TABLE II.
COMPARISON RF TECHNOLOGIES SUITABLE FOR INTELLIGENT ROAD TRANSPORTATION SYSTEMS

Market Name **Standard**	**ZigBee®** **802.15.4**	**---** **GSM/ GPRS/ CDMA / 1xRTT**	**Wi-Fi™** **802.11b**	**Bluetooth ™** **802.15.1**
Application Focus	Monitoring & Control	Wide area voice& data	Web, Email, Video	Cable replacement
System Resources	4KB-32KB	16MB+ 1-7	1MB+	250KB+
Battery Life (days)	100 - 1000+	1-7	1-5	1-7
Network Size	Unlimited	1	32	7
Maximum Data Rate (KBps)	20-250	64-128+	11000	720
Transmission Range [m]	10-1600	1000+	1-100	1-10+
Success Metrics	Reliability, Power, Cost	Reach, Quality	Speed Flexibility	Cost Convenience

- *The second, the system must support a certain throughput, a number of sensor data packets per second.* In low rate ITS applications only a few data packets are expected to be transmitted, so that performance criteria will be satisfied. BER (Bit Error Rate) is expected to be equal and better than 10^{-6}.
- *The third, these data packets are only useful if they are received within a maximum latency period.* It must be verified for every project.
- *The fourth, many systems must operate under severe conditions and intrinsic safety restriction*.

All of four mentioned targets must be satisfy to continue in further design evaluation. For detailed information read [8].

VI. Experimental Part

Figure 4 presents sensor node developed for measuring of traffic flows parameters. Referring to sensing of magnetic field changes and vibration, sensor LSM303 is used. Magnetometer measurement interval is adjustable in range from ±1.3G to ±8.1G. Acceleration-meter range is possible to adjust in interval from ±2g to ±8g. Signal processing and control unit is realized on the basis of 32-bit microcontroller STM32F100 in small pocket LQFP48. MCU is realized on the basis of ARM-Cortex M3 core with maximal frequency 24MHz. SRAM memory capacity is from 4 to 8 KB. Flash memory capacity is from 16 to 128 KB. Memory subsystem is extended by microSD cart whose allows saving of big data content. Basic PCB consists of connector for connection of communication unit. This solution allows to experiment with various communication modules. XBee PRO communication module in 2.4GHz band was used. Sensor node is realized on 2-layer PCB with dimensions 32x38 mm.

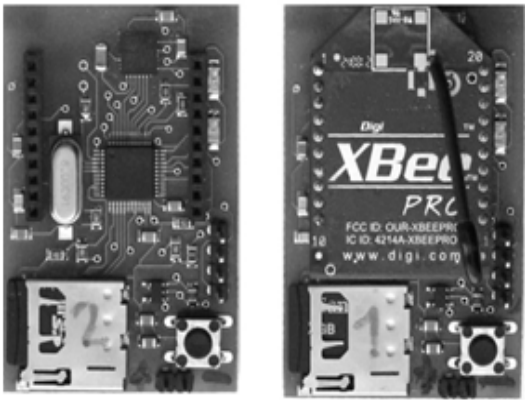


Fig4. Wireless sensor nodes

Proving functionality of the nodes, average speed of vehicle calculated on the basis of measured signals of magnetic induction is presented. Two nodes were well-situated at road side in the distance of 10m from each other. Synchronization of them was secured by RF communication. Both of sensors were sensing changes of Earth magnetic field in axes x, y and z depending on car detection, in case of sampling period 50ms. Measurements were transmitted to PC via communication module XBee. All measured data including time mark were stored on microSD card in the same time. Magnetometer measurement range was set to value ± 2.5 G. Direct components of waveforms were filtered. Cars moving about 1.5m far away from sensor nodes evoked changes of magnetic field induction in order of ten-miliGauss (mG). Amplitude of magnetic induction range is related to the distance between vehicle and a sensor.

Time difference between measured signals is depended on localization of sensor elements and speed of passing vehicle. Average speed of vehicle passing the sensors can be calculated on the basis of the evaluation of time difference between corresponding changes of magnetic induction measured by sensors 1 and 2. Comparing behaviours of magnetic induction both of sensors can be used to filter incommensurate signal elements related to other influences than vehicles movements (pedestrians on pavement, inverse direction moving vehicles, etc.). Average vehicle speed could be stated on the basis of total magnetic induction change (S) calculation:

$$S = \sqrt{(x2 + y2 + z2)} \quad . \tag{1}$$

Figure 5 shows plots of total magnetic induction change for sensor1 and sensor2. Average speed of vehicle between sensors1 and 2 can be derived on the basis of measured signals of magnetic induction. Time shift of maximal induction change between sensor1 and 2, for the first vehicle, is 1.28s. Average speed of first vehicle is (for 10m distance between sensor1 and 2):

$$v = 10/1.28 = 7.8 \quad [m/s] \quad . \tag{2}$$

In case of the second vehicle, for its average speed is valid:

$$v = 10/1.01 = 9.9 \quad [m/s] \quad . \tag{3}$$

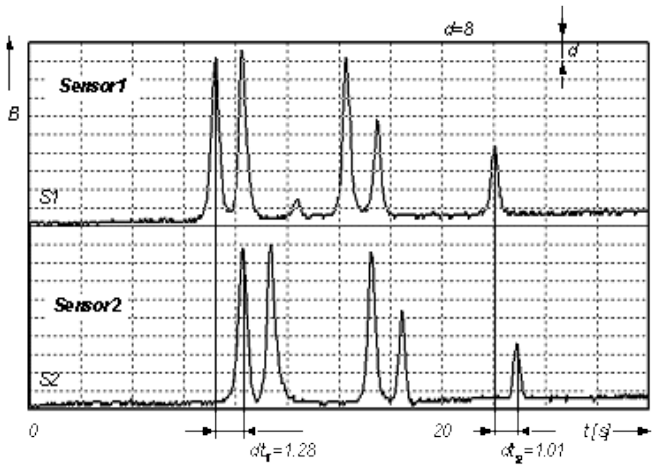


Fig5. Plots of total change of magnetic induction for sensor1 and sensor2

For more precise calculation of time shift could be used more sophistic method of estimation error elimination. It is

necessary to notice that the precision is always limited by real period of sampling frequency. Precision of vehicle average speed calculation between two nodes could be increased by decreasing sampling frequency as well as increasing of distance between sensors.

Presented measurements illustrate usability of sensor nodes with magnetic induction sensing for monitoring of traffic flow parameters. Vehicles classification is presented in literature [9]. Firmware reprogramming is described in [10].

VII. Conclusion

Presented paper is focused on design of WSN road transportation systems for parking management, monitoring of traffic flows, emergency vehicles, weather and environmental conditions. The design must meet many targets and requirements principally discussed through the paper. The system must be value added and complementary to V2X technology as well as ADAS system.

Acknowledgment

Centre of excellence for systems and services of intelligent transport II., ITMS 26220120050 supported by the Research & Development Operational Programme funded by the ERDF.

"Podporujeme výskumné aktivity na Slovensku/Projekt je spolufinancovaný zo zdrojov EÚ"

References

[1] EC: “EU Transport in figures”, *Statistical pocketbook*, ISBN 978-92-79-19508-2, 2011
[2] Junko Yoshida: “Counter Argument: 3 Reasons We Need V2X”, *EETIMES*, 2013
[3] Junko Yoshida: “If a Car´s ReallyAutonomous, Why V2X?“, *EETIMES*, 2013
[4] ETSI EN 302 571 V1.1.1, September 2008: ITS; Radiocommunications equipment operating in the 5 855 MHz to 5 925 MHz frequency band;
[5] *http://www.bb-elec.com/Learning-Center*
[6] Juraj Miček, Ján Kapitulík: “Transmit power control analysis in V2X communication system”, *Communications: scientific letters of the University of Žilina*, Vol. 12, No. 3A, s. 55-59, 2010, ISSN 1335-4205
[7] *http://www.zigbee.org*
[8] Lance Doherty, Jonathan Simon, Thomas Watteyne: “Wireless Sensor Network Challenges and Solutions”, Linear technology, WP001, *http://www.linear.com*
[9] Ondrej Karpiš: “System for Vehicles Classification and Emergency Vehicles Detection”, Proceedings of 11th IFAC/IEEE International Conference on Programmable Devices and Embedded Systems, PdeS, pp.155-159, Brno University of Technology, 2012
[10] Ondrej Karpiš: “Software actualization in Wireless Sensor Networks”, Proceedings in Information and Communication Technologies - International Conference, ICTIC 2012, pp.51-54, EDIS-ŽU, 2012

Európska únia
Európsky fond regionálneho rozvoja
Operačný program VÝSKUM a VÝVOJ
Agentúra Ministerstva školstva, vedy, výskumu a športu SR pre štrukturálne fondy EÚ

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1023–1026
DOI: 10.15439/2014F310
ACSIS, Vol. 2

WSN for Traffic Monitoring using Raspberry Pi Board

Michal Kochláň, *IEEE Student Member*
University of Žilina
Faculty of Management Science and Informatics
Univerzitná 8215/1, 010 26 Žilina, Slovakia
Email: michal.kochlan@fri.uniza.sk

Michal Hodoň, Lukáš Čechovič,
Ján Kapitulík, Matúš Jurečka
University of Žilina
Faculty of Management Science and Informatics
Univerzitná 8215/1, 010 26 Žilina, Slovakia
Email: {michal.hodon, lukas.cechovic, jan.kapitulik,
matus.jurecka}@fri.uniza.sk

***Abstract*—This paper introduces low-cost non-intrusive sensory that can collect traffic data based on Raspberry Pi single board computer. Image information acquired by Raspberry Pi HD camera module is analyzed for moving objects presence. After evaluation of detected object count, size, class and motion vector object properties are sent to server node by RF transceiver. Sensor low-power consumption ensures possibility to operate from battery for an extended period of time.**

I. Introduction

TRAFFIC flow monitoring and analysis has been active research and engineering topic for more than two decades. Main information acquired from traffic flow monitoring includes: traffic volume, vehicle type identification (bike, car, light van, truck) and vehicle speed. Traffic volume data is used for a variety of purposes including historical trend analysis, forecasting, planning for future infrastructure improvements and expansions. Whereas transport remains the largest producer of CO emissions in EU, traffic monitoring becomes important also from the environmental point of view [1]. Also the World Health Organization has officially decreed that inhaling diesel fumes can cause lung cancer and puts diesel plumes in the same category as arsenic, strontium-90 and neutron radiation. [2] This has given traffic monitoring significant importance.

Other traffic data parameters, such as speed and vehicle classification, are becoming more important as a measure of traffic safety and roadway pavement use. Recent traffic flow analysis systems are able to perform vehicle number plate recognition which can provide information about main ways of traffic flow through cities and can help to optimize road infrastructure. Collecting this data can be done using a variety of different technologies. Traffic detection technology methods scoring biggest interest in this area includes: Doppler radar (measures the relative velocity of an object moving through its target range), magnetometer sensors (detects vehicles based on the disruption of the Earth's magnetic field by metal vehicles), video camera (processes images using sophisticated computer algorithms), side-fire radar (side-fire beams placed along a roadway reflect back to the sensor to detect vehicles), pneumatic tubes (transmits information to a counting device after a pulse is created when vehicles drive over a tube).

With the advent of powerful single board computers like: OLinuXino, Galileo, PandaBoard, Raspberry Pi, Odroid and others it is possible to design reliable, low-cost traffic monitoring system. Low-power consumption of these boards ensures possibility to operate from car battery for an extended period of time often more than one week.

This paper introduces low-cost non-intrusive option that can collect traffic data based on Raspberry Pi. We have chosen this board for its easy and powerful HD camera handling, good performance to power consumption ratio affordable price and wide community.

II. System overview

Traffic flow monitoring system consists of Raspberry Pi based sensor nodes used for acquisition of image information from environment and their analysis. Traffic flow parameters including volume, speed and vehicle class are then sent to main server via RF transceivers [5]. RF connection offers also configuration and diagnostics of sensor nodes.

The main priorities while designing the traffic flow monitoring sensors were:

- low cost of sensors
- low power consumption of sensors
- high reliability

"Fig. 1" shows complete Raspberry Pi based sensor node including camera module, 12V battery 9Ah, stem down 5V converter and ARDUPILOT 3DRADIO RF transceiver (without antenna).

Rasberry Pi single board computer offers these key features [3]:

- ARM1176JZF-S core CPU @ 700 MHz
- Broadcom VideoCore IV GPU @ 250 MHz
- 512 MB (shared with GPU)
- 5Mpix Camera module capable of full HD video @ 30fps
- USB, GPIO, UART, I2C and SPI bus
- 700 mA (3.5 W) power rating

This work was supported by Competence Center for research and development in the field of diagnostics and therapy of oncological diseases", ITMS: 26220220153, co-financed from EU sources and European Regional Development Fund

978-83-60810-58-3/$25.00

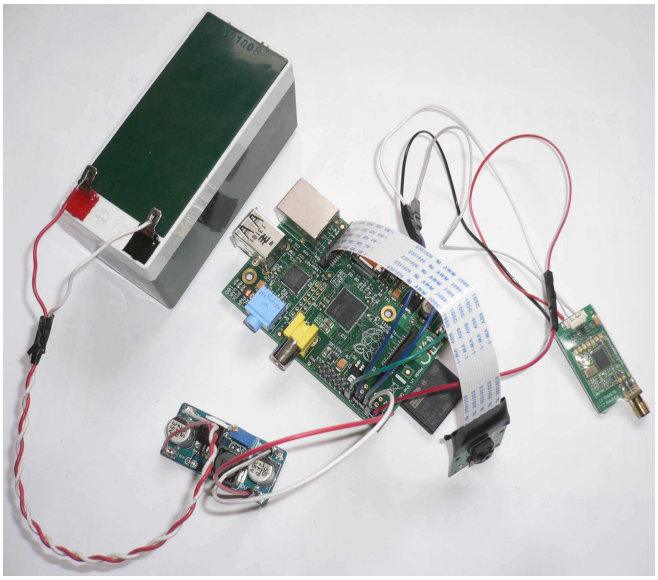


Fig. 1. Raspberry Pi based sensor node

Since processing of high-definition video streams for real-time applications is a challenging task with computationally demanding algorithms it was necessarily to perform as much as possible image processing tasks on GPU. Modern GPUs are very efficient in dealing with computer graphics and their highly parallel structure makes them more effective than general-purpose CPUs for algorithms where processing of large blocks of data is done in parallel.

Raspberry Pi GPU resources can be accessed by Multi-Media Abstraction Layer (MMAL) C library which has been used to implement video data acquisition and processing. Description of the video data processing is shown on "Fig. 2"

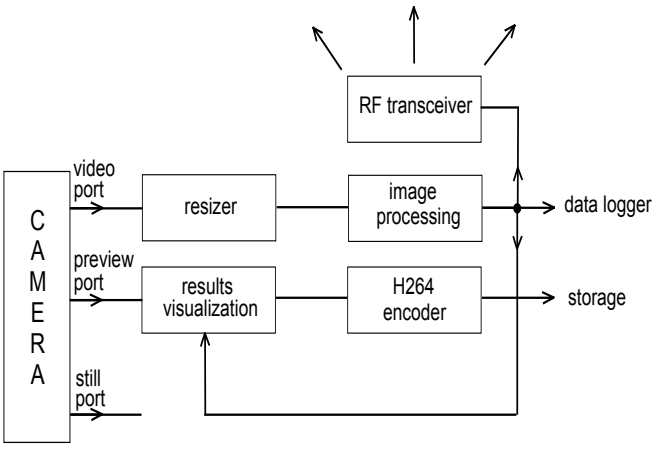


Fig. 2. Video data processing overview

Camera has three output ports (video, preview and still port). Due to the hardware limitations all of these ports must be set to the same resolution, this was in our case 1280x720 at 30 frames per second. Camera still port was not used. Since sufficient resolution for reliable vehicle detection is much lower (640x480 has been used in our system) image processing block has been connected to camera port through resizer block. The results of image processing – detected vehicles are visualized by modifying (drawing into) data stream flowing from camera preview port. Encoder block provides compression of raw YUV video data stream into H264 steam which is stored on Raspberry PI's SD card. For demonstration of system functionality encoder block can be easily replaced by previewer block. Showing the traffic live stream with recognized results and statistics is then possible through HDMI. Traffic flow statistics are stored on SD card and also send by ARDUPILOT 3DRADIO RF transceiver [7] to monitoring center main server for further evaluation.

III. Image Processing

Image processing block ("Fig. 3") performs moving objects detection and classification.

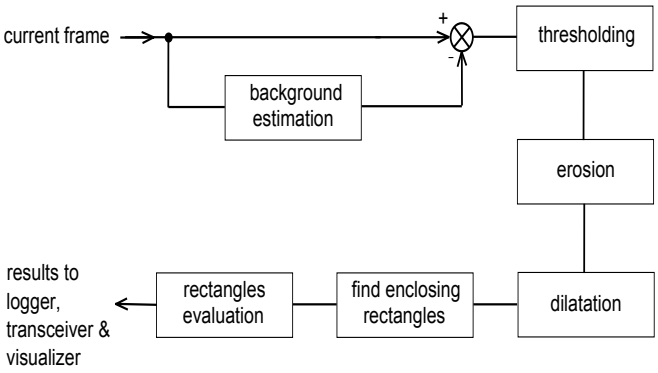


Fig. 3. Image processing block overview

The most important procedure in the moving objects detection task is background estimation. Detection of moving objects is made from the difference between the current frame and a reference frame – background image [8]. Background subtraction is a widely used approach for detecting moving objects in videos from static cameras. Our background estimation model is based on the idea that the small changes in frame sequence (e.g. illumination changes) should be retained while relatively fast changes caused by e.g. moving object should be suppressed. Background image is periodically updated according the following equation:

$$\underline{B}(m) = \underline{B}(m) - f[\underline{F}(m) - \underline{B}(m)] \tag{1}$$

where $\underline{B}(m)$ and $\underline{F}(m)$ are background and current frame matrices in time m, function $f(\cdot)$ can be described by following equation:

$$f(x) = \frac{\gamma \cdot \frac{x}{\alpha}}{\left[1 + \left(\frac{x}{\alpha}\right)^2\right]^2} \tag{2}$$

where function argument x is the difference between current frame and background image pixel values, function parameters were set as follows $\alpha = 64$, $\gamma = \frac{25}{2}$.

Graph of function $f(\cdot)$ is shown on "Fig. 4"

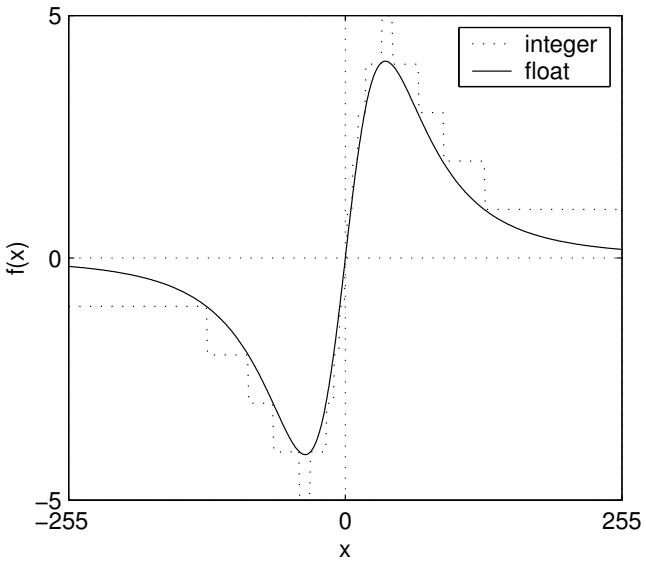


Fig. 4. Image processing block overview

It is clear that the background image function $f(x)$ is raising nearly linear for small x values but for big values of x is decaying.

This background image function has been implemented in integer (shown by dotted line in "Fig.4") which significantly helps to improve speed of whole system.

After subtraction of estimated background image ("Fig.6") from current frame ("Fig.5") the thresholding has been applied so black & white image has been obtained ("Fig.7").

Fig. 5. Original (current) frame

Since most cameras produce a noisy image, motion has been detected in such places, where there is no motion at all. To remove random noisy pixels, erosion filter followed by dilatation has been used. So the resulting image contains only the regions where the actual motion has been present.

Fig. 6. Background image

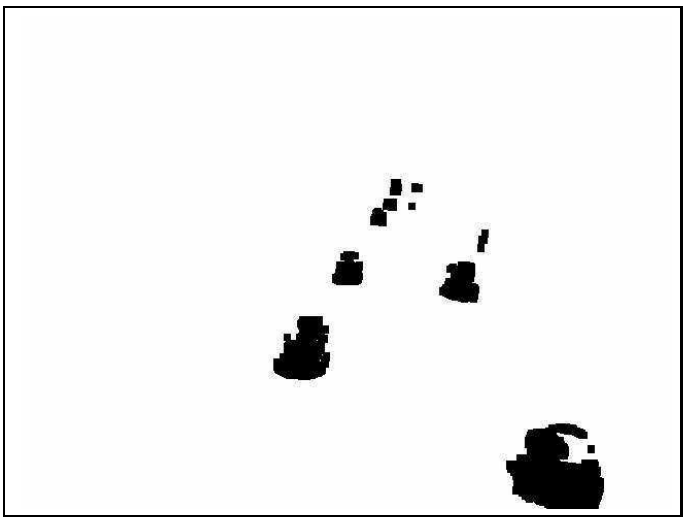


Fig. 7. Image after thresholding

Motion regions are then bounded into minimal enclosing rectangles. These rectangles are then the subject of evaluation where size, position and motion vector is being calculated. Traffic flow informations obtained are stored on SD card, visualized into encoded live video stream ("Fig.9") and transmitted to server node using RF transceiver.

IV. Experiments & results

Described sensor node was subject to one day real traffic monitoring test.

It has been attached to the bridge construction over the road and it was monitoring traffic flow in both direction. Power was supplied from 12V 12Ah rechargeable sealed lead acid battery.

feature	recogniton correctnes in [%]
traffic volume	95.7%
vehicle class	93.2%
vehicle speed	not evaluated

Fig. 8. Recognition correctness

Results obtained are summarized in “Fig.8” Above-mentioned table presents promising recognition results of traffic volume and vehicle class.

Fig. 9. Visualization of recognition result

Unfortunately vehicle speed was not evaluated because of non-functional microwave speed detector which was intended to be the reference for speed comparation. Sensor average current draw during the test was 440mA.

V. Acknowledgement

This work was supported by:

– project "Competence Center for research and development in the field of diagnostics and therapy of oncological diseases", ITMS: 26220220153, co-financed from EU sources and European Regional Development Fund.

"Podporujeme výskumné aktivity na Slovensku/Projekt je spolufinancovaný zo zdrojov EÚ"

– Faculty of Management Science and Informatics institutional grant.

VI. Conclusion

An approach for the traffic monitoring based on Raspberry Pi single board computer was proposed in this paper. Video signal processing and analysis were proposed with the focus on low power consumption and reliability of the system. This is still far from being final version of traffic monitoring system and there is lot for us to improve. In future we plan to perform further experiments including speed evaluation and testing sensor node in different weather conditions.

References

[1] The Harmful Effects of Vehicle Exhaust, webpage: http://www.ehhi.org/reports/exhaust/summary.shtml [April 19, 2014]
[2] World Health Organization says diesel fumes cause cancer, webpage: http://www.autoblog.com/2012/06/13/world-health-organization-says-diesel-fumes-cause-cancer [June 19, 2014]
[3] Raspberry Pi, webpage: http://en.wikipedia.org/wiki/Raspberry_Pi [April 20, 2014]
[4] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing,* New York: MPC, 1992.
[5] J. Miček and O. Karpiš, “Wireless sensor networks - design of smart sensor node,” *International conference on military Technologies ICMT 2011,* Brno, pp. 1109–1116.
[6] O. Karpiš and J. Miček, “Sniper localization using WSN,” *International conference on military Technologies ICMT 2011,* Brno, pp. 1063–1068.
[7] O. Karpiš, “Software actualization in WSN networks,” *Information and Communication Technologies - International Conference,* Žilina, 2012, pp. 51–54.
[8] M. Piccardi *Background subtraction techniques: a review*. IEEE International Conference on Systems, Man and Cybernetics 4., 2004 pp. 3099-3104.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1027–1034
DOI: 10.15439/2014F299
ACSIS, Vol. 2

2.4GHz ISM Band Radio Frequency Signal Indoor Propagation

Michal Kochláň, *IEEE Student Member*
Department of Technical Cybernetics
Faculty of Management Science and Informatics
University of Žilina
Univerzitná 8215/1, 010 26 Žilina, Slovakia
Email: michal.kochlan@fri.uniza.sk

Juraj Miček, Peter Ševčík
Department of Technical Cybernetics
Faculty of Management Science and Informatics
University of Žilina
Univerzitná 8215/1, 010 26 Žilina, Slovakia
Email: {juraj.micek, peter.sevcik}@fri.uniza.sk

***Abstract*—Indoor environment is from the point of the wireless communication an extremely hostile environment. Despite this fact, wireless sensor network applications in the indoor environment are very common. Having signal propagation in a real environment, without considering interferences from other sources, we meet (not only indoors) four basic phenomenons - path-loss, reflection, diffraction and scattering. Each of these effects impact on the spread of the signal and contributes to attenuation and distortion at the receiver side. Detailed description of the electromagnetic wave propagation can theoretically obtain the solution of Maxwell's equations. However, this is too demanding and for practical cases unusable. In practice, to describe the signal propagation, the approximate models are used, which are often based on experimental results. This contribution includes case study on indoor radio frequency signal propagation at 2.4GHz ISM Band with related math, supported by implementation of the propagation models and experimental results.**

I. Introduction

INDOOR wireless sensor network (WSN) applications, despite unfavorable conditions of indoor radio signal propagation, constitute the very attractive area. As an introduction, let us mention at least some of the frequently occurring applications that the authors consider representative.

A. Intelligent Buildings

In recent decades, an increased attention to matters related to sustainable environment has been dedicated. A large part of the environmental problems is related to the energy consumption of the society. It is worth to emphasize that the effective regulation of the overall energy consumption of the society is able to achieve significant advances in the state of the environment [1].

One of the dominant energy consumption components is the operation of residential and non-residential buildings. According to [2], in 2012, the energy consumption of buildings in the U.S. was 39%, followed by 32% energy consumption share of an industry and transportation sector with portion of 29% of the total consumption of the energy. In Europe, the situation is quite similar. In 2010, the buildings operation took 41% of total energy consumption, 32% in transportation and an industry energy consumption shared 25% [3].

Remark 1.1: Note that compared to the U.S., the energy consumption intended for buildings operation in Europe is slightly higher. Also, an interesting fact is that the annual energy consumption of the residential buildings is about 200 kWh/m^2, non-residential buildings are characterized by higher annual energy consumption of about 300 kWh/m^2.

Since 1990, the energy consumption increases annually by around 0.6% in the residential sector and by around 1.5% in the non-residential sector [2], [3]. The presented statistical data show that the highest energy burden in the developed countries lies in operation cost of buildings. Based on the predictions presented in [2], it is expected that within the next 20 years, the energy consumption share of the buildings will not be reduced, just to the contrary, we can assume a slight increase in the energy consumption (increase in about 1%).

Based on the above, it is clear that the operation of buildings consumes huge amounts of energy. Therefore, it is extremely important to address the issues of intelligent buildings control so that the user comfort and the effective energy sources utilization can be ensured.

In order to be able to reduce the energy consumption, it is necessary to know why, where and when the energy consumption occurs. Finding the answers to these questions is possible thanks to the new technologies in the field of ICT (Information and Communication Technologies) such as [4], [5] or [6]. Using new ICT means in buildings we get intelligent monitoring and control systems of the buildings that enable increasing user comfort while cutting energy and environmental burden. Buildings equipped with modern control and monitoring systems are often labeled as "intelligent" or "smart".

An extensive control, communication and monitoring system installed in modern intelligent buildings today is often divided into the following six subsystems [1]:

- *Lighting control* subsystem (lights, blinds, etc.);
- *Heating, Ventilation and Air Conditioning (HVAC)* subsystem - creates a psychometric chart to help determining the optimal environmental parameters;
- *Security and safety* subsystem (entrance authorization,

This work was supported by Centre of excellence for systems and services of intelligent transport II. ITMS 26220120050 supported by the Research & Development Operational Programme funded by the ERDF.

fire alarm, personnel tracking, etc.);
- *Metering* subsystem (electricity, gas, water and other parameters with connection to the energy control);
- *Indoor climate monitoring* subsystem (temperature, humidity, dust, concentration of CO_2, NO_x ("Green Building Monitor^{TM}" - Siemens system to inform staff, clients and visitors about the energy consumption, environmental load of the building and about the state of the building indoor climate);
- *Guest control* subsystem (navigation, information for visitors, access control, motion tracking, etc.).

It is obvious that the provision of these subsystems implies the utilization of the latest technologies in the field of ICT. It is necessary to sense a large number of parameters in the intelligent buildings and based on the values, the actuator elements drive the building so that the optimal operation of the building is assured [1]. Sensors gathering the information create an extensive network with defined rules of communication. Nowadays, more and more, we meet the wireless communication technologies in the area of building control, which, in comparison to a wired network, is characterized by lower installation cost, higher flexibility and scalability [7]. Thus, even in the intelligent building environment, there are gradually used technologies known under the names of WSN or WSAN (Wireless Sensor and Actuator Networks).

B. WSN in the Industry (Industrial Automation)

Industrial applications represent another promising area of an efficient use of the WSNs [8]. Industrial control systems that integrate WSAN offer several advantages over conventional distributed control system [9], [10]. These are in particular easiness of the sensors' and actuators' installation, self-organization of the network, a simple modification, easy expandability of the network, efficient distributed and parallel data processing and lower cost compared to the conventional solutions [11], [12], [13]. There are also some disadvantages of wireless solutions, which we can find in the occurrence of interferences, not only in the industrial environment, unpredictable delays in packets, limited capacity of transmission channels and so on [9], [11]. However, the mentioned drawbacks, are not substantially limiting for the development and the implementation of WSN applications in industry. According to [9], it is essential to address, in particular, the following issues:

- Limited sources of the energy, the limitation of communication attributes, computing and storage capacity of the network nodes (channel capacity, limited communication range, etc.);
- Network operation in high interference industrial environments and related dynamic change of network topology;
- High demands on Quality-of-Service (QoS), especially requirements on packet deliverability in a defined time;
- Effective utilization of data redundancy to increase system reliability and accuracy of the status information of the controlled/monitored system;
- Effective addressing of the related issues to the error rate (BER $\in \langle 10^{-2}; 10^{-6} \rangle$) and variable transmission capacity of individual transmission channels (adaptive modulation schemes, channel coding, etc.);
- Security in industrial applications is a key issue. It is necessary to protect communication from intentional active and/or passive attacks;
- Large-scale deployment and ad-hoc network functionality. Many industrial applications consist of a large number of randomly distributed nodes, so it is advantageous if the network is able to build autonomously the communication links and to control the communication among the nodes;
- Integration with other networks (e.g. the Internet). This point is very important from the perspective of effective control of the enterprise approach to the technological level, which represent also a WSN.

In the field of the industrial automation and in the context of WSAN implementation, there are arising new problems (distributed processing of variables, redundancy utilization to increase system reliability, problems with delayed packets), whose solutions are interesting from the theoretical and application point of view [9]. Based on the current development, a sharp increase in wireless solutions even in industrial applications is expected [14] [15], [16], [17].

C. Health Applications and Senior Assistance Services

Electronic health-care is a broad and interesting application area of WSNs. Let us mention a few representative examples of such WSN utilization [11]:

- Monitoring of vital signs and other selected parameters of patients in hospitals;
- Telemonitoring of the patients without hospitalization;
- Tracking patients, visitors and hospital staff;
- Indoor climate monitoring within the hospital premises;
- Controlling access to medicines and identification of time and kind of medication use through patient node(s).

WSN networks are able to monitor the behavior of older people and those with disabilities and enable to keep track of their health status without significant restrictions on their lives and quickly identify the signs of disease [18], [19]. Because of the above examples, it is clear that in health-care we meet growing number of interesting WSN applications indoors and outdoors - modern applications of WSN that monitor vital signs of a human body, track patients, monitor hospital environment, serve as medical access control systems and many more [20], [21].

Almost all mentioned applications have one thing in common, WSNs operate in indoor environment. Indoor environment is from the perspective of the wireless communication subsystems, an extremely hostile environment [7]. Despite this fact, WSN applications in indoor environments are highly desired.

Further investigations dealing with the properties of the indoor signal propagation assume the wireless sensor node based on Texas Instruments' (TI) CC2511 transceiver that is highly suitable for the indoor environment implementation.

It is possible to list a number of other interesting and prospective application areas of WSNs related to the currently popular term "Internet of Things" (IoT), but due to the limited extent, we do not.

II. Indoor Signal Propagation

If we consider wireless transmission channel for evaluation, we come to the conclusion that its features do not fully meet the media properties for reliable high-speed communication. The transmission channel is sensitive to noise, interference generated by the obstacles, communication distance, etc. [22]. Moreover, those adverse effects vary in time and space at random, as a result of the change in position of the receiver and/or transmitter and dynamic environmental changes.

Having signal propagation in a real environment, without considering interferences from other sources, we meet four basic phenomenons [22]. These are:

- Path-loss;
- Reflection;
- Diffraction;
- Scattering.

Each of these effects impact on the signal propagation and contributes to the path-loss and the distortion of the received signal. Answering the question of which of the above-mentioned effects has the most significant impact on the quality of reception is not possible in general. The answer depends on the wavelength, the environment type and many other specific conditions. It is obvious that the electromagnetic waves do propagate in the real environment attenuated, reflected, diffracted and scattered from and by the terrain, buildings and other objects. The detailed description of the propagation of electromagnetic waves can be theoretically obtained by solving Maxwell's equations with constraints [23]. The detailed description is only possible in case of having all physical characteristics of the objects affecting the propagation in mind and implemented into the equations. Unfortunately, this is computationally demanding and in practical cases not usable. Therefore, in practice, approximate models, which are often based on experimental results, are used to describe the radio frequency wave propagation.

For determining the signal path-loss in the line-of-sight (LOS), a sufficient solution is the radio signal propagation model at outdoor environment that assumes that other propagation effects can be neglected [23].

Remark 2.1: Other propagation effects can be neglected when an open area transmitter and/or receiver antennas are placed at a sufficient height above the ground.

In this case, it is possible to determine the signal power at the receiving antenna out of the Friis equation:

$$P_r = A_e S_r = \frac{\lambda^2}{4\pi} G_r \cdot \frac{G_t P_t}{4\pi d^2} = \left(\frac{\lambda}{4\pi d}\right)^2 G_r G_t P_t, \qquad (1)$$

where,

d is *distance between the transmitter and the receiver*;
λ represents *signal wavelength*;
P_t represents *transmission power*;
P_r represents *power at the receiving antenna*;
G_t represents *transmitting antenna gain*;
G_r represents *receiving antenna gain*;
$A_e = G_r\lambda^2/4\pi$ is *receiving antenna effective area*;
$S_r = G_t P_t/4\pi d^2$ is *signal power density at the receiving antenna*.

If the propagating radio waves reach an surface that is larger than the radio wavelength, then the wave partially reflects and partially penetrates the obstacle material. In case the material is the perfect conductor, the wave reflects into the primary environment without the energy loss. Of course, in this case the reflection law applies. In general, the electromagnetic field strength of the reflected and passing waves can be expressed by the Fresnel reflect and transfer coefficients that depend on the material properties, the wave polarization, the angle of radio wave impact and the frequency. A more detailed analysis of the effect of reflection can be found in [11]. The reflection significantly affects the radio frequency signal propagation in the indoor environment.

The diffraction occurs when there is no direct path (line-of-sight) for the radio waves between the transmitter and the receiver antennas - the transmission path is limited by an obstacle. Due to the diffraction, in the shade of the perfect obstacle with the limited dimensions, the radio signal strength has a non-zero value. The effect of the signal diffraction can be neglected at high frequencies.

Signal scattering is present in cases where the electromagnetic waves pass through or reflect from the objects that have comparable or smaller dimensions than the wavelength. Scattered waves occur when rough surfaces and small objects appear in the space of the transmission channel. The signal level varies at the receiver antenna because of the signal scattering and the levels are usually different from the values predicted by the propagation models that even take into account the path-loss, reflection and diffraction.

Remark 2.2: Note that the larger signal energy is scattered, the more the energy of the reflected waves is reduced. All the above mentioned effects, affect the propagation of the electromagnetic waves indoors. The way how they impact depends on the signal wavelength and the environmental properties and circumstances.

III. Case Study

To investigate the case of the electromagnetic wave propagation we used the hallway scenario illustrated at the Fig. 2. The antennas (the transmitter and the receiver antenna) have been placed at the center line of the corridor in the height $80cm$ above the floor.

For measurements we used the system-on-chip solution based on TI CC2511 (see Fig. 1) that operates in the ISM band starting at the frequency equal to $2.400GHz$ and ending at $2.4835GHz$. This ultra-low power solution integrates a full-speed USB controller, I^2S interface USART and 12-bit A/D converter. However, more important information for us are the radio peripheral attributes. Besides the mentioned frequency range, the radio solution is capable of 2FSK, GFSK and MSK

Fig. 1. Wireless system-on-chip used for experimental results

modulation techniques, programmable transmission power up to $+1dBm$, high receiver sensitivity ($-100dBm$ in average) along with the 128-bit AES security co-processor.

In the first step, we will try to describe the propagation of the electromagnetic waves with Five-Ray propagation model (see Fig. 3), in which we consider a simple reflections from the side walls, floor and ceiling of the hall.

Then the impulse response function of the transmission channel can be written in the form:

$$h(t) = A_1 \cdot \sigma(t - T_1) + A_2 \cdot \delta(t - T_2) + \\ + A_3 \cdot \delta(t - T_3) + A_4 \cdot \delta(t - T_4) + A_5 \cdot \delta(t - T_5) \quad (2)$$

where an expression $A_1 \cdot \sigma(t - T_1)$ represents a received signal portion (from line-of-sight);

expressions $A_2 \cdot \delta(t - T_2)$, $A_3 \cdot \delta(t - T_3)$ and $A_4 \cdot \delta(t - T_4)$ represent reflected signal portions from the both side walls and the hallway floor;

$A_5 \cdot \delta(t - T_5)$ represents a reflected contribution from the hallway ceiling.

Given the same length of the transmission path, it is possible to assume that the reflections from the both side walls and the floor have the same signal time delay (T_2=T_3=T_4). Then, it is possible to rewrite (2) into the following form:

$$h(t) = A_1 \cdot \sigma(t - T_1) + A_v \cdot \delta(t - T_2) + A_5 \cdot \delta(t - T_5) \quad (3)$$

where $A_v = A_2 + A_3 + A_4$.

Then, the frequency response of the transmission channel is:

$$H(j\omega) = A_1 e^{-jT_1\omega} + A_v e^{-jT_2\omega} + A_5 e^{-jT_5\omega}. \quad (4)$$

For the times T_i aplies:

$$T_1(d) = \frac{d}{3 \cdot 10^8} \quad [s], \quad (5)$$

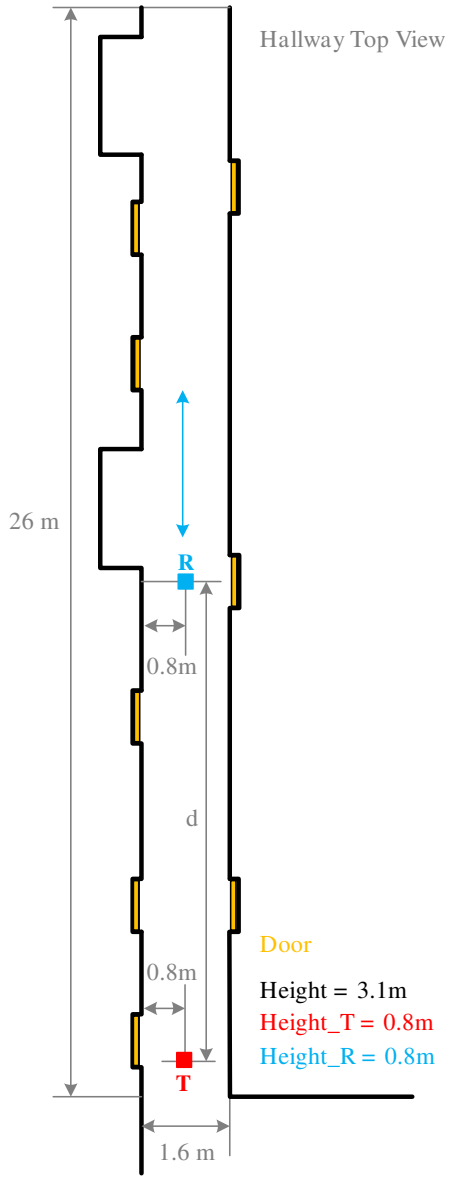


Fig. 2. Topology of the measurement indoor environment

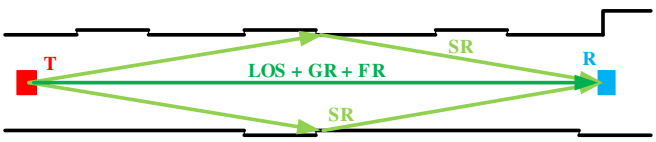


Fig. 3. Five-Ray propagation model illustration

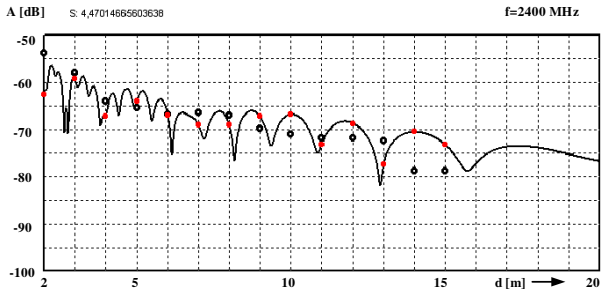


Fig. 4. Transmission power dependency on the distance d

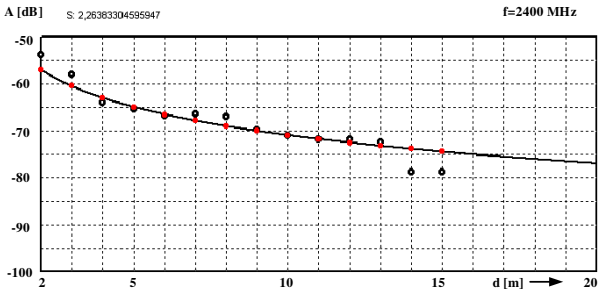


Fig. 5. Transmission power path-loss dependency on the distance d

$$T_2(d) = \frac{\sqrt{d^2 + 4 \cdot 0.8^2}}{3 \cdot 10^8} \quad [s], \tag{6}$$

$$T_5(d) = \frac{\sqrt{d^2 + 4 \cdot 2.3^2}}{3 \cdot 10^8} \quad [s], \tag{7}$$

With respect to the (1) and if assuming that the antennas gain equal 1, it is possible to write for A_1 the following:

$$A_1 = C \cdot \left(\frac{\lambda}{4\pi d}\right)^2, \tag{8}$$

where the constant C depends on the antennas gain (transmitting and receiving antenna).

For all reflected radio waves applies:

$$A_i = K_i \cdot C \cdot \left(\frac{\lambda}{4\pi d_i}\right)^2, \tag{9}$$

where the coefficient K_i is the reflection coefficient that depends on reflection plane material, reflection plane surface, the frequency, etc.)

Based on the mathematical representation of the Five-Ray signal propagation model, the experimental results have been acquired. These results show the dependency of the channel path-loss on the distance between the receiver and the transmitter. The graphical representation of the data with reflection coefficient values set to 0.5 and the frequency $2400MHz$ is shown on the Fig. 4.At the same time, the Fig. 4 shows the measured values in the specific measurement points (black circles). The red dots represent the values predicted by the propagation model at the specific measurement points.

As the measure of the propagation model match with the experimental results, we selected the relationship (10). Having the zero-mean error of the propagation model this relationship represents a selective standard error deviation of the model.

$$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(M_i - M_o)^2}, \tag{10}$$

where M_i represent the measured values at the specific measurement points;

M_{o_i} represent the model calculated values at the specific measurement points.

When comparing the measured values with the values obtained from the Five-Ray propagation model, we received an average standard deviation value equal $S \approx 4.5dB$. We can conclude that the indoor environment is too complex for the transmission path with strong scattering and multiple electromagnetic waves reflections. Therefore, it is impossible to describe these effects using any analytical propagation model, unlike the possibility in case of the urban environment (urban mikro-cell).

Let's analyze an option of describing the electromagnetic wave propagation through the Friis equation to calculate the path-loss of the electromagnetic waves in the free space. Then, the mathematical model describing the signal path-loss dependency on the distance between the transmitting and the receiving antenna will be in the form:

$$A_{dB}(d) = 10log_{10}\left(\frac{P_r}{P_t}\right) = \\ = 20log_{10}(\lambda) - 20log_{10}(d) + K, \tag{11}$$

where P_t is the transmission power;

P_r is the power at the receiving antenna output;

K is a constant representing the antennas' parameters (receiving and transmitting antenna; $K = 20log_{10}(4\pi) - 10log_{10}(G_rG_t)$)

A graphical illustration of the $2400MHz$ electromagnetic wave path-loss dependency on the distance between the transmitting and receiving antenna d is shown in the Fig. 5. Similarly to the Fig. 4, this figure shows the measured values in each specific measurement point (black circles) and the propagation model calculated values (red dots). The propagation model is defined in (11).

Likewise, as in the previous case of the Five-Ray propagation model, for this one we determined the standard error deviation of the model $S = 2.26dB$ based on the model calculated values and measured values.

According to [11], in most cases, we can use a simplified model of the wave propagation in the form:

$$A(d) = K \cdot \left(\frac{d_0}{d}\right)^{\gamma}, \tag{12}$$

TABLE I
TYPICAL VALUES OF γ EXPONENT FOR DIFFERENT ENVIRONMENTS

Environment	γ value
Urban area (macro-cells)	3.7 - 6.5
Urban area (micro-cells)	2.7 - 3.5
Office space (same floor)	1.6 - 3.5
Office space (different floors)	2 - 6
Commerce	1.8 - 2.2
Industrial	1.6 - 3.3
Household	3

$$A_{dB}(d) = 10\ log\ K - 10\ \gamma\ log\left(\frac{d}{d_0}\right), \tag{13}$$

where K is a function of the average environment path-loss and takes into account the characteristics of the antennas;

d_0 is the reference distance that for the indoor environment should be in the interval $1 - 10m$. (for 2.4GHz frequency, it is recommended to set the value d_0 at the lower interval boundary, for outdoor environment d_0 should be from the interval $10 - 100m$).

The value of the coefficient K may be set approximately in the reference distance d_0 based on the Friis relationship:

$$K = \left(\frac{\lambda}{4\pi d_0}\right)^2, \tag{14}$$

Better way of the path-loss value $K(d_0)$ determination in the reference distance for certain conditions is to measure it. Similarly, also the γ value depends on the specific environmental conditions and can be set by measurements. The easiest way is to minimize the sum of difference squares of the measurements in the specific measurement points d_i and model predicted values ($\sum(M_o(d_i) - M_i)^2$), where M_i represent the measured values).

Available literature on propagation models gives approximate values of the γ exponent - Table I.

Considering that the random environmental factors contribute to the path-loss propagation given by the value E_{dB}, then the relationship (9) turns into the following form:

$$A_{dB}(d) = 10\ log\ K - 10\ \gamma\ log\left(\frac{d}{d_0}\right) + E_{dB}, \tag{15}$$

where E_{dB} is a random variable with a Gaussian distribution with zero mean and standard deviation S. Then, based on the experimental measurements it is possible to create the mathematical model (9) that minimizes the sum of difference squares of the measured values and the model calculated values. It can be proved that for this criteria based on the $N + 1$ different measurements in different distances $d_i = 0, 1, 2, ..., N$ it is possible to determine γ value as follows:

$$\gamma = \frac{\sum_1^N \left(K(d_0) - M(d_i) \cdot log_{10}\left(\frac{d_i}{d_0}\right)\right)}{10\sum_1^N \left(log_{10}\left(\frac{d_i}{d_0}\right)\right)^2}, \tag{16}$$

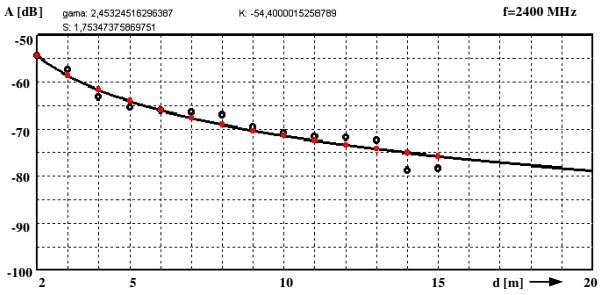


Fig. 6. Transmission power path-loss dependency on the distance d (black circles represent the measured values and the red dots represent the predicted values from the propagation model (17))

where $K(d_0$ is the path-loss in the reference distance d_0 (it can be obtained by measurements eventually it is possible to determine its approximate value according to (10),

$M(d_i)$ is the measured path-loss value at a distance d_i from the transmitter.

Remark 3.1: Note that the above model is applicable only for distances $d > d_0$.

Based on the (11) and (12) a transmission channel model can be written in the form:

$$A_{dB}(d) = C - 10\ \gamma\ log\left(\frac{d}{2}\right), \tag{17}$$

where the reference distance has been chosen $d_0 = 2m$. Fig. 6 illustrates the obtained results.

Similarly, as in the previous cases, based on the measured and calculated values the standard error deviation of the model has been determined $S = 1.753dB$. It is obvious that the present model is the most appropriate for determining approximate path-loss of radio signal propagation in indoor environments.

Remark 3.2: However, in practical implementation of the latter model, prior to implementation, it is necessary to perform sample measurements in the the investigated environment.

During the verification of the model, there have been measurements in 9 frequency channels performed. The investigated channels were spaced by $10MHz$ from the range beginning at $2400MHz$ up to $2480MHz$. The obtained parameters of the propagation model for different frequency channels are shown in Table II.

Remark 3.3: Propagation model used in Table II is in the form: $A_{dB}(d) = C - 10\ \gamma\ log_{10}(d/2), P_t = +1dBm, 2FSKmodulation$.

It is obvious that for each channel in the investigated ISM frequency band, we receive a propagation model characterized by different parameters. This situation arises from the fact that the values obtained during the measurements in each channel are slightly different. The path-loss in the individual channels (1-9) at distances of $2 - 15m$ is illustrated in Fig. 7.

TABLE II
PROPAGATION MODEL PARAMETERS

Channel	Frequency [MHz]	γ	C
1	2400	2.5492	-53.9
2	2410	2.4096	-53.5
3	2420	2.9357	-50,5
4	2430	2.6661	-53.0
5	2440	2.6640	-50.1
6	2450	2.5445	-51.8
7	2460	2.6750	-49.3
8	2470	3.0334	-50.1
9	2480	3.4085	-49.7

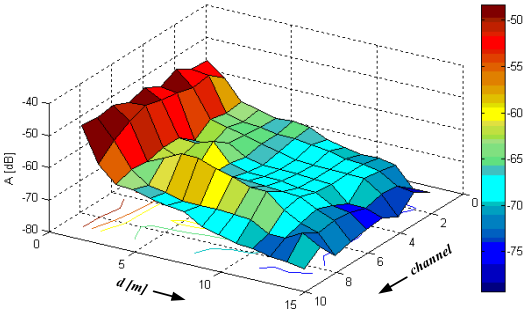


Fig. 7. Measured channel path-loss dependency on the distance d varying from $2m$ to $15m$

The situation where each channel is characterized by a model with optimal γ parameter and C according to Table II is illustrated in Fig. 8. This figure models path-loss propagation in the $2400 - 2480MHz$ frequency range (channels 1-9).

Remark 3.4: Note that the standard error deviation of the model has been $S = 2.46dB$.

Let's simplify rather complicated approach for the modeling of the wave propagation that uses different models for every

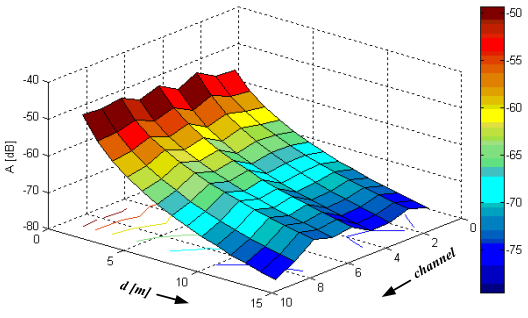


Fig. 8. Modeled channel path-loss dependency on the distance d varying from $2m$ to $15m$

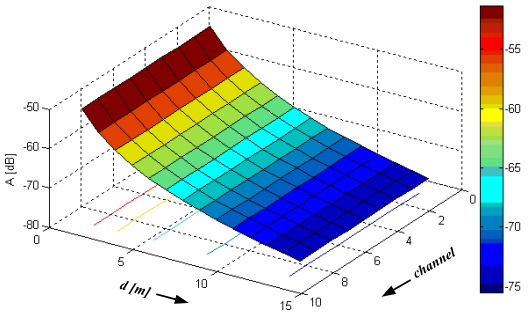


Fig. 9. Modeled path-loss dependency on the distance d varying from $2m$ to $15m$

single channel. Let's try to substitute all 9 previous propagation models with one valid for the entire frequency range $2400-2480MHz$. During the new model description, average values of the coefficients γ_p=2.7629 and C_p=-51.32 are used. Then, the model characterizing the whole frequency range is in the following form (18):

$$A_{dB}(d) = -27.629\ log\left(\frac{d}{2}\right) - 51.32. \tag{18}$$

Fig. 9 is based on the model values defined by (18) and shows the dependency of path-loss propagation in the frequency range $2400 - 2480MHz$ at different distances d.

Remark 3.5: Note that the standard error deviation of the model in this last case is $S = 2.9dB$. It is obvious that there is a slight deterioration in the accuracy of the model compared to the previous one.

IV. CONCLUSION

A comprehensive description of the electromagnetic wave propagation solve Maxwell's equations. However, this approach is computationally demanding and impractical. In real life situations, the approximate models are deployed to describe the signal propagation. These models are often based on the experimental results. This paper shows case study on indoor radio frequency signal propagation at 2.4GHz ISM Band supported by the related math, implementation of the propagation models and experimental results.

To investigate the case of the electromagnetic wave propagation, a specific hallway scenario has been used. The measurements were performed by the system-on-chip solution based on TI CC2511 that operates in the ISM band starting at the frequency equal to $2.400GHz$ and ending at $2.4835GHz$.

Based on the evaluation of the conducted experiments we can conclude that Multi-Ray (Five-Ray) propagation model describing a signal path-loss in indoor environment is unusable. Model based on Friis relationship is indeed easy to apply, requires no measurements, but in the case of indoor environment is applicable only in case of line-of-sight between the receiver and the transmitter. This model does not reflect the

characteristics and topology of the environment and therefore its use is not recommended in indoor environments. Simplified model of propagation of radio waves can be used to determine the approximate communication range. The successful application requires performing the sample measurements of the path-loss.

Let's point out that the radio frequency signal propagation models are dependent on specific environmental conditions. The achieved results cannot be applied directly. It is necessary to carry out experimental measurements of path-loss at multiple locations of protected area by measuring and using the relations (15) and (16) and the optimal model parameters. Then, based on the simulation experiments with propagation models, it is possible to find a suitable deployment of the WSN nodes. Based on path-loss measurements at different distances, we can also conclude that typical signal fading due to the multi-path propagation is caused by greater distortion in the indoor environment and occur to a lesser extent than in the urban areas. However, potential signal loss cannot be suppressed by the frequency channel alternating. By changing the channels (using frequency diversity), it is possible to suppress the effect of interferences from other sources, which are now in the indoor environments strongly present at the $2.4GHz$ ISM band.

ACKNOWLEDGMENT

This contribution/publication is the result of the project implementation *Centre of excellence for systems and services of intelligent transport II.* ITMS 26220120050 supported by the Research & Development Operational Programme funded by the ERDF.

"Podporujeme výskumné aktivity na Slovensku/Projekt je spolufinancovaný zo zdrojov EÚ"

REFERENCES

[1] J. K. W. Wong, L. Heng, and S. W. Wang, "Intelligent building research: a review", *Automation in Construction,* 14.1 (2005), pp. 143–159.

[2] U.S. Department of Energy, "2011 Building Energy Data Book", March 2012.

[3] European Commission, "Report from the Commission to the European Parliament and the Council", Brussels 18.4.2013.

[4] M. Hyben, and M. Hodoň, "Low-cost Command-recognition Device", *Teleinformatics Review,* ISSN 2300-5149, Vol. 36, No. 3, 2013, pp. 19–28.

[5] P. Ševčík, and O. Kovář, "Power unit based on supercapacitors and solar cell module", *SCIECONF 2013 : the 1st international virtual scientific conference*, 10.-14. June 2013: proceedings in scientific conference, ISSN 1339-3561, Žilina: University of Žilina, 2013, ISBN 978-80-554-0726-5, pp. 468–471.

[6] O. Karpiš, "Software actualization in Wireless Sensor Networks", *ICTIC 2012: proceedings in information and communication technologies*, 19-23.3.2012, Žilina: University of Žilina, 2012, ISBN 978-80-554-0513-1, pp. 51–54.

[7] F.L. Lewis, "Wireless sensor networks", *Smart environments: technologies, protocols, and applications*, 2004, pp. 11–46.

[8] O. Karpiš, "System for vehicles classification and emergency vehicles detection", *PDeS 2012: proceedings of 11th IFAC/IEEE international conference on programmable devices and embedded systems*, Brno, May 23th - 25th, 2012, ISBN 978-3-902823-21-2, pp. 155–159.

[9] V.C.Gungor, and G.P.Hancke: Industrial Wireles Sensor Networks: Challenges, Design Principles and Technical Approaches, IEEE Transactions in Industrial Electronics, vol. 56, No.10, 2009.

[10] O. Karpiš, "Solar-cell based powering of a node for traffic monitoring", *IOSR journal of engineering (IOSRJEN)*, ISSN 2278-8719, Vol. 3, No. 4, 2013, pp. 28–32.

[11] A. Goldsmith, "Wireless Communication", *Cambridge University Press*, 2005, ISBN-13 978-0-511-13675-7.

[12] M. Hudik, "Performance optimization of broadcast collective operation on multi-core cluster", *ICSC 2012: tenth international conference on soft computing applied in computer and economic environments*, January 20, Kunovice, Czech Republic, 2012, ISBN 978-80-7314-279-7, pp. 51–55.

[13] J. Púchyová, "Behaviour of multiagent system with defined goal", *Information Sciences and Technologies: bulletin of the ACM Slovakia*, ISSN 1338-1237, Vol. 5, No. 4, 2013, pp. 15–25.

[14] K. S. Low, and M. J. Er, "Wireless Sensor Networks for Industrial Environments", *Computational Intelligence for Modelling, Control and Automation*, 2005, 28-30 Nov. 2005, ISBN 0-7695-2504-0, pp. 271–276.

[15] M. Hodoň, and L. Čechovič, "Hall-efect based sensor for inertial navigation systems", *ICTIC 2013: proceedings in conference of informatics and management sciences*, 25.-29. March 2013, Žilina, Slovak Republic, ISSN 1339-231X, ISBN 978-80-554-0648-0, pp. 359–362.

[16] M. Jurečka, and L. Čechovič, "Application of pulse coupled neural network in speaker identification", *MEMSTECH 2012: perspective technologies and methods in MEMS design*, Lviv, Polyana, Ukraine, 18-21 April 2012, ISBN 978-617-607-229-4, pp. 125–128.

[17] J. Papán, M. Jurečka, and J. Púchyová, "WSN for forest monitoring to prevent illegal logging", *FedCSIS: proceedings of the Federated conference on computer science and information systems*, September 9-12, 2012, Wroclaw, Poland, ISBN 978-83-60810-51-4.

[18] J. Miček, O. Karpiš and P. Ševčík, "Body area network: analysis and application areas", International journal of engineering research and development (IJERD), ISSN 2278-800X, Vol. 6, No. 8, 2013, pp. 22–26.

[19] J. Púchyová, M. Kochláň and M. Hodoň, "Development of special smartphone-based Body Area Network: Energy Requirements", Federated conference on computer science and information systems (FedCSIS), September 8-11, 2013. Krakow, Poland, IEEE, 2013, ISBN 978-1-4673-4471-5, pp. 915–920.

[20] D. Laqua, and P. Husar, "Intelligent Power Management enables Autonomous Power Supply of Sensor Systems for Modern Prostheses", *in Biomedizinische Technik/Biomedical Engineering (Impact Factor: 1.16)*, September, 2012, pp. 247–250. DOI:10.1515/bmt-2012-4055.

[21] A. Hofmann, D. Laqua, and P. Husar, "Piezoelectric Based Energy Management System for Powering Intelligent Implants and Prostheses", Biomedizinische Technik/Biomedical Engineering (Impact Factor: 1.16). September, 2012, pp. 263–266. DOI:10.1515/bmt-2012-4265.

[22] J. Miček and M. Kochláň, "Energy-efficient communication systems of wireless sensor networks", *Studia informatica universalis*, ISSN 1621-7545, 2013, Vol. 11, No. 1, pp. 69–86.

[23] D. Suciu, "A Study of RF Link and Coverage in ZigBee", *Scientific Bulletin of the "Petru Maior"*, University of Targu Mures, Vol.7, No.1, 2010, ISSN 1841-9267.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1035–1042
DOI: 10.15439/2014F328
ACSIS, Vol. 2

Mixed-Mode Wireless Indoor Positioning System Using Proximity Detection and Database Correlation

Piotr Wawrzyniak, Piotr Korbel, Piotr Skulimowski and Paweł Poryzała
Institute of Electronics
Lodz University of Technology
Lodz, Poland
piotr.wawrzyniak@dokt.p.lodz.pl, piotr.korbel@p.lodz.pl, piotr.skulimowski@p.lodz.pl, pawel.poryzala@p.lodz.pl

***Abstract*—The paper presents a prototype mixed-mode wireless indoor positioning and navigation system. The main goal of the system is to provide accurate and reliable location information for visually impaired users. The system also enables access to location related context information. The radio nodes of the network can operate in two power modes providing basis for both rough and precise user positioning. The paper describes an overview of the system architecture and the user interface suited to the needs of the visually impaired. Then, the details of the implemented positioning methods are presented. Finally, the results of experimental evaluation of positioning accuracy obtained using different classification strategies are discussed.**

***Index Terms*—Context-aware services, indoor radio communication, location services, personal communication networks, pervasive computing, radio navigation, wireless sensor networks**

I. Introduction

Recently, a number of wireless indoor positioning systems have been developed and numerous research works are reported in the literature. One of the first indoor positioning systems that used radio beacons and Received Signal Strength Indicator (RSSI) measurements was the RADAR system developed by Microsoft Research [1]. From that time the problem of indoor positioning and navigation has been widely addressed around the world [2-14]. Hence maintaining low deployment and maintenance costs is among the most important objectives of the research, majority of the solutions reported in the literature rely on radio signal strength measurements [23-32].

Received signal strength (RSS) measurements can be incorporated to location services in several ways. First of all, radio wave indoor propagation models can be used to determine the possible location of the terminal. This approach requires detailed description of the propagation environment and thus is difficult to implement. Another approach involves the use of database search methods to calculate user terminal position. Therefore, it is necessary to provide reference RSSI measurements (i.e. measurements taken in predefined locations) that are stored in the reference database, and which are then used by location estimation algorithms.

This work was partially supported by the National Centre for Research and Development of Poland under grant no. NR-02 0083-10 in years 2010-2013

One of the application areas of indoor positioning systems is aiding the visually impaired and blind in navigation [22] and orientation and thus in their everyday activities. Almost all electronic travel aids (ETA) for the visually impaired [15-20] require accurate information on current user location. Obtaining precise information on user location can for example facilitate access to public services offered in large buildings (e.g. city halls, hospitals) by aiding to locate rooms or by giving a remote guidance on how to reach the destination. Contemporary satellite navigation systems like GPS provide positioning services sufficient for successful navigation in outdoor scenarios. However, the use of satellite positioning systems is usually impossible in indoor and densely built-up urban areas only. Thus, most of ETAs offered on the market make use of local networks of reference stations that transmit infrared [15] or radio signals [16-18]. The transmitters are used to identify various points of interest (POI) like bus stops, entrances to public buildings, etc.

In the paper, we present a mixed-mode RSSI-based wireless positioning system developed as a part of a complex solution aiding the visually impaired in indoor navigation. The remainder of the paper is organized as follows. Section II presents an overview of the positioning system architecture. Section III provides details of the proposed positioning approach, while Section IV gives details of the user interface module dedicated for visually impaired users. Section V describes experiments conducted to evaluate the performance of the system, while Section VI discusses the results of these experiments. Section VII summarizes our work.

II. System Architecture

The general architecture of the proposed indoor positioning system consists of a local localization server, a local database server and an optional global localization server, as previously described in [21]. A wide range of portable user devices (smartphones, notebooks, tablets etc.) may be used as system terminals, however the terminals should have the capability to measure strength of the signals transmitted by system reference stations mounted inside a building. Thus, a dedicated software or hardware is necessary to read the measurement data and pass the results to the local positioning server.

The key tasks of the local positioning server include

- storing information about the layout of the area (a digital map) it serves (e.g. of an office building);

978-83-60810-58-3/$25.00

- making use of the local database engine to store reference measurement data;
- computing the probable user location based on the RSSI measurement values reported by the terminal;
- management of communication with the global localization server, if available.

The general architecture of the proposed system is shown in Fig. 1. All the components of the system communicate and exchange data using XML/JSON and SOAP-based Web Services.

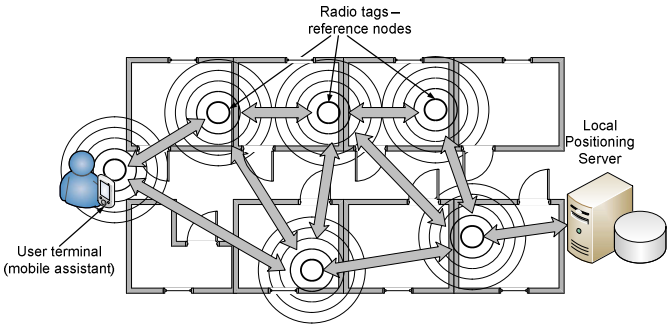


Fig. 1. General architecture of the indoor positioning system.

III. POSITIONING TECHNIQUES

A variety of techniques can be employed to estimate the position of a wireless network terminal. Majority of systems described in the literature base on measurements of signal parameters transmitted by system reference stations. Then, the distances of the terminal to at least some of the reference nodes can be calculated and the approximate position of the terminal can be estimated. Among the most commonly used signal properties are propagation time, angle of arrival, and signal strength, however due to low deployment and maintenance costs, majority of the solutions reported in the literature rely on received signal strength measurements [2, 23-32].

The most straightforward method to estimate the position of a radio terminal is to determine whether it is within the coverage of some reference station. The accuracy of positioning with this approach strongly depends on the range of reference transmitters. However, when reference stations transmit signals with relatively low power, the position of the user terminal may be well approximated by the known location of the reference transmitter. This approach is called *proximity detection*. Practical implementation of this positioning technique involves installation of many reference nodes, often called *radio tags*. However, due to simple tag's construction the overall system installation cost might remain low. This technique offers good accuracy, however it strongly depends on the number of installed reference tags. The idea of positioning system using proximity detection is shown in Fig. 2.

As previously mentioned, distance estimation techniques involving radio wave propagation modeling are widely used in positioning systems. However, due very high complexity of indoor radio wave propagation environment, applicability of these methods is limited to outdoor areas. In typical indoor scenarios, strong multipath propagation effects make it impossible to unambiguously relate measured signal parameter value to a distance from the transmitter. Even along short propagation paths, signal parameters may exhibit very strong variability. Another factor limiting performance of positioning methods is time variability of indoor radio channel characteristics. For example, depending of the time of the day, the offices may either be crowded or almost empty what may result in significant changes of the reported values.

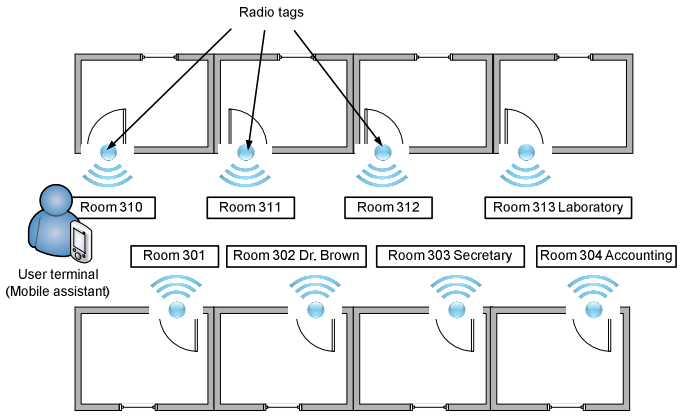


Fig. 2. Proximity detection – presentation of indoor location of the terminal and position related context information (room number).

Therefore, there is a need to search for new positioning methods for indoor applications. One of the approaches that is adequate for indoor systems assumes the use of correlation analysis of reported signal parameter values with some reference data recorded at predefined locations. As database search methods (Fig. 3) rely on evaluation of similarity of measured signal characteristics at actual location to the reference datasets, these methods are not so prone to multipath and shadowing effects as the methods based on radio wave propagation modeling.

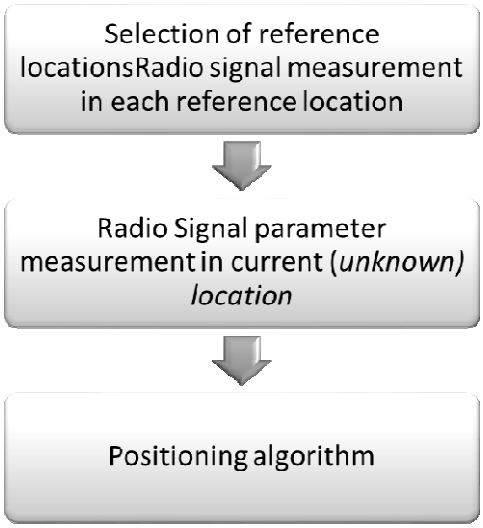


Fig. 3. Block diagram of DCM positioning method

Despite of the fact that database correlation methods may be based on analysis of any available signal parameters, most of practical implementations involve received signal strength

measurements. The advantage of the use of RSSI is that most of contemporary radio receivers provide possibility to monitor RSSI level and a wide range of devices can be used with a positioning system without a need to implement any hardware modifications.

The use of database search methods makes it also possible to reduce the influence of RSSI time variability by the use of normalization to the value read from a given reference signal source.

A. *Proposed Positioning Technique*

The proposed implementation of a wireless indoor positioning system involves received signal strength (RSSI) measurements to estimate terminal position. It makes use of the advantages of both aforementioned approaches, i.e. proximity detection and database search methods.

Hence proximity detection is most effective and accurate when area served by a single reference tag is relatively small, the tags should be equipped with radio transmitters supporting low transmit power modes. On the other hand, database search method accuracy increases with the number of sources of reference signals. In that case, the nodes should be capable to transmit reference signals over relatively large areas. Moreover, the coverage areas of neighboring transmitters should overlap.

It is worth mentioning, that position information in the form of absolute geographical coordinates of the user is not the most expected output from indoor navigation and positioning systems. Geographical coordinates are more suitable for outdoor positioning, mainly due to easy integration with GIS systems.

Moreover, in indoor applications accurate and reliable altitude estimation is required. Although in outdoor scenarios the use of absolute altitude above ground or sea level as altitude descriptor is the most convenient, in indoor scenarios floor index should be considered as the natural way of expressing in-building altitude of travelling people.

Therefore, the proposed indoor positioning system makes use of area-based context-related positioning. Area-based positioning systems provide end users with context information related to the current zone of the building. The system output data set includes but is not limited to:

- floor index or name (if applicable),
- zone within a building (e.g. “north wing”),
- room or office number or its name (e.g. “kitchen” or “auditory no. 416”),
- additional site-related information (like name of current lecture in an auditory room).

Web based application for management of context information is presented in Fig. 4.

Moreover, the proposed system returns absolute coordinates of the user terminal to ensure backward compatibility.

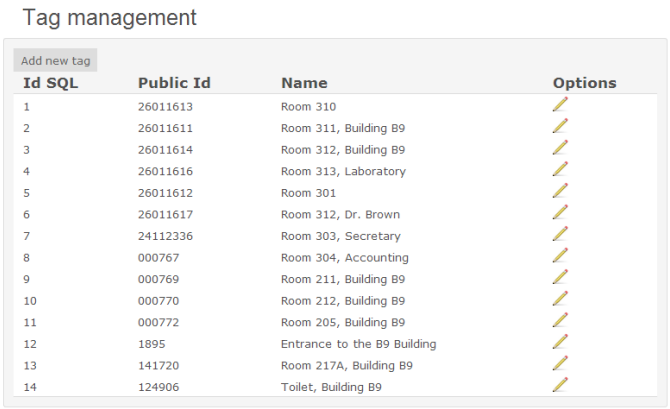


Fig. 4. Web application interface for the management of the radio tags database. Each tag has its unique id and public id (address). Each of the database entries can be edited or deleted.

IV. Positioning Network Interface Module

To enable access to the dedicated positioning network from popular mobile devices (smartphone, tablet, etc.), a dedicated interfacing PilotIE (Fig. 6) device was designed. Proprietary communication protocol with the network of radio beacons enables variable output power settings and provides access to the received signal strength (RSSI) and link quality (LQI) indicators. PilotIE is designed as a device interfacing the radio beacons scattered in the building and the user’s mobile device (via Bluetooth).

Another feature of the PilotIE device is the ability to control selected functions of dedicated mobile applications. Most of the modern smartphones are equipped with a touchscreen and they do not have physical keys. Interaction with a mobile phone requires the use of both hands and may be especially uncomfortable for blind or visually impaired users, who at the same time are often using a white cane. In our solution mobile phone can be hidden in a pocket, which significantly improves the safety of the blind and protect the device in case of falling. Because modern mobile phones are usually equipped with very good quality speech synthesizers, voice messages could be transmitted to the users via the hands-free set. Built-in physical keyboard allows to change systems parameters, like speech speed or volume. It also enables the possibility to re-read the previous message.

Device is built based on the Texas Instruments low-power, 16-bit microcontroller (MSP430F5438A) with the external CC1101 radio transceiver and the TiWi-uB2 Bluetooth 2.1+EDR and BLE 4.0 module. Bluetooth dual mode module can be switched to work in the new, energy efficient Bluetooth 4.0 Low Energy standard (only available in the newest mobile devices) or classic Bluetooth 2.1+EDR mode (most of older devices). It is equipped with two status LEDs and eight user buttons. Device is assembled in a 78mm × 48mm × 20mm enclosure with a single BL-5C (1020 mAh) Li-Ion battery.

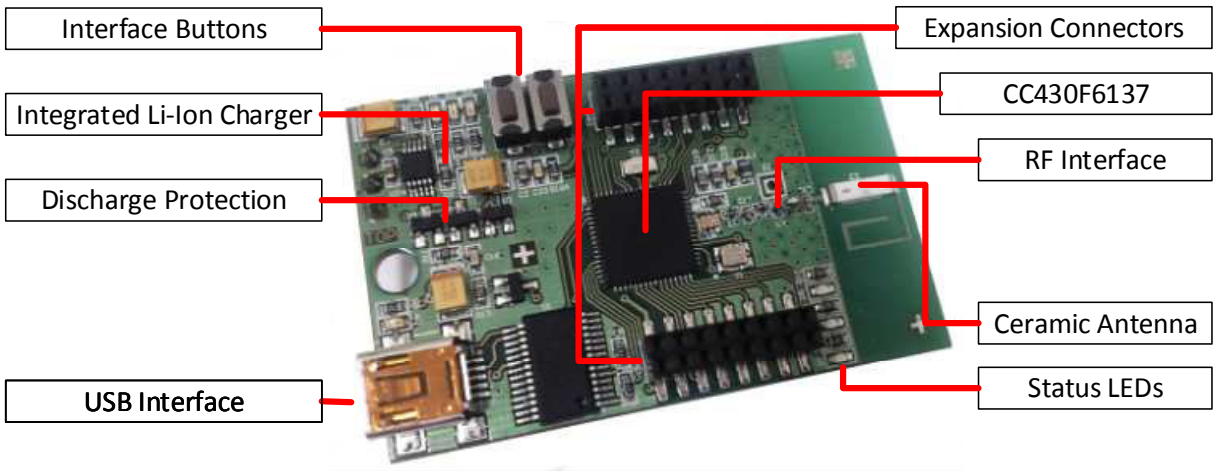


Fig. 5 Radio beacon unit – construction details.

Fig. 6. PilotIE – an interface module for positioning network and mobile phone communications.

V. Experiment Setup

In order to verify the proposed positioning algorithms a number of experiments were undertaken with the use of self-made radio beacons and dedicated user interface module (Fig. 6).

Our prototype radio beacon modules (Fig. 5 and Fig. 7) were built based on the ultralow-power microcontroller system-on-chip (CC430 family, developed by Texas Instruments) with integrated RF transceiver core (CC1101) for the sub-1-GHz industrial, scientific and medical (ISM) radio bands. The CC430F6137 core MCU features 32KB of flash memory, 4KB of RAM, two 16-bit timers, a high-performance 12-bit analog-to-digital converter (ADC), universal serial communication interfaces, a real-time clock module and 44 I/O pins.

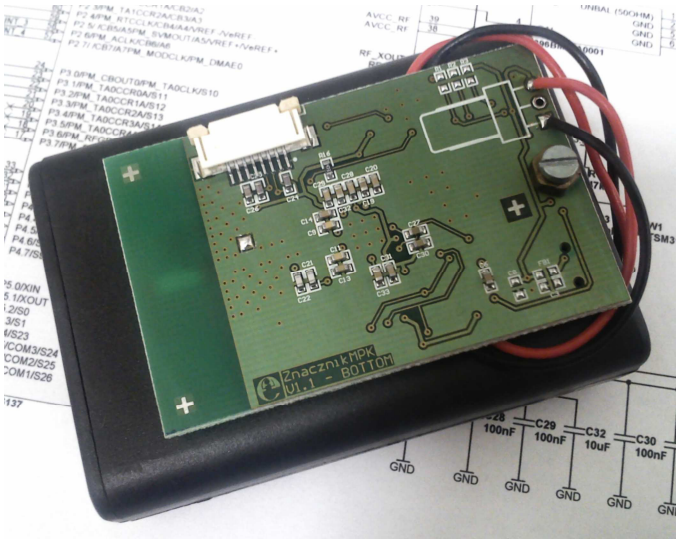


Fig. 7. Radio beacon unit.

Integrated radio module ensures easy implementation of proprietary communication protocols with variable output power settings and provides access to the received signal strength (RSSI) and link quality (LQI) indicators. CC1101 RF transceiver is able to work with programmable data rates ranging from 0.6 to 500 kBaud, programmable output power levels from -30 to +12 dBm, and high sensitivity (up to -117 dBm at 0.6 kBaud for 1% Packet Error Rate).

Radio beacon was optimized to achieve extended battery life. Integrated lithium-ion battery charger and battery discharge protection circuit enables the use of cheap Li-Ion batteries. Device is able to communicate with the PC over the FT232R USB UART interface bridge. Three programmable LEDs and a single TACT switch are provided for status indication and optional configuration. All unused A/D inputs, I2C/SPI buses, GPIOs, and external interrupt inputs are routed out to the expansion connector on top of the module for possible future applications.

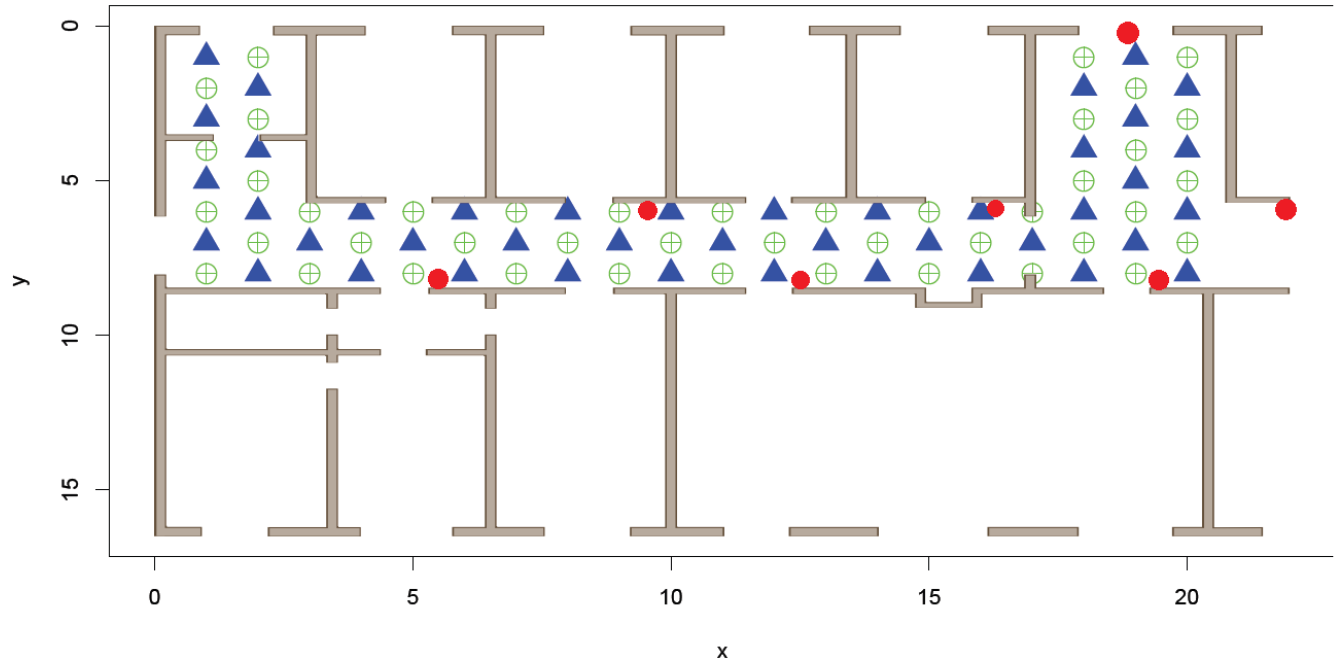


Fig. 8. Experimental setup - RSSI measurement locations (blue - training subset, green - test subset, red - radio beacons)

Device integrates a small ceramic antenna (Johanson Technology 0868AT43A0020, with a peak gain of -1.0 dBi and average gain of -4.0 dBi) or can be used with an optional unipol wire antenna. In the presented application radio beacons are pre-configured for the 868MHz band with the data rate of 38.4 kBaud and 2-GFSK modulation (20 kHz deviation).

Device was designed on a small PCB (55 × 35 mm) which perfectly fits 3xAA battery holder or can be mounted in the KM-27 housing (64.5mm × 46.8mm × 22.3mm) with a single BL-5C (1020mAh) Li-Ion battery.

Each of the fourteen radio beacons used in experiments was equipped with omnidirectional wired antenna. The nodes were distributed across test bed area (seven locations are marked in Fig. 8). All the beacons transmitted packets with two power levels: -30 dBm and -12 dBm. This approach benefits in ability to combine proximity-based methods with database search within a single machine learning algorithm.

Experiment was settled in office building with long and narrow corridor in its central part and number of offices at the sides of that corridor. RSSI measurements were collected in test points arranged in two grids. Fourteen radio beacons were placed in the experiment area. Placement of radio beacons and locations of the measurement points is shown in Fig. 8.

While one dataset was used for training the algorithms, the second one was intended to be used to verify positioning accuracy. Nevertheless, each measurement record contained 10 RSSI readouts taken with the sampling interval of 2.5 seconds.

On the basis of previous research [21] two positioning algorithms were developed. The first one employs Multilayer Perceptron Artificial Neural Network (MLP). The other one was Random Forests (RF) [33]. Moreover, k-Nearest Neighbors (kNN) algorithm with $k=7$ was used as a reference method. All the algorithms were trained using records from learning dataset and positioning accuracy was verified with separate testing set. We examined and compared positioning accuracy for the following three scenarios:

- Using DCM-only system, i.e. only the subsets of RSSI measurements for packets transmitted with the power of -12 dBm were used.
- Using proximity-only system, i.e. only the subsets of RSSI measurements for packets transmitted with the power of -30 dBm were used.
- Using proposed fusion method, i.e. complete datasets containing RSSI measurements for packets transmitted with both -12 dBm and -30 dBm were used.

VI. EXPERIMENT RESULTS

Experimental results show that Multi-Layer Perceptron Neural Networks underperform Random Forests in indoor positioning applications. Random Forests classifier provides the lowest overall average error as well as each of the quartile errors. It is worth to notice that Random Forests classifier achieves better results when mixed-mode input data (i.e. proximity-based and DCM-like) are used. This is in opposite to MLP, where the highest accuracy can be reached with the use of proximity mode only. It is also noticeable that in case of MLP classifier, the biggest averaged positioning error was achieved for the mixed mode.

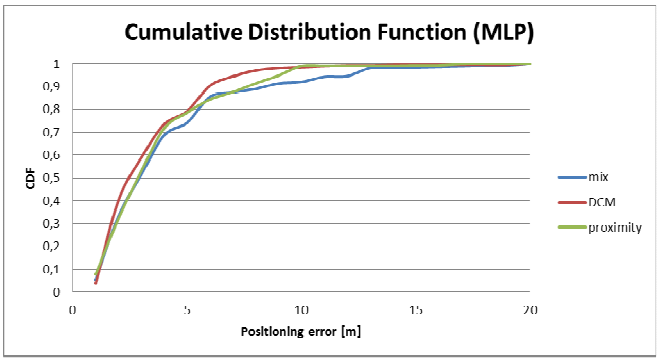


Fig. 9. Cumulative Distribution Function (CDF) of positioning error for MultiLayer Perceptron Artificial Neural Network classifier

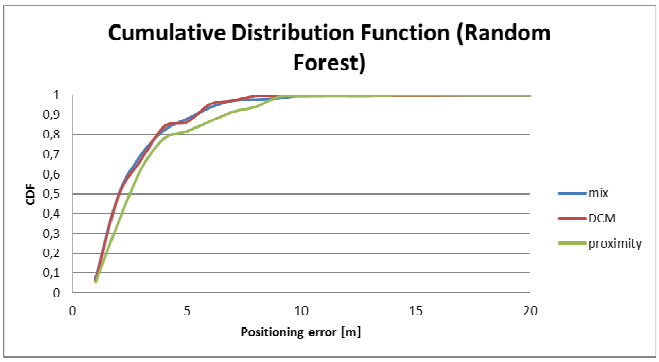


Fig. 10. Cumulative Distribution Function (CDF) of positioning error for Random Forests classifier

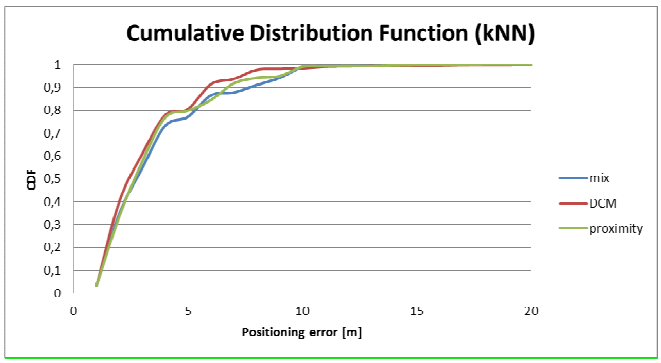


Fig. 11. Cumulative Distribution Function (CDF) of positioning error for k-Nearest Neighbour classifier

Simple kNN classifier outperforms MLP when average error is considered. On the other hand, the maximum positioning error (4th quartile) for kNN classifier was as high as 19.42 meters for DCM-like dataset, which is significantly higher than for Random Forest classifier.

Random Forest classifier takes all the benefits from combined proximity and DCM modes and outranked MLP and kNN in all accuracy indicators i.e. mean positioning error and the values for all considered quartiles.

Averaged positioning error and error quartiles for the three examined approaches and two proposed algorithms were summarized in Table I.

When CDF functions are considered (Fig. 9 to Fig. 11) a number of features might be observed. First of all CDF for Random Forest classifier outperforms CDFs for MLP and kNN while growth rate in the area of positioning error ranging from 0 to 10 meters is considered. This results in high accuracy in real applications, which is reflected in low average positioning error. It can be also observed that RF underperforms when proximity-only mode is used, however combining proximity mode with DCM methodology results in increasing positioning accuracy. On the other hand, combining proximity and DCM methodologies results in significant increase of location estimation inaccuracy when MLP classifier is involved.

Besides positioning accuracy, Random Forests are also computationally very efficient when compared to MultiLayer Perceptron network, especially when the training time is considered.

TABLE I.
ERROR INDICATORS OF THE TESTED POSITIONING METHODS

	-30 dBm			-12dBm			mixed -30 dBm and -12 dBm		
Indicator	MLP	RF	kNN	MLP	RF	kNN	MLP	RF	kNN
Average positioning error [m]	3.16	2.79	2.95	2.83	2.27	2.70	3.53	2.32	3.10
Average positioning error (25 %) [m]	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Average positioning error (50 %) [m]	2.24	2.24	2.24	2.24	2.00	2.24	2.24	2.00	2.24
Average positioning error (75 %) [m]	4.12	3.16	3.16	4.12	3.00	3.00	5.00	3.00	4.12
Average positioning error (100%) [m]	17.00	13.34	17.03	19.10	16.03	19.42	19.00	14.00	14.04

VII. SUMMARY AND CONCLUSION

In this paper indoor positioning system for short range radio communications network has been proposed. The system uses radio beacons transmitting packets with two different power levels of -12 dBm and -30 dBm. This approach gives possibility for joint use of positioning methods based on proximity detection and database search. While packets transmitted with output power of -12 dBm covers relatively large area, transmit power of -30 dBm strongly distinguish small areas of interest.

The positioning system combines proximity sensing and database search methods. The experiments conducted in a large office building resulted in average positioning error not exceeding 2.32 meters when Random Forest classifier was used with combined proximity sensing and database search methods. Since the system uses custom device coupling radio beacons with modern smartphones via Bluetooth connectivity, it is possible to present the results to the users in the form of voice messages.

Future development works assume incorporation of multi-system opportunistic positioning. As the result, the positioning algorithms will benefit from the use of data from generally available radio networks, like public Wi-Fi or mobile cellular telephony networks.

ACKNOWLEDGMENT

This work was partially supported by the National Centre for Research and Development of Poland under grant no. NR-02 0083-10 in years 2010-2013.

REFERENCES

[1] P. Bahl, V.N. Padmanabhan, "RADAR: an in-building RF-based user location and tracking system," INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE , vol.2, pp.775-784 vol.2, 2000, http://dx.doi.org/10.1109/INFCOM.2000.832252

[2] Y. Gu, A. Lo, I. Niemegeers, "A Survey of Indoor Positioning Systems for Wireless Personal Networks," IEEE Communications Surveys & Tutorials, No. 1, 2009, pp.13–32, http://dx.doi.org/10.1109/SURV.2009.090103

[3] M. Mendalka, K. Bizewski, Ł. Kulas, and K. Nyka, "Pattern matching localization In ZigBee wireless sensor networks," Proc. 18th International Conference on Microwave, Radar and Wireless Communications MIKON-2010, Vilnius, Lithuania, 2010.

[4] N. Patwari, J.N. Ash, S. Kyperountas, A.O. Hero III, R.L. Moses, and N.S. Correal, "Locating the Nodes: Cooperative Localization in Wireless Sensor Networks," IEEE Signal Processing Magazine, 22 No. 4, 2005, pp. 54–69, http://dx.doi.org/10.1109/MSP.2005.1458287

[5] Ekahau Real Time Location System (RTLS). http://www.ekahau.com/products/real-time-location-system/overview.html. Accessed 20 May 2013.

[6] A. Kushki, K.N. Plataniotis, A.N. Venetsanopoulos, "WLAN Positioning Systems," Cambridge University Press, Cambridge, 2012.

[7] G. Gonçalo G, H. Sarmento, "Indoor Location System using ZigBee Technology," Proc. Third International Conference on Sensor Technologies and Applications SENSORCOMM 2009, Athens, Greece, 2009, http://dx.doi.org/10.1109/SENSORCOMM.2009.31

[8] K. Radecki, P. Lewicki, and J. Marski, „Lokalizacja terminala radiowego z wyjściem RSSI wewnątrz korytarza budynku," Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne, 6, 2011, pp. 479–482. (*in Polish*)

[9] I. Maly, Z. Mikovec, and J. Vystrcil, "Interactive Analytical Tool for Usability Analysis of Mobile Indoor Navigation Application," Proc. 3rd International Conference on Human System Interaction, Rzeszów, Poland, 2010, pp. 259–266, http://dx.doi.org/10.1109/HSI.2010.5514559

[10] Jun-geun Park et al., "Growing an organic indoor location system," Proceedings of the 8th international conference on Mobile systems, applications, and services, ACM, 2010, pp. 271–284, http://dx.doi.org/10.1145/1814433.1814461

[11] J. Hightower, G. Borriello, "Particle filters for location estimation in ubiquitous computing: A case study," Proc. UbiComp 2004: Ubiquitous Computing, 2004, pp. 88–106.

[12] L. Zekeng, I. Barakos, and S. Poslad, "Indoor location and orientation determination for wireless personal area networks," Mobile Entity Localization and Tracking in GPS-less Environments, 2009, pp. 91–105.

[13] C. Laoudias, G. Constantinou, M. Constantinides, S. Nicolaou, D. Zeinalipour-Yazti, C. Panayiotou, "The airplace indoor positioning platform for android smartphones," Proc. 13th International Conference on Mobile Data Management (MDM'12), 2012, http://dx.doi.org/10.1109/MDM.2012.68

[14] L.A. Guerrero, F. Vasquez, and S.F. Ochoa, "An Indoor Navigation System for the Visually Impaired," Sensors No. 12, 2012, pp.8236–8258, http://dx.doi.org/10.3390/s120608236

[15] Talking Signs. http://www.talkingsigns.com/. Accessed 21 May 2013

[16] K. Radecki, K. Łukaszewicz, „Lokalne radiowe systemy orientacji dla osób niewidomych w środowisku miejskim," Proc. Krajowa Konferencja Radiokomunikacji, Radiodyfuzji i Telewizji KKRRiT 2004, Warsaw, Poland, 2004. (*in Polish*)

[17] S. Bohonos, A. Lee, A. Malik, C. Thai, and R. Manduchi, "Universal Real-Time Navigational Assistance (URNA): An Urban Bluetooth Beacon for the Blind," Proc. 1st ACM SIGMOBILE International Workshop on Systems and Networking Support for Healthcare and Assisted Living Environment, New York, 2007, pp. 83-88, http://dx.doi.org/10.1145/1248054.1248080

[18] J. Marski, P. Bajurko, K. Radecki, T. Buczkowski, „Miniaturowe radiolatarnie i terminale z sygnalizacją RSSI do wspomagania orientacji osób niewidomych," Przegląd telekomunikacyjny i wiadomości telekomunikacyjne, 6, 2010, pp. 320-323 (*in Polish*)

[19] P. Barański, M. Polańczyk, and P. Strumiłło, "A Remote Guidance System for the Blind," Proc. 12th IEEE International Conference on e-Health Networking, Applications and Services HealthCom 2010, Lyon, France, 2010, http://dx.doi.org/10.1109/HEALTH.2010.5556539

[20] M. Polańczyk, P. Skulimowski, B. Sujecki, and D. Sulmowski, "Personal Navigation System for the Blind based on Points of Interest," Proc. II Forum Innowacji Młodych Badaczy 2011, Łódź, Poland, 2011.

[21] P. Wawrzyniak, P. Korbel, "Wireless indoor positioning system for the visually impaired," Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on, pp.907-910, 8-11 Sept. 2013.

[22] P. Strumiłło, "Electronic Interfaces Aiding the Visually Impaired in Environmental Access, Mobility and Navigation," Proc. 3rd International Conference on Human System Interaction, Rzeszów, Poland, 2010, pp. 17–24, http://dx.doi.org/10.1109/HSI.2010.5514595

[23] Wong Siew Mooi, Tan Chong Eng, Huda bt. Nik Zulkifli, "Efficient RFID tag placement framework for in building navigation system for the blind," Information and Telecommunication Technologies (APSITT), 2010 8th Asia-Pacific Symposium on, pp.1-6, 15-18 June 2010.

[24] K. Piwowarczyk, P. Korbel, T. Kacprzak, "Analysis of the influence of radio beacon placement on the accuracy of indoor positioning system," Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on, pp.889-894, 8-11 Sept. 2013.

[25] N. Pritt, "Indoor location with Wi-Fi fingerprinting," Applied Imagery Pattern Recognition Workshop: Sensing for Control and Augmentation, 2013 IEEE AIPR, pp.1,8, 23-25 Oct. 2013, http://dx.doi.org/10.1109/AIPR.2013.6749334

[26] B. Taylor, Dah-Jye Lee; Dong Zhang; Guangming Xiong, "Smart phone-based Indoor guidance system for the visually impaired," Control

Automation Robotics & Vision (ICARCV), 2012 12th International Conference on, pp.871-876, 5-7 Dec. 2012, http://dx.doi.org/10.1109/ICARCV.2012.6485272

[27] P. S. Anindya, E.A. Wan, "RSSI-Based Indoor Localization and Tracking Using Sigma-Point Kalman Smoothers," Selected Topics in Signal Processing, IEEE Journal of, vol.3, no.5, pp.860,873, Oct. 2009

[28] A.V. Bosisio, "Performances of an RSSI-based positioning and tracking algorithm," Indoor Positioning and Indoor Navigation (IPIN), 2011 International Conference on, pp.1,8, 21-23 Sept. 2011, http://dx.doi.org/10.1109/IPIN.2011.6071952

[29] M.O. Gani, C. OBrien, S.I. Ahamed, R.O. Smith, "RSSI Based Indoor Localization for Smartphone Using Fixed and Mobile Wireless Node," Computer Software and Applications Conference (COMPSAC), 2013 IEEE 37th Annual, pp.110,117, 22-26 July 2013, http://dx.doi.org/10.1109/COMPSAC.2013.18

[30] S.C.Ergen, H.S.Tetikol, MKontik, R.Sevlian, R.Rajgopal, and P.Varaiya, "RSSI-Fingerprinting-Based Mobile Phone Localization With Route Constraints," IEEE Transactions on Vehicular Technology, Vol. 63, No.1, January 2014.

[31] A.Zanella, A.Bardella, "RSS-Based Ranging by Multichannel RSS Averaging," IEEE Wireless Communications Letters,Vol. 3, No. 1, February 2014.

[32] M.Ficco, C.Esposito, and A.Napolitano, "Calibrating Indoor Positioning Systems with Low Efforts," IEEE Transactions on Mobile Computing, Vol. 13, No.4, April 2014, http://dx.doi.org/10.1109/TMC.2013.29

[33] L.Breiman, "Random Forests," Machine Learning, 45, pp. 5-32, Kluwer Academic Publishers, 2001.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1043–1047
DOI: 10.15439/2014F210
ACSIS, Vol. 2

Tool-supported Requirements-based Topology Design for Wireless Sensor Networks

Stefan Lange, Jürgen Lösche, Krzysztof Piotrowski
IHP
Im Technologiepark 25
15236 Frankfurt(Oder)
Germany
Email: {lange|loesche|piotrowski}@ihp-microelectronics.com

***Abstract*—Planing the topology of wireless sensor networks (WSN) for a specific application is a complex task. Each application defines requirements to its WSN. Some of these requirements have to been fulfilled by the wireless technology, e.g., energy consumption and throughput, and some by the network topology, e.g., redundancy and latency. Topology makes restrictions to the wireless technology and the wireless technology makes restrictions to the network topology.**

In this paper we present an algorithm to select a network topology and a wireless technology depending on application's requirements automatically.

The algorithm is part of the Sens4U approach, which aims to simplify and possibly automate the process of building WSN applications and support applications development done by non-WSN-experts.

I. Introduction

WIRELESS SENSOR NETWORK (WSN) in enviromental monitoring consists of a large number of nodes with sensors and a wireless network device. These nodes are deployed over a large area, take measurements, and send the data to a sink node where the data is stored and can be analyzed. Planing of such networks is a complex and error prone task.

There are two approaches to deploy the nodes. For the first approach the sensor nodes are randomly distributed over the area, e.g., abandoned from airplane. After arriving final positions all nodes determine their geographic coordinates automatically and set up a network topology autonomously. Advantage of the approach is needlessness of a proper design.

However, the main disadvantage of this approach is the higher demand of the network nodes for systems resources. Dedicated hardware and software components are required to obtain the geographic position of the node. Furthermore, each localization method has an inherent inaccuracy which leads into divergences between the measured and the real coordinates of the node. Moreover, maintenance of network is complicated due to a lack of documentation.

The second approach contains a design phase to prepare the deployment. In this phase the positions of the network nodes are determined. During the deployment phase each sensor node is installed at its predefined location. Thus, deployment of nodes is well documented.

The toolchain introduced in the project Sens4u[1] follows the second approach. The project aims to simplify and possibly automatize the process of building a WSN application. It brings together WSN-experts and non-WSN-experts. WSN-experts can develop modules for WSN register them to the Module Pool. Non-WSN-experts are enabled to specify and build their WSN. The Sens4u toolchain transforms the WSN application specification into an implementation using the Module Pool. Thus, as a result of the project usage of WSN applications have been made available for a wide spectrum of scenarious. The proposed toolchain is given in Figure 1.

In the design flow the customer expertise into the application domain expertise is owned by the actual customer and the, at least basic, WSN expertise is owned by the integrator. The integrator role is introduced to support the customer in requirement specification. The customer explains the target application to the integrator. They identify features of the application and provides these to the planning tool as input. The planning tool generates the set of technical requirements containing the required functionalities and their required parameters. The set is forwarded to the expert system. In this component the hardware/software configuration is generated, based on the available modules in the module pool and the technical requirements. An important part of the configuration is the proposed network topology. On the one hand, several technical requirements reduce the number of allowed topologies. On the other hand, not all topologies are supported by all wireless technologies. In addition, the selection of a wireless technology results can cause side effects by including several hardware and software modules for the wireless technology. Therefore choosing a topology for the WSN application is a challenging task.

This paper focus strictly on topology selection in context of the Sens4U toolchain. All other aspects are outlined briefly only.

The remaining part of the paper is structured as follows. The following section describes the concept in detail. Section 4 represents details about the Proof-Of-Concept followed by an application example in Section 5. A section with an overview of related work follows. The paper ends with a conclusion and an outlook for future work.

978-83-60810-58-3/$25.00

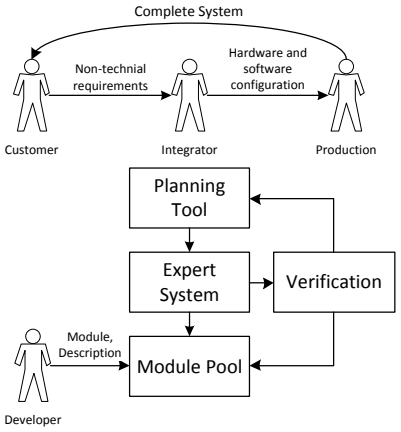


Fig. 1. The detailed tool-chain-oriented development flow and user roles.[1]

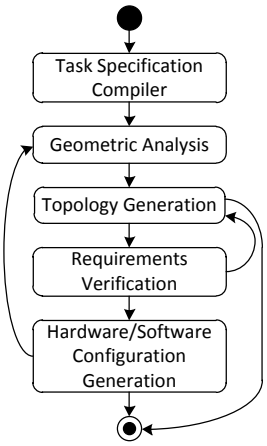


Fig. 2. Data flow in the Expert System

II. Concept

A. Input Data

The expert system receives specifications of WSN applications from the planning tool. In this paper a network specification is definied as a tuple $S_0 = (M, s, \{R_0, R_1\}, p, t, B, R)$ consisting of

1) A set of measurement points M.
2) A sink $s \in M$.
3) Two reference points R_0 and R_1 with given geographically coordinates. These points are used to map the coordinates of the specification to geographically coordinates by applying cross-multiplication.
4) A function $p : M \cup \{R_0, R_1\} \rightarrow (x, y)$ which maps measurement points to coordinates.
5) A function $t : M \rightarrow T$ which maps measurement points to measurement task definitions.
6) A polygon B describing the outer borders of the area.
7) A set of requirements R. Current status is that equations $key = value$ are supported. The supported keys are given in Table I. If no requriement defines a value for a key, the default value is used for the key.

B. Architecture of the Expert System

The expert system component is the core of the Sens4U toolchain. The expert system is composed out of five subcomponents which are given in Figure 2. The figure also shows the dataflow between the subcomponents. The functionality of each subcomponent is described in the following text.

- The *Task Specification Compiler* builds from the task definition of each measurement point an application modell. In addition it calculates the expected datarate during measurement operation. The result is extended to tuple $S_1 = (M, s, \{R_0, R_1\}, p, t, d, B, R)$. The function $d : M \rightarrow \mathbb{Q}$ maps each measurement point to the calculated datarate. The functionality dealing with the generated application modell is not covered by this paper. For this reason the applciation modell is not declared as an element in S_1.
- During the *Geometric Analysis* the network specification is analyzed. The set M is treated as a fully connected undirected graph G_0. The distance between its incident nodes is assigned to each edge. A graph G is created as a copy from G_0 but without all edges with a distance longer then $maxrange$. For G the values in Table II are calculated.
- *Topology Generation*: This subcomponent creates topology suggestion according to the information received from the geomtric analysis. If no topology suggestion can be made, the process exists returning an empty result. The exact behaviour of this subcomponent is described in the text later.
- The component *Requirements Verification* looks for violations of requirements caused by the topology. If a violation is found the topology suggestion is rejected and topology generation will try to create another suggestion. The utilization of defensive programming simplifies the development of new algorithm for creation of topology suggestions and defining new requirement keys. Due to this addtional requirements verification the algorithms need only evaluate subsets of R.
- The *Hardware/Software Configuration Generation* trys to create hardware/software configuration for the wireless sensor nodes. It looks up the module pool for a hardware/software plattform which firstly is able to implement the application modell and secondly supports a wireless network technology able to form the given topology suggestion. This step can result in three different states.
  1) Firstly, there can be no solution. In this case the expert system terminates and returns with no solution.
  2) Secondly, the Expert System have found one or more possible solutions. In this case the process ends and returns the solutions to the planning tool.
  3) The third case means that there can be solutions with additional requirements. The topology suggestion is discarded and the additional

TABLE I
KEY/VALUE PAIRS SUPPORTED AS REQUIREMENTS

Key	Description	Default Value
$maxrange$	The maximum range of the wireless network technology.	$+\infty$
$rmaxdegree$	The maximum degree of a network node in the topology.	$+\infty$
$maxhops$	The maximum number of hops between a network data and the data sink.	$+\infty$
$maxdatarate$	The maximum throughput supported by the network technology.	$+\infty$
$minredundancy$	The minimum path redundancy of the topology.	1

TABLE II
PROPERTIES DETERMINED IN STEP 1.

Key	Description
$maxdistance$	The largest geometrical distances between two measurement points in G_0.
$diameter$	The longest path between two nodes in G.
$pmaxdegree$	The biggest degree of a node in G.
$components$	The number of components in G.
$nodescount$	The number of vertices in G.

requirements are added to R. Then the process restarts at the geometric analysis.

C. Subcomponent Topology Generation

In this subcomponent two tasks have to be processed. The network nodes are placed and a topology suggestion for the network is created. The topology suggestion is stored as a Directed Acyclic Graph(DAG) $G_T = (V_T, A_T)$. Each vertex $v \in V_T$ represent a network node. Each directed edge $d \in A_T$ describes a suggested network link. The data flows from source to target of each edge. The root node of G_T is the data sink.

The rules how a topology suggestion is derived from a network specification are implemented in several Topology Creators. The Topology Creators are registered at process chain a priori. The Topology Creator is selected according to the requirements in M and the properties of G. For this purpose each topology creator stores a set of constraints for requirements and properties. Only under the given constraints the topology creator can create a topology suggestion.

A prolog engine is used to identify topology generators unable to fulfill given reguirements. Constraints, properties, and requirements are loaded as predicates to the prolog engine. From the set of topology generators fulfilling requirements one is randomly choosen and executed. The result is sent to requirements verification component. If no topology generator is able to fulfill the requirements, the expert system terminates and return an empty result to the planning tool. The behaviour described above is given in Algorithm 1.

In the parargaphs folloqing two Topology Generators are introduced in detail.

a) Single-Hop Star Topology Creator: This generator implements the most simple way to create a topology suggestion. On each measurement point a network nodes is placed and all nodes are connected directly to the sink. This algorithm works only if each node is within range of the sink.

The constraints of the Single-Hop Star Topology Generator are given in Figure 3. Lines 1 to 3 define constants. First parameter gives the name , second parameter gives the value, and third parameter gives the unit of the constant. Lines 4 to 7 define the constraints. First parameter of a constraint gives the name of the generator. The other parameters of each constraint define an equation between requirements, properties and constants. In Line 4 a constraint limiting reachable redundancy is given. A star topology does not provide any redundancy which implies a maximum value of 1 for the requirement $redundancy$. Due to, the generator does not place any repeating nodes and generates a Single-Hop topology, line 5 declares that G_1 must consists of only one component. Line 6 gives an essential but not sufficient constraint over the diameter of G. The diameter of a star is 2. Due to, the diameter of G must be less or equal to 2. However, the constraint does no check for the sink as the central node. From all measurement points data should be sent to the sink. As follows, the sink node will have a degree equal to the number of measurement points. Line 7 ensures that the requirement $rmaxdegree$ allows such a topology.

b) Multi-Hop Tree Topology: This generator gets a minimum spanning tree from G_1 as topology suggestion. As described for the previous generator the nodes are placed at the positions of the measurement points. The constraints are given in Figure 4. This definitions differs in two points. There is no constraint on $rmaxdegree$. The generator has to check for degree of each in node in the spanning tree itself. If no valid spanning tree can be found the generator has to return an empty result. Even, there is no constraint on $diameter$. The existence of a spanning tree in a graph can not be derived from equation with the diameter of that graph.

c) Topology Generator Selection: The Prolog source to find topology generators, which can not fulfill the given requirements, is shown in Figure 5. Determining the suitability of a generator needs a proof of all constraints. Showing the inadequacy needs one failed proof of constraint, only. Thats why $choose(T)$ checks for violations and has all unusable topology generators as its result. Each valid value for T in the formular $choose(T)$ gives an topology generator which cannot be used with given requirements.

```
constant('StarTopology.maxpcomponents',1,'1').
constant('StarTopology.redundancy',1,'1').
constant('StarTopology.maxdiameter',2,'1').
constraint('StarTopology','redundancy','==','StarTopology.redundancy').
constraint('StarTopology','components','==','StarTopology.maxcomponents').
constraint('StarTopology','diameter','<=','StarTopology.maxdiameter').
constraint('StarTopology','rmaxdegree','<=','nodescount')
```

Fig. 3. Prolog source for the constraints of the Single-Hop Star Topology Creator

```
constant('TreeTopology.maxcomponents',1,'1').
constant('TreeTopology.redundancy',1,'1').
constraint('TreeTopology','minredundancy','==','TreeTopology.redundancy').
constraint('TreeTopology','components','==','TreeTopology.maxcomponents').
```

Fig. 4. Prolog source for the constraints of the Multi-Hop Tree Topology Creator

```
violates('==',A,B):-eval(A,VA,B,VB),VA \= VB.
violates('!=',A,B):-eval(A,VA,B,VB),VA == VB.
violates('<=',A,B):-eval(A,VA,B,VB),VA > VB.
violates('>=',A,B):-eval(A,VA,B,VB),VA < VB.
violates('<',A,B):-eval(A,VA,B,VB),VA >= VB.
violates('>',A,B):-eval(A,VA,B,VB),VA =< VB.
choose(T):-constrant(T,A,R,B),violates(R,A,B).
```

Fig. 5. Prolog source to find topology generators violating the requirements.

Fig. 6. GUI of the expert system.

III. Proof-Of-Concept

At first this section gives an overview on the implementation of the expert system. Secondly, the result of processing a application specification is presented.

A. Implementation

The expert system has been implemented as a web application based on Java Server Pages running on an apache-tomcat server. The user interface is shown in Figure 6. The left side of the window contains a list with uploaded application specifications are or being processed. When a request get the state "PROCESSED" the data for the planning tool can be downnloaded using the link in the last row. The right side of the window contains the upload dialog. Here data from the planning tool is imported to the expert system.

The process described in the previous subsection has been implemented as part of Network-Analyzer in project Sens4u. The data is stored in GEXF[2] format. Requirements are stored in RuleML[3], which is inserted in GEXF file.

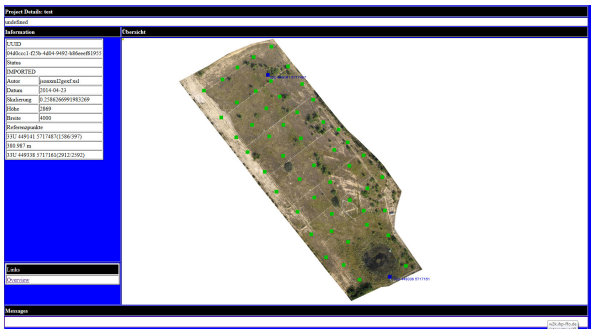


Fig. 7. GUI of the expert system.

B. Example

In context of the project a wireless sensor network is planned at the artificial catchment Hühnerwasser[1]. The application specification received from the planning tool and imported to the expert system is given in Figure 7.

In figure 8 the topology suggestion produced by the Sens4U toolchain is shown. There are still no requirements defined for the application, the topology is a star.

[1]http://www.tu-cottbus.de/projekte/en/oekosysteme/startseite.html

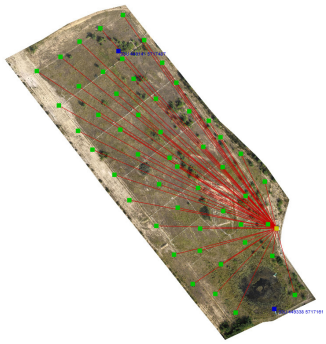


Fig. 8. A generated star topology for Hühnerwasser WSN application.

IV. Related Work

In [4] the software environment POWER for planning and deploying wireless sensor networks is introduced. POWER implements an iterative process. Firstly, nodes are placed in a virtual environment. Secondly, the network in virtual environment is simualeted and evaluated. If an optimal solution is found the process terminates at this point. If not, the network in the virtual environment is optimized in the third step and the process starts again. Unlike the Sens4U toolchain POWER do not enable non-WSN-experts to specify its WSN application.

The framework FLEXOR[5] defines a software architecture for wireless sensor nodes and development environment with tools dealing with WSN applications based on the FLEXOR software architecture. Like Sens4U it addresses non-WSN-experts as users of WSN applications. However, while the expert system decides about hardware and software components to use, FLEXOR supports software modules only. Furthermore, the FLEXOR-toolchain does not support topology design.

Algorithm 1 Selection of a topology generator

Require: TC is the set of all registered topology generators.
Reset Prolog Engine
Load R to Prolog Engine
for each $tc \in TC$ **do**
  Load constants of tc to Prolog Engine
  Load constraints of tc to Prolog Engine
end for
$TC_0 \leftarrow$ Query Prolog Engine
$TC_1 \leftarrow TC \setminus TC_0$
if $TC_1 = \emptyset$ **then**
  return $\emptyset$
else
  $G_T \leftarrow \emptyset$
  while $G_T \neq \emptyset$ **do**
    $tc_0 \leftarrow$ first element in TC_1
    $TC_1 \leftarrow TC_1 \setminus \{tc_0\}$
    $G_T \leftarrow$ Call tc_0
  end while
end if

[6] defines a design flow and an user model for a component-based and composition-driven design process. The goal is also let non-WSN-experts specify and develop WSN applications.

GENSEN[7] is a topology generator for realiastic WSN deployments for the network simulator NS2[8]. Input data are not an application specification, but three parameters for node distribution. The parameters specify the distribution strategy, the number of different atenna orientations, and the number of different energy levels. GENSEN is an example for a family of topology generators for simulations.

V. Conclusion and Future Work

This paper presents a solution for the difficult technical problem of topology selection that occurs in each WSN application. The approach itself and its integration into the Sens4U-toolchain are described in detail. An implementation is introduced as Proof-Of-Concept and an example calculation is given.

Several issues are still open. Sensor node placement can be prohibited by restricted zones or can become expensive by problem areas. Likewise, disturbance zones can exist where wireless communication is impossible. For these cases more sophisticated strategies for node placement have to be developed. Complex measurement tasks can be distributed in more than one node. Furthermore, generators for more effective and reliable topologies must be implemented to fulfill survivability resiliance requirements.

Acknowledgment

This work is part of the project Sens4U (Sensorknoten für Umweltmonitoring) and was founded by the Bundesministerium für Bildung und Forschung (BMBF) under grant 03WKP26A.

References

[1] K. Piotrowski and S. Peter, "Sens4u: Wireless sensor network applications for environment monitoring made easy," in *SESENA*, C. Julien and K. Wehrle, Eds. IEEE, 2013. doi: 10.1109/SESENA.2013.6612264 pp. 37–42.

[2] GEXF Working Group, "Gexf 1.2draft primer," Mar 2012.

[3] H. Boley, "The ruleml family of web rule languages," in *Principles and Practice of Semantic Web Reasoning*, ser. Lecture Notes in Computer Science, J. Alferes, J. Bailey, W. May, and U. Schwertel, Eds., vol. 4187. Springer Berlin Heidelberg, 2006. doi: 10.1007/11853107_1 pp. 1–17.

[4] J. Li, Y. Bai, H. Ji, J. Ma, Y. Tian, and D. Qian, "Power: Planning and deployment platform for wireless sensor networks," in *Grid and Cooperative Computing Workshops, 2006. GCCW '06. Fifth International Conference on*, 2006. doi: 10.1109/GCCW.2006.73 pp. 432–436.

[5] A. Forster, K. Garg, D. Puccinelli, and S. Giordano, "Flexor: User friendly wireless sensor network development and deployment," in *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2012 IEEE International Symposium on a*, 2012. doi: 10.1109/WoWMoM.2012.6263698 pp. 1–9.

[6] S. Peter and P. Langendorfer, "Tool-supported methodology for component-based design of wireless sensor network applications," in *Computer Software and Applications Conference Workshops (COMPSACW), 2012 IEEE 36th Annual*, 2012. doi: 10.1109/COMPSACW.2012.98 pp. 526–531.

[7] T. Camilo, J. S. Silva, A. Rodrigues, and F. Boavida, "Gensen: A topology generator for real wireless sensor networks deployment," in *Proceedings of the 5th IFIP WG 10.2 International Conference on Software Technologies for Embedded and Ubiquitous Systems*, ser. SEUS'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 436–445.

[8] K. Fall and K. Varadhan, Eds., *The ns Manual*. The VINT Project, 2011.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1049–1057
DOI: 10.15439/2014F221
ACSIS, Vol. 2

An Energy Conservative Wireless Sensor Network Model for Object Tracking

Gokcer Peynirci
Institute of Applied Sciences,
Yasar University, Izmir, Turkey
Email:
gokcer.peynirci1@stu.yasar.edu.tr

Ilker Korkmaz
Dept. of Computer Engineering,
Izmir University of Economics,
Izmir, Turkey
Email: ilker.korkmaz@ieu.edu.tr

Muharrem Gurgen
Siemens Gebze R&D Center,
Kocaeli, Turkey
Email:
muharrem.gurgen@siemens.com

***Abstract*—This study aims to find the relationship between energy consumption level and object tracking success in an object tracking sensor network (OTSN). Convenient use of energy proposes a great challenge for wireless sensor network (WSN) design and the balance between successful object tracking and low energy consumption is a tight one. To address this issue, we propose a new network operation scheme for object tracking, implement this scheme in Network Simulator 2 (ns-2) and present the obtained results of the conducted simulation experiments. The simulation results show that the proposed method can be used to track objects in a WSN network in an energy conservative manner.**

I. Introduction

WIRELESS Sensor Networks (WSN) are mainly composed of a large number of sensor nodes as a network structure whose nodes are limited in terms of memory, processor and power resources. Nodes that are distributed in the environment can communicate unrestrained in short distances and adapt to its environment. By communicating with each other, they compose a variable topology layout and take form of a network themselves [1].

A sensor network provides a set of high level information processing tasks such as event detection, environmental monitoring, object tracking, or classification [2]. Sensor networks today have numerous application areas including health, military, home and various commercial applications [1]. Object tracking in WSNs have become one of the killer applications in this aspect. Although achieving object tracking using WSNs may be relatively hard depending on the object's speed, size, sensors' quality, and also the environmental conditions, it may become widespread with the right technology and methodologies in place. As a top view of the general solution to object tracking problem, a group of sensor nodes can be utilized to sense the environment for an object's location and track it along its journey in a predefined area. Meanwhile, knowing that the most critical constraint of a WSN is its lifetime, we see any study devoted to increasing it as well worth the effort.

During object tracking, a key factor is minimising energy consumption while making sure that the object is monitored. A number of metrics can be monitored to determine the energy consumption such as average number of active sensors during network operation and average length of operation per sensor node. These two provide a good performance assessment in determining energy efficiency of a WSN. There are also some other important metrics that help us determine whether an object tracking application is successful. We need to make sure the time period between the moment an object enters a monitored area and the moment it is detected by the network is kept at a minimal. Another metric is the precision of the object tracking data. This can mainly depend on the following two factors: the localization method in use and the schema used within the network for the tracking the object.

A WSN is quite an ad-hoc type of network which is built for a specific aim. The network should be designed and built with this aim in mind, in our case object tracking, the protocols and methods utilized should be in line with this aim.

WSNs vary in the way they are designed, the equipment used and ultimate goal of operation. Moreover, the way they are deployed to the area, routing protocol used or prediction algorithm used may make the difference between success and failure in object tracking scenarios. Each of these variables can affect the performance of the particular network application in question.

Some researchers [3]–[6] chose to tackle the problem of energy efficiency in object tracking in WSNs from a prediction based point of view. They based their hypothesis upon the fact that if we can predict the object's location for a given period of time, this enables us to activate only the minimal set of sensors thereby minimising energy consumption of each node.

To accomplish effective object tracking, different points of views have emerged such as using a tree-based formation [7], [8] a cluster-based formation [9] or a prediction-based method [3]–[6]. The advantage of tree based methods is in their network coverage rate and minimization of energy consumption under ideal situations. The cluster based methods aim to balance energy consumption among clusters and clusters can be formed dynamically in order to track objects. While the prediction based methods can be used to achieve successful object tracking, it calls for high rates of prediction accuracy and precision to limit energy consumption during an object tracking operation. The prediction based mechanisms can also be added onto cluster-based formations to propose a mixed or multi-phase approach for object tracking. Besides, some heuristics about predictions may also be considered to preserve energy. A short literature review on tree-based, cluster-based and

978-83-60810-58-3/$25.00

prediction-based approaches to object tracking in WSNs can be obtained from [10].

The rest of the paper is organized as follows: In Section II, Literature Review, we discuss some of the protocols and methodologies invented for object tracking in WSNs. In Section III, Factors Affecting Object Tracking Success, the main performance metrics to evaluate the object tracking scenarios in WSNs are explained. Section IV, Proposed Network Model and Operation, clarifies the whole operation process of the sensor network in our approach to track the corresponding object within the predefined environmental boundaries. Section V, Simulation, details the simulation environment considered in our object tracking scenario. Section VI, Experiment Results, gives the results and evaluations of the experiments conducted for different network topologies. Finally, Section VII, Conclusion and Future Work, concludes the paper.

II. Literature Review

This section mainly presents a literature review in target, object, or location tracking in WSN. Some background on the use of energy in an efficient manner in general WSN applications is reviewed as well.

In the Leach protocol [9], the cluster head (CH) in each cluster serves as the main node for data processing and transmission. This results in quick energy drain for the cluster head. To overcome this problem, they proposed randomized rotation of the cluster head role amongst the nodes of a cluster where each node takes on the cluster head role based on a threshold value determined by a probabilistic function. This threshold value T(n) is basically given by the following [9, 11, 12] :

$$T(n) = \frac{P}{1 - P * \left(r \bmod \frac{1}{P} \right)} \quad \forall n \in G$$
$$T(n) = 0 \quad \forall n \notin G \tag{1}$$

where n is node number for which the threshold value is to be computed, r is the current round, P is the percentage of cluster heads and G is the set of nodes that have been cluster heads in the last 1/P rounds. The operation of the Leach protocol is divided by rounds. Each round starts with a set-up phase where the clusters are organised, followed by a steady phase where the collected data is sent from the clusters to the base station. Based on the probabilistic threshold value computed by each node in distributed sensor network area, the use of energy in each cluster head node gets balanced and the network lifetime gets increased by making the nodes survive longer with probabilistic energy conservation.

Two clustering approaches for object tracking are given in [8] and [13]. In [8], a tree-structured cluster is created following the entry of an object in the monitored area. An explicit leader election mechanism is used that selects the sensor closest to the object as the CH. Afterwards, a minimum cost tree is created that includes all the sensors within a predefined range. The tree is set to be reconfigured when distance between target and CH exceeds a pre-set value. In [13], in contrast to using dynamic clusters, the use of static clusters is proposed, where each cluster is activated based on detection of a target. The currently elected CH uses linear prediction to determine whether to keep on tracking or to switch the tracking task to another CH.

There are some protocols that take into account quality of data for nodes which aim to reduce the amount of data being transferred in energy efficient target tracking scenarios in WSN [14]. The redundant data, which is the one collected by closely stationed nodes need only be transferred once to the cluster head. For this purpose two algorithms were proposed in [14]: Reduced Area Reporting (RARE-Area) and Reduction of Active Node Redundancy (RARE-Node). The first one limits the number of nodes taking part in object tracking by monitoring the data quality. Sensor data is assigned with a weight and the nodes that have a weight above the threshold value can participate in tracking. The second one aims to reduce the amount of redundant data by means of identifying spatial relationships between neighbouring nodes.

In Dual Prediction-based Reporting [3], both the sensor nodes and the base station make predictions about object movement to track the object. When the base station makes an error in its prediction, it is corrected by the readings of the sensor nodes.

In the Prediction-based Energy Saving scheme (PES) [4], firstly object movement is predicted to determine the suitable nodes called target nodes. After this, the selected nodes are awakened based on energy and performance metrics. Finally, a recovery mechanism is carried out if the object is missing. This recovery mechanism depends on two modes, namely ALL_NBR [4] which wakes up all the nodes surrounding the estimated route of the moving object and if this one fails, flooding recovery that wakes up all the nodes in the network in a more aggresive fashion is used.

In [10], the authors propose a distributed tracking algorithm which is run at each node of the network. This protocol distinguishes between inner nodes and the border nodes and keeps the border nodes at active state the whole time. They also propose a three level recovery system based on the positions of the nodes in the monitored area.

Another prediction-based energy-efficient target tracking protocol (PET) was proposed in [5] to derive the travelling path of the target and utilize the target's moving patterns for energy saving. Cooperation amongst sensor nodes is the key characteristic for this protocol. A linear predictor is used to predict the target's next location. As not all sensors may have useful information, sensor nodes with the best data possible are selected in order to conserve energy.

In [15] and [16], the prediction is based on the object's movement direction. In [4] and [17], the first node which senses the object wakes its one hop neighbours at first, if the object cannot be located then some more-hop neighbours are awaken, if this also fails, all the nodes are awaken at the worst case. An alternative way to choose the appropriate

nodes to wake is selecting the nodes that have more energy for the recovery process [18].

Another study [6] proposes a prediction algorithm that is divided in clustering and prediction stages. They keep all nodes inactive except the selected CH nodes and when one of CH nodes senses an object, it becomes active and activates three more nodes using the tracker node selection algorithm. The activated nodes carry out the tracking until the current CH node selects the nearest CH node as new current CH node to carry out tracking.

Reference [19] proposes an energy efficient technique to predict the future movements of a mobile object using its inherited behavioural movement data patterns stored. The proposed prediction based tracking technique using sequential pattern (PTSP) offers object tracking with the efforts of a minimum number of sensor nodes in the network; meanwhile the rest of the nodes sleep to preserve the total energy.

Another target tracking approach based on a hybrid predictive model is proposed in [20] to be used in grid wireless sensor networks intrinsically. The proposed model divides the surveillance area into grids and applies a hybrid approach combining the Markov chain and the Grey Theory to predict the target path probabilistically.

A performance comparison between different kinds of tracking algorithms for tracking an object with relatively fast speed in wireless sensor networks is given in [21]. The use of cluster based versus spanning tree based target tracking algortihms are compared mainly. The corresponding cluster-based tracking algorithms involved in the comparisons use etiher a static network where the clusters are formed at deployment time, or a dynamic network where the clusters and the backbones are constructed dynamically in case of an event. Reference [21] also compares the results of adding different filtering techniques, i.e., linear, extended Kalman, and particle filters, into a proposed dynamic lookahead tree based tracking algorithm. The corresponding spanning tree based tracking algorithm is used for degrading the target miss ratio; moreover the filters are used to raise the prediction accuracy.

Another up-to-date proposal for a fast and energy efficient target tracking model based on location prediction is presented in [22]. It is pointed that the proposed method leads to a good accuracy with low energy consumption and it has low missing rates compared to linear and extended Kalman filter predictors.

III. Factors Affecting Object Tracking Success

With the knowledge of the approaches and mechanisms presented in Section II, this section explains the factors and the main performance metrics to evaluate the object tracking scenarios in WSNs.

The monitoring scheme deployed in a WSN setup is equally important regarding energy consumption and object tracking success rate. The sensors can be set to monitor their surroundings in scheduled monitoring mode, where all the sensor nodes are well synchronised to the base station. A dynamic clustering monitoring may be employed where sensor nodes are organised into cluster of nodes reporting to a common cluster head. A prediction based monitoring may be employed which uses a wake-up mechanism to activate specific sensors specified to be within sensing range of the object. This requires a recovery mechanism that is activated in case the object is missed. A prediction based monitoring can only be successful given the internal object location prediction algorithm is producing reliable results, otherwise the sensors will be depleted due to having to carry out a large number of recovery operations. Moreover, the object's trajectory can become the main factor affecting network lifetime, assuming an object can follow similar paths each time it enters the monitored area. There are some main factors effecting object tracking success which are explained in the following subsections.

A. Energy Consumption Efficiency

Energy consumption is directly related to how the network is designed and how it operates. Every communication made between each node in the network incurs a cost in terms of energy to nodes in the network thus should be minimised. The efficiency of the prediction algorithm as well as the mechanism used for the task of object tracking can be taken as base points to measure the energy efficiency of a sensor network.

B. Accuracy of Target Tracking

In order to keep a low probability of missing the object and for an effective target tracking application a good degree of target tracking accuracy should be achieved. This measure is also directly related to the infrastructural mechanisms used for tracking the object, i.e., using prediction based, cluster based or tree based approaches.

C. Scalability

Some applications may require huge numbers of sensor nodes. This can present different challenges compared to networks with smaller number of nodes. Scalability is about how well the network copes with high numbers of nodes. This study made use of a fixed number of nodes, which is relatively small to take into account scalability issues of the network.

D. Interconnectivity

Nodes in a sensor network need to be interconnected for the network to function properly. If there are nodes which have no route to forward packets to the access point, data collected by them will be of no use. Maximum connectivity should be achieved in the deployment stage and it should be preserved as much as possible through energy savings and congestion control.

E. Network Lifetime

Each node in a wireless sensor network is powered by a battery. This means when most of the nodes deplete their batteries the network will not be functional. Therefore it is crucial that nodes make the best use of their batteries by turning off (sleep mode) their microcontroller and transceiver when these are not needed.

IV. Proposed Network Model And Operation

WSNs typically consist of a large number of sensor nodes dispersed in a large field. This study involves simulations made using nodes distributed in 50 x 50 m^2 and 100 x 100 m^2 fields populated with a fixed number of nodes. Comparisons between two deployment methods: grid and random were made. None of the nodes are mobile and each node in the network is set to be equal in terms of initial charge level, computation capability and communication range, namely the network only involves homogeneous sensor nodes. The sensor nodes used in the simulations have identical features to Mica2 [23] motes in terms of energy usage in active and sleep modes. Sensor nodes don't require the use of a GPS device and network lifetime diminishes when any node runs out of energy. Due to the way the ns-2 simulation is configured (area of the simulation is set in the ".tcl" scenario file) the object stays inside the boundaries of the monitored area for the duration of the simulation.

A network built for the purpose of object tracking has basically two subsets of tasks to accomplish, namely monitoring the area and reporting of the object's location to the sink node. If considered on the node level, each node is tasked to listen to its environment and report about object movement when necessary. In the most energy draining scenario, all the nodes have to be listening all the time for potential object movements. If we employ a node wise and network wise mechanism where all the nodes perform in a specific way, we can limit the time necessary to have a node in listening state. According to this specification the network can be in one of the three states: not tracking, tracking or recovering. The recovery mode works by first waking up the neighbouring nodes of the latest active nodes and goes on to wake up all the nodes in the network following a spiral route. This recovery mechanism is similar to [4] but not necessarily the same. The border nodes are the ones that are active continuously in order to make sure there is no object to track in the monitored area.

In order to track an object in the network the first requirement is to sense it and to sense the object at least some sensor nodes should be awake, for instance border nodes or randomly selected nodes. If all nodes are awake the whole time, the network lifetime shortens. If some of the nodes are sleeping in some conditions to save their power, the object could be lost again after having been found. In this situation, a recovery phase is initiated that aims to relocate the object however this leads to extra energy consumption. To achieve energy conservation and minimise object missing rate, some nodes can be awaken before the object enters their sensing territories and some nodes can be put into sleep mode after the object leaves their boundaries'. This mechanism can be triggered as follows: the active node that is sensing the object activates some other sleeping nodes. To select the appropriate nodes to put to sleep or activate requires a good prediction about the object and achieving the best possible prediction estimation means more energy can be conserved.

The nodes that have the possibility to locate the object are first predicted on the sink node. The sink node awakens a sensor node according to the initial prediction results. This is similar to the Wang's model [18], however in their model, when the predicted nodes fail to sense the object, neighborhood nodes are awakened according to the results of the genetic algorithm they use and are awakened on a one by one basis, while we take an approach in which they are awakened based on one of the three approaches we implemented.

A. Prediction Model

Our prediction algorithm is executed at the sink node, based upon the information received from tracking nodes. All the nodes depend on the sink node to determine which state they are in.

The prediction algorithm uses the well-known formula to compute velocity, which is:

$$v = \frac{\sqrt{\left(x_i - x_{i-1}\right)^2 + \left(y_i - y_{i-1}\right)^2}}{t_i - t_{i-1}} \tag{2}$$

We use a linear prediction method as our prediction mechanism that we coded inside the ns-2 code framework. According to the prediction algorithm results, a new set of sensor nodes are given the wake-up signal.

B. Tracking Model

When a border node detects an object in its sensing range, it initially awakens neighbour nodes so that they can carry out localization. The result of this localization is sent to the base station, which in turn uses this data to predict the location of the object. Once the newly activated nodes start sensing the object, they send it to the base station and the base station concludes that the prediction was correct and puts the previous group of nodes into sleep state. A new prediction is made in the same manner, and this loop continues until the object is lost.

The proposed tracking mechanism is based on measuring received signal strength (RSS) at the tracking nodes. RSS which is known to decrease exponentially based upon distance to the tracked object is calculated by [24]:

$$r_i = a.\| x - x_i \|^{-\alpha} + n_i \tag{3}$$

where r_i is the value of the RSS in the i^{th} sensor node, a is the strength of the signal emitted from the target, x refers to the real (yet to be found) coordinates of the target, x_i is the known coordinates of the i^{th} sensor, α is the attenuation coefficient and lastly n_i denotes the white Gaussian noise with zero-mean and variance σ^2.

C. Recovery Model

When the object is lost, the recovery mechanism is triggered which activates all the nodes starting from the neighbour nodes of the last active nodes, after activating all the nodes, if still there is no object detected, the base station concludes the object had left the area, and takes all the nodes except the border nodes back to sleep mode.

There is a second mode of operation of the proposed system which is power saving mode. In this mode, the network acts in a way to further limit the energy consumption by increasing node deactivation frequency and decreasing number of nodes involved in the recovery process at the cost of less accurate tracking. Our simulations focused on finding the trade-off between energy conservation and accuracy of target tracking along with how the two perform for objects travelling at high and low speeds. Those two alternative modes are referred to as "Powersaving" and "Non Powersaving" while we present the measurements of our simulations via graphical results in Section VI.

The recovery mechanism we propose is analogous to the proposed model [4], in which they try to cover misprediction of object's speed and movement in their recovery phase. They propose three models: Heuristic DESTINATION, Heuristic ROUTE and Heuristic ALL NBR. Each of these has different energy efficiencies. The first one assumes the speed and direction of the object is predicted correctly and wakes up one node on the predicted path. The second one assumes the speed of the object is mispredicted and the current node informs nodes on the predicted path. It assumes the direction is correctly predicted. The third one assumes both the speed and the direction of the object are mispredicted and the current node informs the neighboring nodes that surround the route. The sink node, wakes up a node (current node) based on the prediction result, and when this node loses track of the object, it first informs a neighbour node and based on the result it gets from the other node, it either informs the sink node or doesn't. If it informs the sink node, this shows that this node also failed to sense the object, if it doesn't inform, it means the node sensed the object and there was no need for further recovery. The node that sensed the object now becomes the current node.

The difference of our model compared to the energy conservative approach proposed in [4] is that the prediction is done on the sink node and the sink node determines the current node except in situations where the object is lost temporarily and found (neighbour node becomes the current node) or when a node detects a new object by chance in the area. In addition to this, we adapt a more aggressive approach for recovery where we begin to wake up all the nodes as well the ones on the predicted path.

The difference we propose compared to the recovery mechanism of [4] is that, we define a spiral route that begins from the closest neighbour of the last current node (nearest place where the object was lost) that continues to wake up all the nodes in the network until the object is found. We aim to minimise the time it takes to wake up all the nodes in this manner and put them to sleep if the object had already left the area. We also aim to make transitions between these states (prediction, tracking, recovery) as fast as possible in order to make sure the object is tracked for the time it is inside the network and also to increase energy efficieny.

V. SIMULATION

The IEEE 802.15.4 [25] medium access control (MAC) protocol is used for our scenario implementation on ns-2.34 [26]. Different from the IEEE 802.11 [27] protocol which is used for WLAN networks, IEEE 802.15.4 protocol is a low tier, ad hoc, terrestrial, wireless standard for wireless networks and other ad hoc networks such as WSNs. The main simulation parameters used in our different scenarios are given in Table I below.

TABLE I.
SIMULATION PARAMETERS

Description	**Value**
Simulation Environment	ns-2.34
PHY-MAC Layer	IEEE 802.15.4
Field Size	50 m x 50 m, 100 m x 100 m
Tracking Node Number	21
Sink Node Number	1
Sensor Node Deployment	Uniformly Distributed, Randomly Distributed
Energy Consumption (Active Mode)	8 mA
Energy Consumption (Sleep Mode)	< 15 uA
Communication Range	40 m
Sensing Range	~15 m
Velocity of Target	5 m/s - 16 m/s
Duration of Simulations	200 s
Number of Trials	5
Number of Tracked Target	1

A sensor network consisting of 21 tracking nodes uniformly distributed in a field of 50 x 50 m^2 and a sink node at the lower left-hand side is considered. All the sensor nodes are homogeneous and immobile. The topological deployment of the nodes can be seen below in Fig. 1. The tracked object is denoted by a full circle and the sink node is

TABLE II.
STATE TRANSITION TABLE OF THE FSM

	Time_Out	Wake_Up	RSS> Threshold	RSS < Threshold	On_Predicted _Path	Object_Found	Object_Lost	Object_ Left_ Area
SLEEP	-	SENSE	-	-	-	-	-	-
SENSE	SLEEP	-	DETECTED	-	-	-	-	-
DETECTED	-	-	DETECTED	WAIT_MESSAGE	-	-	-	-
WAIT_ MESSAGE	-	-	-	-	SENSE	-	RECOVERY	-
RECOVERY	-	-	DETECTED	-	-	SLEEP	-	SLEEP

designated as a pentagon to differentiate it from the ordinary tracking sensor nodes. Sink node, in our proposal, is responsible for running the prediction algorithm and activating the sensor nodes according to the prediction results.

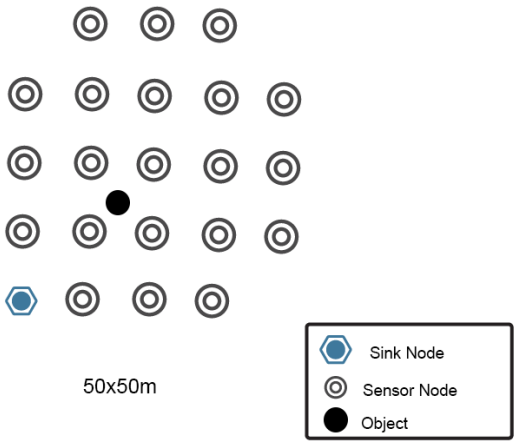


Fig. 1 Main Network Topology

In preparation of the simulation, modifications to the ns-2 code base and implementations of additional functions were defined. An application layer based on the proposed scenario was implemented on top of ns-2.34's IEEE 802.15.4 MAC layer. For transport layer, we modified the message agent which sends packets in similar way to UDP. The most important function is the process message which processes incoming messages and changes the states of nodes. We completely rewrote the message processing functions along with all the functions on the application layer.

Each node is simulated as having five different states during network operation. It is drawn as a finite state machine (FSM) and coded inside the ns-2 code base along with the prediction algorithm. The finite state machine is depicted in Fig. 2, a corresponding state transition table is provided in Table II and detailed descriptions of each state and message used are given below.

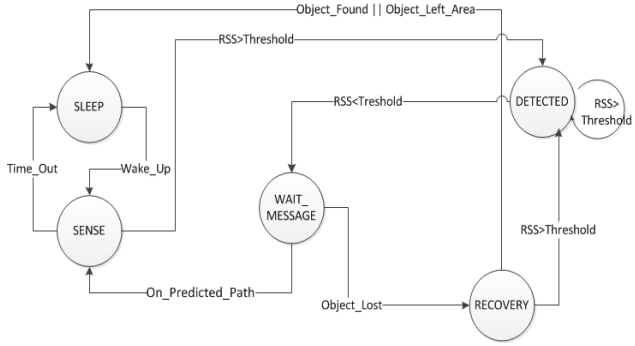


Fig. 2 Finite State Machine

1) States:

SLEEP: In this state, sensor node does not receive or send any messages and keeps energy efficiency at maximum. In order to receive messages from the network, sensor nodes change state to SENSE periodically.

SENSE: In sense state, a sensor node can send and receive messages and periodically changes its state to SLEEP. The sensor nodes selected by the prediction algorithm are in this state and they are expected to have the object in close proximity. If a sensor node senses an object in this state, it changes its state to DETECTED.

DETECTED: Sensor nodes stay in this state as long as their RSS value is greater than the threshold value. If a sensor node cannot sense the object anymore then it changes its state to wait message.

WAIT_MESSAGE: In this state, sensor nodes wait for the sense message of the next node on the predicted path. If it receives On_Predicted_Path message then it changes its state to SENSE and if it receives Object_Lost message it changes to RECOVERY.

RECOVERY: In recovery state, unless Object_Found or Object_Left_Area messages aren't received from the sink node, sensor node stays in this state. If RSS value is greater than threshold value then it changes its state to DETECTED and sends data about object's position.

2) Messages:

Time_Out: When the predefined timeout value in a node runs out the node is given the timeout signal.

Wake_Up: Sink node sends this message to nodes in the predicted path.

On_Predicted_Path: This message is for nodes that no longer sense the object but may still be on the predicted path and be required to continue to sense.

RSS>Threshold: When the RSS value of a node is above the threshold value it means it can start tracking the object.

RSS<Threshold: When the RSS value of a node is below the threshold value it means it no longer tracks the object.

Object_Lost: This message is sent to specified nodes depending on the stage of the recovery mode to inform that recovery mode is initiated.

Object_Found: This message is sent in recovery mode to searching nodes which were unsuccessful in sensing the object.

Object_Left_Area: This message is sent to all sensor nodes if the recovery mode fails to recover the object in a specified time.

VI. Experiment Results

Different scenarios were setup and each one involved objects with two different speeds: 5 m/s and 16 m/s. The first one stands for low speed objects and the latter stands for high speed objects respectively. The main purpose was to compute average energy loss and average recovery time after measuring their absolute values by running each simulation scenario for 5 times. The experiments were also conducted with medium speeds, i.e., 8 m/s and 12 m/s.

At the end of each simulation, data on total energy spent and time lengths of recovery for each object were obtained. The data gathered from the simulation was analysed and put on graphs for better visualization. The object moves following linear paths and bounces off at area boundaries. For localization, trilateration [28] is used to determine target

position along with RSS value to measure distance to the target.

Simulations were mostly conducted on one (main) topology and later on the simulation was extended to include three more different topologies. The decision to extend the simulation emerged from the need to compare obtained results with results from a different deployment model (random). The main topology used is given in Fig. 1. This first topology (referred to as Topology1) includes nodes that are uniformly distributed in the area.

In the second topology (Topology2), the field was increased to an area of 100 x 100 m^2, increasing the length between nodes but keeping the same formation as the first topology. The third topology (Topology3) covers randomly distributed nodes in a 50 x 50 m^2 field. The last topology (Topology4) involves nodes randomly distributed in a field of 100 x 100 m^2.

By conducting various simulation experiments, the aim was to measure the success of our approach and to compare the results for low speed and high speed objects. In addition, two modes of operation have been taken into consideration, namely powersaving and non powersaving modes for the operation of the network. Each node in the network has the same starting energy of 20 Joules at the beginning of any simulation. The average remaining energy per node, average recovery delay, and packet delivery ratio results of simulations conducted on Topology1 are given in Fig. 3, Fig. 4, and Fig. 5.

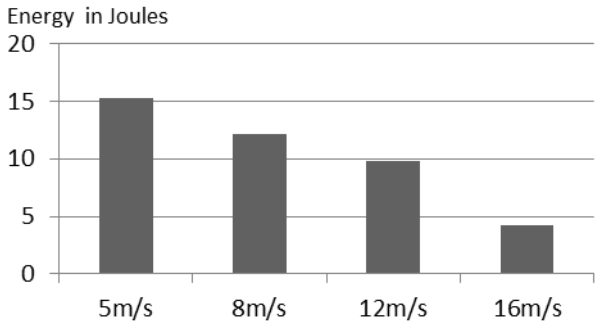


Fig. 3 Average Remaining Energy per Node

In Fig. 3, the displayed energy levels regarding the related speed values of the moving object belong to the average remaining energy per node. The remaining energy value of each node is measured at the end of each simulation and the corresponding average remaining energy value per node is calculated for the whole network. The average values of the calculation results for 5 repeated simulations are given in Fig. 3. The relationship between object speed and the total remaining energy of all the nodes can be seen in Fig. 3, object's speed has a direct relationship to energy consumption levels. This is due to increased object loss rates for objects travelling at high speeds. As the recovery frequency increases, so does the total energy consumption. Taking the results shown in Fig. 3 as an example, for an object travelling at 16 m/s the average remaining energy drops as low as 4.2 Joules, where the energy consumption in the whole network is greater than the other cases for slower speeds.

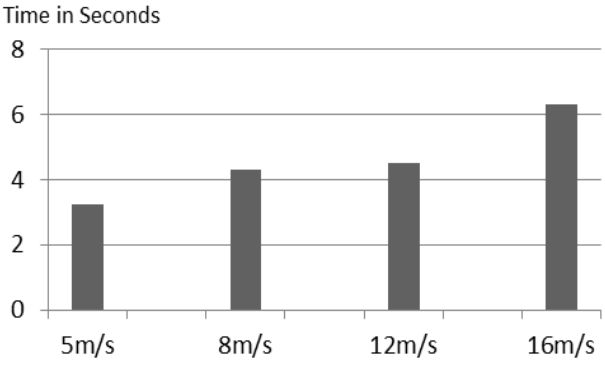


Fig. 4 Average Recovery Time

In Fig. 4, we can see the direct correlation between object speed and average recovery time in seconds. For objects travelling at 5 m/s average recovery time is as low as 3.23 seconds and for objects travelling at 16 m/s the corresponding duration is 6.32 seconds which is significantly higher compared to the former. For objects travelling at 8 m/s and 12 m/s average recovery times of 4.32 and 4.39 seconds are observed respectively.

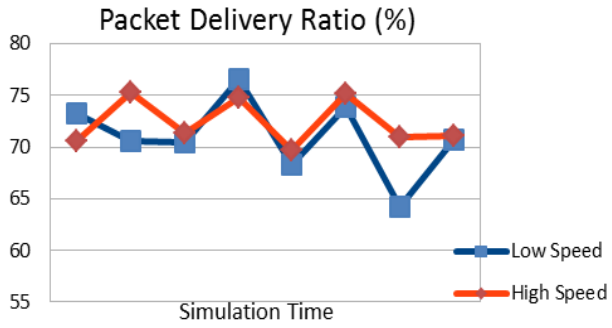


Fig. 5 Ratio of Packets Successfully Delivered

Fig. 5 shows the packet delivery success ratio in the network. Since the prediction in our model is done on the sink node, the tracking object's location data is delivered to the sink when it's detected by any node or group of nodes. Meanwhile, the data may be lost in the network due to the interference and collision of the packets. Besides, IEEE 802.15.4 PHY and MAC layers inherit the packet drops of the wireless media. To us, on average a 72% success ratio for the packet delivery seen in Fig. 5 makes sense regarding the collisions intrinsically available in the wireless channel. We also deduce from Fig. 5 that the oscillation of the packet delivery success during the whole simulation is due to random variation of the number of collided packets at random times.

By adding three other topologies, as mentioned before, we extended our study; Fig. 6 and Fig. 7 show the comparison of results obtained for objects travelling at low speeds and high speeds, respectively. The aforementioned "Powersaving" and "Non Powersaving" terms in Fig. 6 and Fig. 7 refer to the two different running modes of the system explained in Subsection C (Recovery Model) of Section IV.

When running in the non powersaving mode, the system does its best to track and detect the object and uses its all resources to achieve this task. This means the system rigorously awakens all the nodes if necessary regarding the procedures of the recovery algorithm used. On the other hand, if the system is adjusted to run in powersaving mode, the recovery task will be switched into a way to limit the energy consumption at the cost of less accurate tracking. The simulations of which the results are depicted in Fig. 6 and Fig. 7 focus on finding the trade-off between energy conservation and accuracy of target tracking along with how the two perform for objects travelling at high and low speeds.

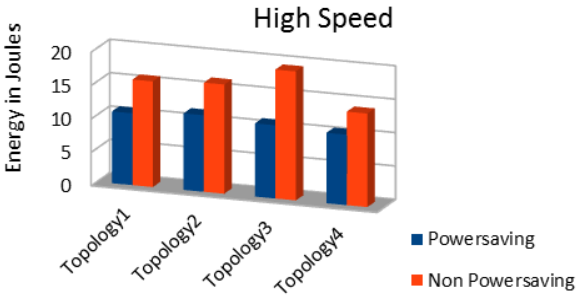


Fig. 6 Average Energy Consumption for High Speed Object

The graph in Fig. 6 shows the amounts of average energy consumed per node with different topological simulation setups for objects travelling at high speeds. Regarding each different topology, energy preservation algorithm makes a considerable difference in terms of energy savings in the network. The disadvantage of powersaving mode compared to non powersaving mode is reduced tracking accuracy. All four topologies have similar energy consumption levels for powersaving mode, whereas for non powersaving mode, energy consumption is more varied across different topological setups. Topology3 has the highest amount of total energy consumption for non powersaving mode.

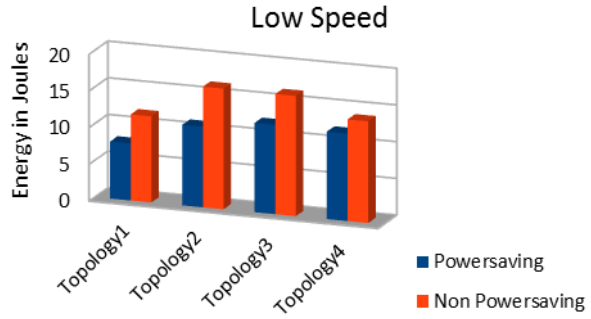


Fig. 7 Average Energy Consumption for Low Speed Object

The top-view of the graph in Fig. 7 seems similar with the one in Fig. 6, except that the exact results are different for the case that the object moves at a low speed. According to Fig. 7, all eight different simulation scenarios (four topologies within two operating modes) show varied energy levels. The minimum energy consumption in total for both running modes is observed in Topology1, whereas the maximum energy consumption is measured in Topology3 among the eight different scenarios.

The disadvantage of powersaving mode is its reduced tracking accuracy due to the adjusted awakening algorithm involved in the recovery model to limit energy consumption. As a result of the trade-off observed between tracking accuracy and energy consumption, the advantage of powersaving mode is significantly less energy consumption. As evident in Fig. 6 and Fig. 7, regarding the energy consumption levels, powersaving mode has a distinctive advantage both for objects travelling at high speeds and low speeds. It can also be inferred from Fig. 6 and Fig. 7 that non powersaving scenarios in which the tracking object moves with high speeds usually consume more energy than the non powersaving scenarios involving a tracking object with low speeds. This is mainly because of the higher number of recovery stages incurred in high speed scenarios. By using powersaving mode in such high speed scenarios, we can limit energy consumption to the levels observed in low speed scenarios.

VII. Conclusion and Future Work

This paper presents a network simulation based study aiming to answer the question of "how can energy efficiency be improved in an OTSN through implementing various scenarios for object tracking". Having presented our initial work at the national conference of Academic Computing in Turkey [29], we extend our work to include the proposed network operation and simulation results. This network setup was simulated in ns-2.34 and associated graphs of obtained results are given.

Different scenarios were considered where target speeds and moving patterns vary. The main focus has been on energy conservation which has been mostly achieved by reducing the average energy consumed by sensor nodes.

As future work, we wish to improve our tracking method to track multiple objects simultaneously. We also think that, a dynamic clustering of nodes before or after the prediction phase may make a difference in terms of improving tracking success and energy efficiency.

Acknowledgment

This project initially started as a senior project at Izmir University of Economics involving six people in total. We would like to thank to Utkan Surgevil, Yetkin Hafizoglu and Nihal Pacaman for their efforts.

References

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," Computer Networks, vol. 38, pp. 393-422, 2002, http://dx.doi.org/10.1016/S1389-1286(01)00302-4

[2] H. Karl and A. Willig, Protocols and Architectures for Wireless Sensor Networks: John Wiley & Sons, 2005, http://dx.doi.org/10.1002/0470095121

[3] Y. Xu, J. Winter, and W.-C. Lee, "Dual prediction-based reporting for object tracking sensor networks," in The First Annual International Conference on Mobile and Ubiquitous

Systems: Networking and Services, MOBIQUITOUS 2004, Aug. 2004, pp. 154-163, http://dx.doi.org/10.1109/MOBIQ.2004.1331722

[4] Y. Xu, J. Winter, and W.-C. Lee, "Prediction-based strategies for energy saving in object tracking sensor networks," in *Proceedings of IEEE International Conference on Mobile Data Management*, 2004, pp. 346-357, http://dx.doi.org/10.1109/MDM.2004.1263084

[5] M. Z. A. Bhuiyan, G.-J. Wang, L. Zhang, and Y. Peng, "Prediction-based energy-efficient target tracking protocol in wireless sensor networks," *Journal of Central South University of Technology,* vol. 17(2), pp. 340-348, 2010, http://dx.doi.org/10.1007/s11771-010-0051-1

[6] F. Deldar and M. H. Yaghmaee, "Designing an energy efficient prediction-based algorithm for target tracking in wireless sensor networks," *2011 International Conference on Wireless Communications and Signal Processing (WCSP),* Nov. 2011, pp. 1-6, http://dx.doi.org/10.1109/WCSP.2011.6096835

[7] H. T. Kung and D. Vlah, "Efficient location tracking using sensor networks," *2003 IEEE Wireless Communications and Networking, WCNC 2003,* March 2003, pp. 1954-1961 vol. 3, http://dx.doi.org/10.1109/WCNC.2003.1200686

[8] W. Z. W. Zhang and G. C. G. Cao, "DCTC: dynamic convoy tree-based collaboration for target tracking in sensor networks," *IEEE Transactions on Wireless Communications,* vol. 3(5), pp. 1689-1701, 2004, http://dx.doi.org/10.1109/TWC.2004.833443

[9] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, Jan. 2000, http://dx.doi.org/10.1109/HICSS.2000.926982

[10] S. P. M. Tran and T. A. Yang, "OCO: Optimized communication & organization for target tracking in wireless sensor networks," in *Proceedings of the IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC'06)*, June 2006, pp. 428-435, http://dx.doi.org/10.1109/SUTC.2006.1636209

[11] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," *Wireless Communications, IEEE Transactions on,* vol. 1(4), pp. 660-670, 2002, http://dx.doi.org/10.1109/TWC.2002.804190

[12] M. J. Handy, M. Haase, and D. Timmermann, "Low energy adaptive clustering hierarchy with deterministic cluster-head selection," in *Mobile and Wireless Communications Network, 2002. 4th International Workshop on*, 2002, pp. 368-372, http://dx.doi.org/10.1109/MWCN.2002.1045790

[13] H. Yang and B. Sikdar, "A protocol for tracking mobile targets using sensor networks," in *Proceedings of 2003 IEEE International Workshop on Sensor Network Protocols and Applications*, May 2003, pp. 71-81, http://dx.doi.org/10.1109/SNPA.2003.1203358

[14] M. Guo, E. Olule, G. Wang, and S. Guo, "Designing energy efficient target tracking protocol with quality monitoring in wireless sensor networks," *The Journal of Supercomputing,* vol. 51(2), pp. 131-148, 2010, http://dx.doi.org/10.1007/s11227-009-0278-5

[15] F. Zhao, J. Shin, and J. Reich, "Information driven dynamic sensor collaboration," *IEEE Signal Processing Magazine,* vol. 19(2), pp. 61-72, 2002, http://dx.doi.org/10.1109/79.985685

[16] P. V. Pahalawatta, T. N. Pappas, and A. K. Katsaggelos, "Optimal sensor selection for video-based target tracking in a wireless sensor network," in *2004 International Conference on Image Processing, ICIP '04,* Oct. 2004, pp. 3073-3076 Vol. 5, http://dx.doi.org/10.1109/ICIP.2004.1421762

[17] Z. Guo, M. Zhou, and L. Zakrevski, "Optimal tracking interval for predictive tracking in wireless sensor network," *Communications Letters, IEEE,* vol. 9(9), pp. 805-807, 2005, http://dx.doi.org/10.1109/LCOMM.2005.1506709

[18] X. Wang, L. Ding, D. Bi, and S. Wang, "Energy-efficient optimization of reorganization-enabled wireless sensor networks," *Sensors,* vol. 7(9), pp. 1793-1816, 2007, http://dx.doi.org/10.3390/s7091793

[19] S. Samarah, M. Al-Hajri, and A. Boukerche, "A predictive energy-efficient technique to support object-tracking sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 60(2), pp. 656-663, 2011, http://dx.doi.org/10.1109/TVT.2010.2102375

[20] Y.-L. Chen, Y.-C. Lin, and T.-C. Sun, "A prediction scheme for object tracking in grid wireless sensor networks," in *Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS)*, July 2013, pp. 360-364, http://dx.doi.org/10.1109/IMIS.2013.67

[21] A. Alaybeyoglu, A. Kantarci, and K. Erciyes, "A dynamic lookahead tree based tracking algorithm for wireless sensor networks using particle filtering technique," *Computers & Electrical Engineering*, vol. 40(2), pp. 374-383, 2014, http://dx.doi.org/10.1016/j.compeleceng.2013.06.014

[22] M. Mirsadeghi and A. Mahani, "Energy efficient fast predictor for WSN-based target tracking," *Annals of Telecommunications*, pp. 1-9, 2014, http://dx.doi.org/10.1007/s12243-014-0430-y

[23] Mica2 Datasheet, http://www.eol.ucar.edu/isf/facilities/isa/internal/CrossBow/DataSheets/mica2.pdf . (Last accessed: March 2014)

[24] W.-P. Chen, J. C. Hou, and L. Sha, "Dynamic clustering for acoustic target tracking in wireless sensor networks," *IEEE Transactions on Mobile Computing,* vol. 3(3), pp. 258-271, 2004, http://dx.doi.org/10.1109/TMC.2004.22

[25] Wireless medium access control and physical layer specifications for low-rate wireless personal area networks. IEEE Standard, 802.15.4-2003, May 2003. ISBN 0-7381-3677-5.

[26] Network Simulator 2 (ns-2), www.isi.edu/nsnam/ns/. (Last accessed: March 2014)

[27] Wireless medium access control and physical layer specifications for low-rate wireless personal area networks. IEEE Standard, 802.11-2007, June 2007. ISBN 0-7381-5656-6.

[28] B. Krishnamachari, S. B. Wicker, and R. Bejar, "Phase transition phenomena in wireless ad hoc networks," in *IEEE Global Telecommunications Conference, GLOBECOM '01*, Nov. 2001, pp. 2921-2925 vol.5, http://dx.doi.org/10.1109/GLOCOM.2001.965963

[29] G. Peynirci, M. Gürgen, İ. Korkmaz, Y. Hafizoğlu, U. Sürgevil, and N. Paçaman, "An object tracking scenario using wireless sensor networks," presented at the *Academic Computing*, Turkey, 2010.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1059–1064
DOI: 10.15439/2014F252
ACSIS, Vol. 2

Power aware MOM for telemetry-oriented applications using GPRS-enabled embedded devices – levee monitoring use case

Tomasz Szydlo *, Piotr Nawrocki † and Robert Brzoza-Woch ‡ and Krzysztof Zielinski §
*AGH University of Science and Technology
Faculty of Computer Science, Electronics and Telecommunications
Department of Computer Science
al. A. Mickiewicza 30, 30-059 Krakow, Poland
e-mail:tomasz.szydlo@agh.edu.pl
†e-mail:piter@agh.edu.pl
‡e-mail:rabw@agh.edu.pl
§e-mail:kz@agh.edu.pl

***Abstract*—The paper proposes the concept of adaptive message aggregation for telemetry applications that use GPRS connectivity. The method optimizes the power consumed during data transmission, what is useful in the modern telemetry devices powered from renewable energy sources. The concept has been verified in the levee monitoring scenario.**

I. Introduction

THE PURPOSE of telemetry systems is to transparently convey measurement information from a remotely located sensor to receiving equipment for further processing and visualization. Development and miniaturization of electronic devices has allowed for the high penetration of telemetry solutions in the surrounding world in order to increase the quality of life. In typical telemetry solutions, remote stations are powered from external power sources and use industrial communication protocols such as Modbus to gather data from these devices to the central system. The communication to the remote location is achieved by the GPRS network as a low cost and easily accessible communication layer [1]. These protocols provide data by polling, so it requires the remote stations to be available all the time.

Currently, such devices are designed to be powered by energy harvesting thus they must be power efficient and they might temporarily go asleep to preserve power [2], [3]. Because of the differences between these types of devices, the legacy polling protocols might not be effective. We argue that the communication protocols should (1) leverage the power usage characteristic of GPRS technology, (2) handle the sleepy nodes and (3) provide high level addressing of nodes. We think that these requirements might be fulfilled by the message oriented communication. Sending messages across channels decreases the complexity of the end application, thereby allowing the developer of the application to focus on true application functionality instead of the intricate needs of communication protocols. *Message-oriented middleware* (MOM)[4] allows application modules to be distributed over heterogeneous platforms and reduces the complexity of developing applications that span multiple operating systems and network protocols. The middleware creates a distributed communications layer that insulates the application developer from the details of the various operating systems and network interfaces. Message-oriented middleware may provide reliable asynchronous communication mechanisms that might be used to carry i.e. measurement data or other remote communication messages.

In the paper we propose the concept of adaptive message aggregation method for MQTT-SN protocol that optimizes power used by GPRS wireless connection during data transmission. The preliminary research (presented later in the paper) showed that sending data using short IP packets consumes much more energy than using longer packets. Because of the fact that messages containing measurements are relatively small, we propose the concept of adaptive data aggregation prior to sending via GPRS.

The research presented in this paper is a part of ISMOP [5] research project which objectives span construction of an artificial levee, design of wireless sensors for levee instrumentation, development of a sensor communication infrastructure, and a software platform for execution management, data management [6] and decision support [7]. Scientific and industrial consortium in the ISMOP project conducts research on a comprehensive monitoring system enabling evaluation of current and forecasted state of flood levees. This paper focuses on issues related to the organization of data acquired from the sensors located in the levees in order to optimize the power consumed by GPRS modem during data transmission to the central system for later analysis.

The paper is organized as follows. Section II discusses the motivating scenario, where Section III presents the related work. Section IV presents the concept of power-aware adaptive message aggregation, which is then evaluated in the use case

978-83-60810-58-3/$25.00

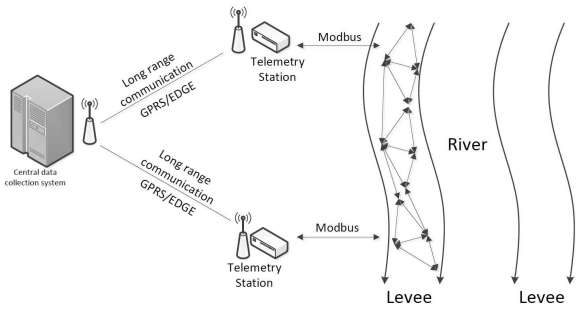


Fig. 1. Telemetry system for flood control levees

in Section V. Finally, the paper is summarized and further research steps are presented.

II. Motivating scenario

Recently, the importance of sensor network for monitoring various areas, objects or devices, has significantly increased. One of the areas in which telemetry and sensor network begin to fulfill a major role is monitoring systems for hydrologic engineering facilities in particular dams and flood levees[8], [9]. The overall concept of hydrological monitoring facilities was the starting point for the assumptions and implementation of the ISMOP research project which will result in guidelines for creating a telemetry system that enables continuous monitoring of levees. Research addresses the collection of massive measurement data in continuous mode, optimized transmission methodology, interpretation and analysis of monitored data with computer simulation and finally providing visualized results for the relevant authorities.

The biggest threat in the context of the levees is their leaking which could lead to their break. A typical application in this area is monitoring system of flood levees containing sensors that measure, among other things, temperature inside the levee. This measurement allows us to detect the rate of change of temperature in the levee. In the case of leakage when the water entering inside of the levee, water is the carrier of heat energy and increases or decreases the temperature in the levee. During the continuous measurement of temperature anomalies can be observed for a specific point in the levee, which deviate from the results of measurements made by the nearby sensors and observed the trend changes. This may indicate the presence of too much water in the levee and thus the danger of burst of the levee. The temperature and other physical values (such as pore pressure) allow to determine the condition of the levees and detect potential risks of leakage. An example scenario of telemetry system containing sensors placed inside the levees and the central data collection system is shown in Figure 1.

The role of the telemetry station is to acquire data from sensor networks which contain information about levee condition and to transmit these data to the central station. The data is generated by the sensor network for an *epoch*, which results in bursts of data each time the epoch changes. Apart from that, telemetry station also sends periodically information about its current condition e.g. battery level, CPU usage and others. The data from sensor networks, due to their importance, should be reliably delivered while the condition information may be transmitted on the best effort basis.

The use of telecommunications networks (including GPRS) for communication from/to telemetry stations is associated with specific energy consumption. It depends mostly on the time during which radio circuits are powered on and on the number of transmitted data. The motivation to our research is to reduce energy consumption during communication using GPRS from/to telemetry station.

III. Related Work

Energy consumption of a wireless transmission device greatly depends on such factors as chosen communication standard, protocols used, and amount of transmitted data. Providing a medium-range or wide-area network connectivity requires a different approach. It is not a demanding task provided that a network infrastructure is available with appropriate SLA (Service Level Agreement) guaranties [10]. However, in remote areas a cellular connection is a common solution for industrial telemetry systems. Typical activities on a smartphone platform (sending a message, making a voice call, transmission over GPRS, etc.) are evaluated in [11]. An in-depth analysis o energy requirements for GPRS and UMTS services is provided in [12] and [13].

High-level protocols over cellular network also have an impact on overall energy requirements of a system [14]. A review of various middleware protocols for telemetry applications can be found in [15]. Message-oriented Middleware (MOM) is widely used as a communication layer for a variety of information systems which require event-driven message and data exchange, and more loose coupling than e.g. remote procedure calls. Examples of commonly utilized technologies for MOM are:

- Java Message Service (JMS) [16],
- Data Distribution Service [17],
- Extensible Messaging and Presence Protocol (XMPP) [18],
- MQ Telemetry Transport (MQTT) and its variation, MQTT-SN [19].

These technologies provide several other functionalities such as transaction management, broker clustering, additional message paradigms including point-to-point, publish/subscribe and request-response. Nevertheless, only MQTT has been designed especially for transferring telemetry-style binary data from the pervasive devices with limited computational resources. It should be noted that utilizing the MQTT protocol over a standard TCP connection may provide redundant message delivery guaranties. As TCP is intended to provide a reliable link and has built-in retransmission mechanisms, setting the MQTT's QoS parameter to 1 or 2 provides another (redundant) layer of persistence. In contrast, those higher levels of QoS seem very useful in MQTT-SN variation which by design uses UDP datagrams. In our research we have chosen MQTT-SN messaging protocol (formerly MQTT-S [20]) because it is promising due to its simplicity. MQTT-SN clients can

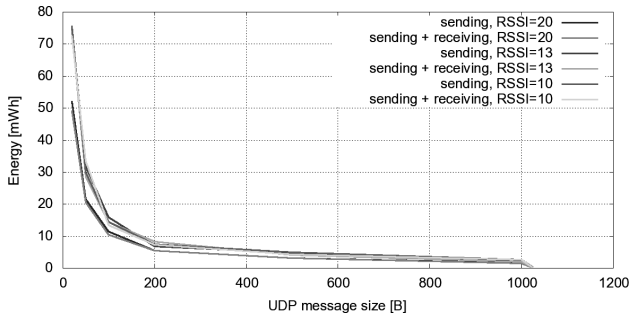


Fig. 2. Power necessary to send 10kB of data using GPRS communication as a function of packet size

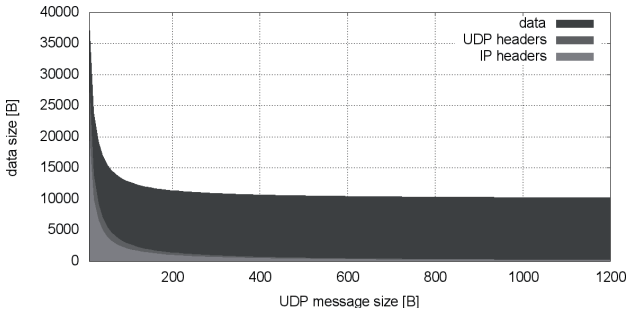


Fig. 3. Overall data necessary to send 10kB of data as a function of packet size

be implemented in resource-constrained hardware (embedded systems), and there are available plenty of its implementations.

However, it is difficult to find any power efficiency considerations for MQTT-SN. By far many solutions dedicated to the MOM technology have been optimized to limit the data transfer and save energy. The MQTT-SN is an example of protocol that was optimized in terms of quantity of data to be transmitted. It results from using short-distance wireless protocols for sensor-based data transmission, including the IEEE 802.15.4 protocol. All these issues, not previously mentioned in MOM solutions, and particularly in the MQTT-SN protocol, have been analyzed in this article and relevant solutions have been suggested.

IV. Adaptive message aggregation

The main goal of the research was to decrease the amount of energy necessary to send the data using MOM over GPRS connectivity. The results of the base research aimed to analyze how much energy is used by GPRS modem as a function of packet data size is depicted in Figure 2.

The nonlinearity in the power consumption is caused by two factors: overhead of the appropriate headers of OSI/ISO stack and purely technical considerations related to the physical communication with the GPRS modem in embedded devices (e.g. the time of data preparation, inter frame gaps and others). Figure 3 shows the overall data size that are necessary to send 10 KB of data payload. The overhead is related to the headers of UDP and IP protocols for each packet. Consequently, from an energy consumption point of view, it is much better for fixed amount of data to send it using as large packet size as possible.

On the other hand, the amount of measurement data that need to be transmitted is usually small i.e. typically 20 B. The previous measurements show that sending such small packets would be inefficient. Based on these data, we propose the concept of data aggregation before transmission.

Our concept, as presented in Figure 4, can be applied to MQTT messages with QoS 0 and 1. In QoS 0, messages are not acknowledged, so client may aggregate several messages before sending them. In the second case, with QoS 1, all the messages had to be acknowledged so, we propose also aggregating the acknowledgment packets on the broker side.

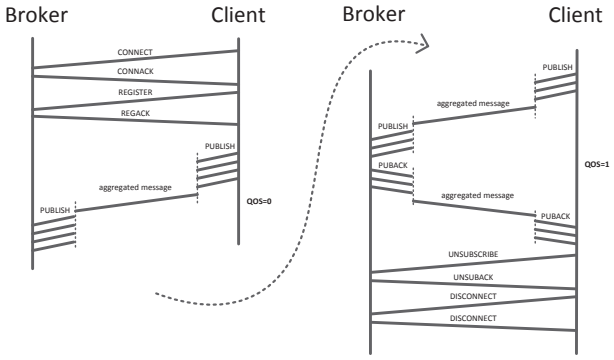


Fig. 4. Message aggregation concept for MQTT-SN

We have assumed that larger, important data to send appear in the bursts, while less important data are sent on regular basis in small chunks. This is dictated by the fact that underlying sensor network (installed in the levee) wakes up in time intervals to preserve power. The naive approach to data aggregation is presented in Figure 5. The main idea is that new messages are not send immediately but first copied to the buffer B with fixed length L and then sent as an aggregated packet. There are two conditions that decide of sending data: buffer is overflowed or buffer timeout T^I has ended. During the *period 1* and *3* aggregated messages are sent because of the timeout condition, while during the *period 2* aggregated message is sent because of the overflow condition. The drawback of the method is that in the *period 3*, the messages that belong to the burst are sent with longer delay then previous ones because overflow condition did not occurred. Such a situation is unwanted if the data has to be analyzed in the real-time.

In our method, we propose adaptive timeout calculation that adjust itself to the frequency of incoming data. The main concept is that buffer overflow situation decreases the buffer timeout meaning that messages should be send faster, while decreasing the frequency of incoming new data recovers the timeout to its previous value. Such a policy results in the situation that messages belonging to the data burst are received by the broker in the burst as well. The concept is depicted in

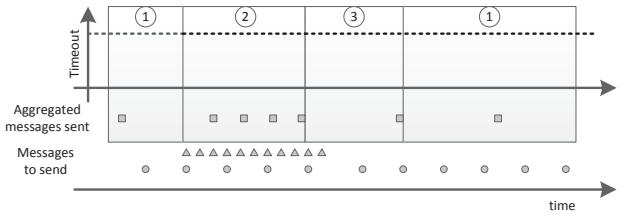


Fig. 5. Naive approach to data aggregation (time periods are marked with numbers, telemetry data are represented by small grey circles, triangles represent bursts of measurement data, packed and transmitted data are depicted as squares)

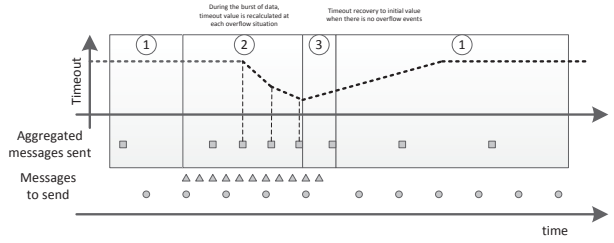


Fig. 6. Adaptive approach to data aggregation (time periods are marked with numbers, telemetry data are represented by small grey circles, triangles represent bursts of measurement data, packed and transmitted data are depicted as squares)

Figure 6. During the *period 3*, the aggregated message is send earlier than in naive approach to the aggregation.

As can be noticed, the presented adaptive message aggregation does not generate any additional overhead. On the contrary, it contributes to reducing the overhead introduced by the OSI/ISO network communication stack when transmitting small chunks of data. Another aspect is the latency of data arrival. It results directly from the fact, that data needs to be packed in a buffer. In our solution, the latency is controlled with the variable timeout value. The timeout automatically decreases as more data arrives. This can be utilized to keep latency and overhead ratio at reasonable levels.

More formally, the algorithm is composed of the two parts and might be presented as follows. The input to the algorithm is provided by four values: T^I is the initial timeout value, T^R is the recovery time to the initial value, the factor α, and a buffer length L. In the MOM client there is global timer T that represents actual timeout value – at time t this value is denoted as T_t. When the aggregated message buffer B of length L is created at time t, it has assigned timeout that equals T_t. Length of the buffers is constant.

First part of the algorithm is responsible for recovering (i.e. increasing) timeout T to the initial value of T^I and is formulated as follow: for each time k, the timeout T_{k+1} is calculated using the equation 1, where ΔT is the time step.

$$T_{k+1} = \min(T_k + \frac{T^I}{T^R}\Delta T, T^I) \tag{1}$$

The second part of the algorithm is responsible for decreasing timeout T to the value that is similar to the time of sending overflowed buffers when data burst is observed. The equation 2 is used only when the overflow of the buffer is observed. Value d in the equation is the time from the last overflow event.

$$T_{k+1} = \min(\alpha T_k + (1-\alpha)d, T^I) \tag{2}$$

Having in mind, that data usually comes from remote telemetry stations to the central point, we propose to use adaptive aggregation method on the client side to aggregate messages, and to use naive aggregation approach on the broker side to aggregate acknowledgments. The evaluation results of the proposed algorithm are presented in the next section.

V. Evaluation

We have evaluated the proposed concept on the scenario similar to the one presented in the previous section. We have assumed that data from levee monitoring sensors are gathered and sent in two stages:

- at the beginning, for QoS 1 in MQTT/MQTT-SN, 1000 PUBLISH messages with a length of 20 B are sent and received confirmation of these messages (PUBACK),
- later, for QoS 0 in MQTT/MQTT-SN, in 12 minutes epoch and for every 30 s PUBLISH messages with a length of 20 B are transmitted.

During tests we used a popular GPRS modem (SIM900D) and, in order to verify the results obtained, we also used an industrial GPRS Modem (Wavecom Fastrack Supreme 20). In order to develop test software we extend implementation of MQTT-SN – Eclipse Mosquitto [21] (which we call *A-MQTT-SN*) to support adaptation.

Above presented testing scenario was carried out for three cases using:

- MQTT protocol (Eclipse Mosquitto),
- MQTT-SN protocol (Eclipse Mosquitto),
- A-MQTT-SN protocol with adaptation for sent and received data (message type PUBLISH and PUBACK).

The adaptive aggregation algorithm for A-MQTT-SN was initiated with values: $TI = 120\ s$, $TR = 240\ s$, $\alpha = 0.5$, and $L = 1000\ B$. The naive aggregation algorithm was initiated with values: $T_I = 2\ s$ and $L = 250\ B$. The values are application-specific, and should be tailored for different conditions, such as: amount of transmitted data, real-time boundaries and the maximal accepted latency by the application.

The measurements were made with a custom multichannel current and voltage sensing module and tailored for energy measurements of various embedded devices. Data for all of the presented tests was acquired from the GPRS modems (Class 10) connecting to public GSM network with a throughput of 25 Kb/s (2 timeslots in uplink direction).

The result of these tests is shown in the following figures:

- for MQTT protocol (using TCP and PPP protocols) – Figure 7,
- for MQTT-SN protocol (using UDP, PPP protocols and AT commands on the GPRS modem) – Figure 8,

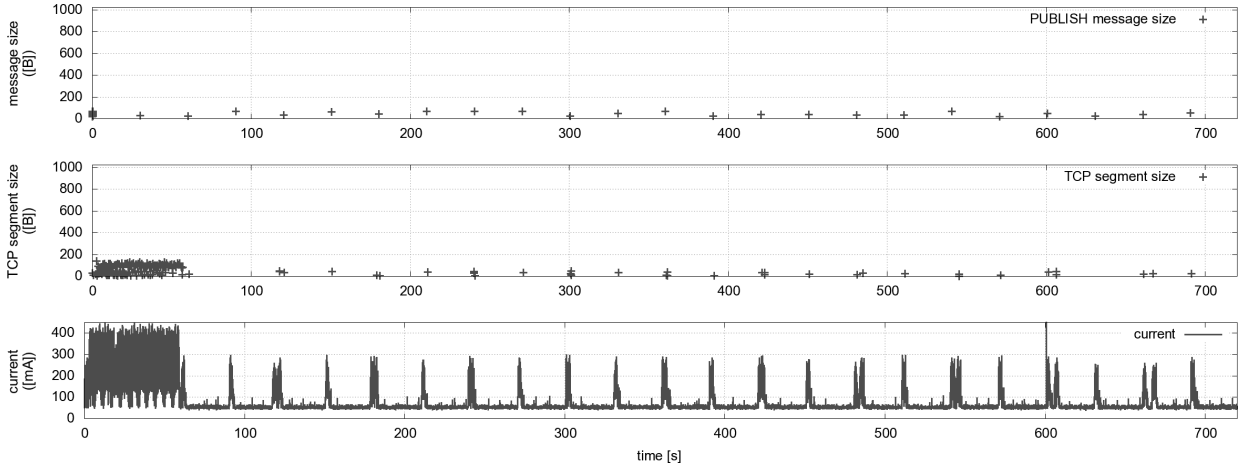


Fig. 7. The current consumption for the MQTT protocol

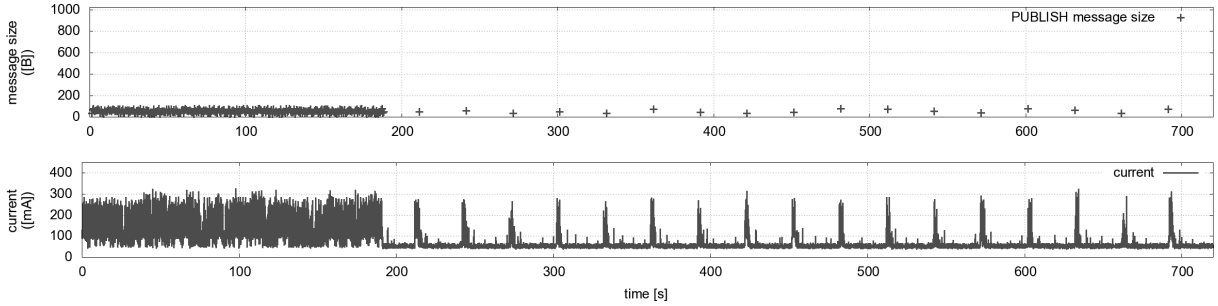


Fig. 8. The current consumption for the MQTT-SN protocol

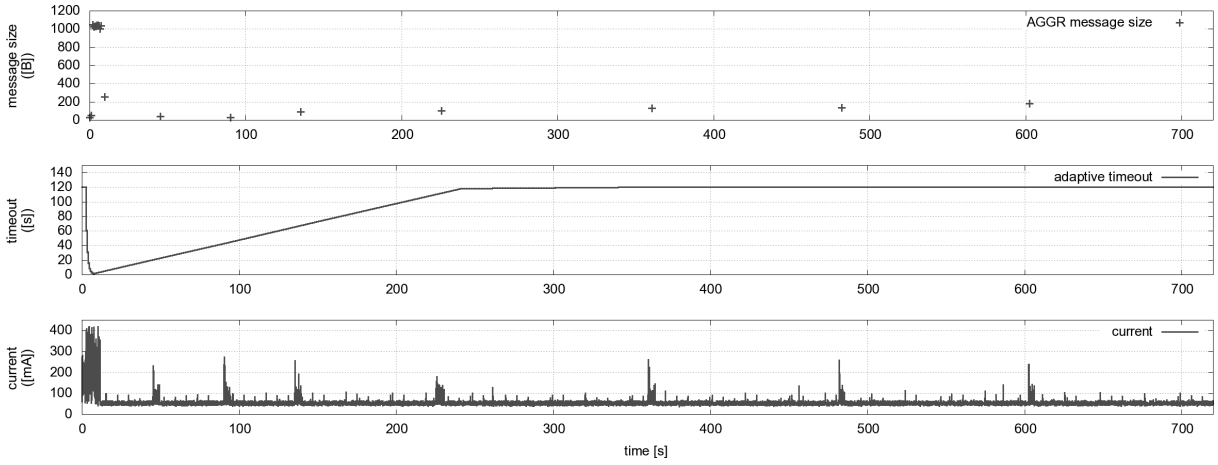


Fig. 9. The power consumption for the A-MQTT-SN protocol

- for A-MQTT-SN protocol (using UDP, PPP protocols and AT commands on the GPRS modem) – Figure 9.

Analyzing the results we can observe that the power consumption of a GPRS modem for data transmission is higher for MQTT and MQTT-SN than A-MQTT-SN protocol. In our opinion higher power consumption for MQTT protocol is the result of using TCP and its complexity (call setup, retransmissions). It can be seen that in the case of sending

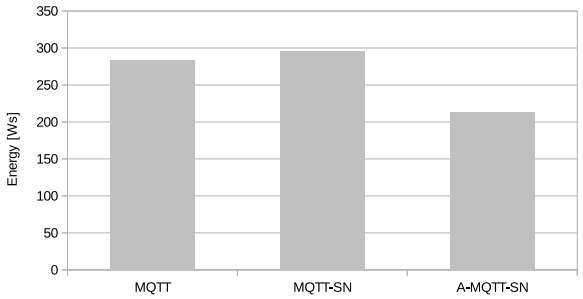


Fig. 10. Energy consumption for MQTT, MQTT-SN and A-MQTT-SN protocols

MQTT messages with QoS 0 should not be a retransmission of messages which, however, are made because of the use of TCP protocol. In the second case, where we used UDP and MQTT-SN protocols, increased energy consumption via modem is related to the size of transmitted data. In case of sending a large number of small packets the overhead associated with packet header is visible. The developed A-MQTT-SN protocol variation aggregates data transmitted and significantly decreases the number of headers and thus gives the best result as regards the energy consumption during transmission by GPRS modem (Figure 10).

VI. Summary and Future Work

The paper discusses the problem of sending the data using GPRS connectivity from remote telemetry stations. We have analyzed various communication protocols and finally selected MQTT-SN. The paper proposed the extensions to the communication protocol that adjust its behavior to the GPRS connectivity profile in order to decrease the data transmission-related energy consumption.

The motivating scenario presented in the paper is only one of the possible applications of our concept. The solutions might be successfully applied to e.g. multilayer telemetry solutions where due to the sleepy nodes, data have to be pushed rarely but efficiently.

Acknowledgment

The research presented in this paper was partially supported by the National Centre for Research and Development (NCBiR) under Grant No. PBS1/B9/18/2013 and by the Polish Ministry of Science and Higher Education under AGH University of Science and Technology Grant 11.11.230.015 (statutory project).

References

[1] M. Szmechta and P. Aksamit, "Modeling packet delay distributions in an industrial telemetry system," in *Computational Intelligence and Intelligent Informatics (ISCIII), 2011 5th International Symposium on*. IEEE, 2011, pp. 71–74.

[2] T. Szydlo, S. Gut, and B. Puto, "Smart Applications: Discovering and interacting with constrained resources IPv6 enabled devices," *Przeglad Elektrotechniczny*, pp. 221–226, 06 2013.

[3] T. Szydlo, P. Suder, and J. Bibro, "Message Oriented Communication For IPV6 Enabled Pervasive Devices," *Computer Science*, vol. 14, no. 4, 2013.

[4] E. Curry, "Message-Oriented Middleware," in *Middleware for Communications*, Q. H. Mahmoud, Ed. Chichester, England: John Wiley and Sons, 2004, ch. 1, pp. 1–28.

[5] "ISMOP Project," www.ismop.edu.pl, 2013, accessed: 2014-04-19.

[6] A. Piórkowski and A. Leśniak, "Using data stream management systems in the design of monitoring system for flood embankments," *Studia Informatica*, vol. 35, no. 2, pp. 297–310, 2014.

[7] M. Chuchro, M. Lupa, A. Pięta, A. Piórkowski, and A. Leśniak, "A concept of time windows length selection in stream databases in the context of sensor networks monitoring," in *New Trends in Databases and Information Systems, Proceedings of 18th East-European Conference on Advances in Databases and Information Systems (in print)*, ser. Advances in Intelligent Systems and Computing. Springer, 2015.

[8] B. Balis, T. Bartynski, M. Bubak, G. Dyk, T. Gubala, and M. Kasztelnik, "A development and execution environment for early warning systems for natural disasters," in *Cluster, Cloud and Grid Computing (CCGrid), 2013 13th IEEE/ACM International Symposium on*, May 2013, pp. 575–582.

[9] B. Balis, M. Kasztelnik, M. Bubak, T. Bartynski, T. Gubaa, P. Nowakowski, and J. Broekhuijsen, "The urbanflood common information space for early warning systems," *Procedia Computer Science*, vol. 4, no. 0, pp. 96 – 105, 2011, proceedings of the International Conference on Computational Science, {ICCS} 2011.

[10] J. Kosinski, P. Nawrocki, D. Radziszowski, K. Zielinski, S. Zielinski, G. Przybylski, and P. Wnek, "SLA Monitoring and Management Framework for Telecommunication Services," in *Networking and Services, 2008. ICNS 2008. Fourth International Conference on*, J. Bi, K. Chin, C. Dini, L. Lehmann, and D. C. Pheanis, Eds. IEEE Computer Society, 2008, pp. 170–175.

[11] G. P. Perrucci, F. H. Fitzek, and J. Widmer, "Survey on energy consumption entities on the smartphone platform," in *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*. IEEE, 2011, pp. 1–6.

[12] A. Sikora, A. Yunitasari, and M. Dold, "GPRS and UMTS services for ultra low energy M2M-communication," in *Intelligent Data Acquisition and Advanced Computing Systems (IDAACS), 2013 IEEE 7th International Conference on*, vol. 1. IEEE, 2013, pp. 494–498.

[13] F. Pauls, S. Krone, W. Nitzold, G. Fettweis, and C. Flores, "Evaluation of Efficient Modes of Operation of GSM/GPRS Modules for M2M Communications," in *Vehicular Technology Conference (VTC Fall), 2013 IEEE 78th*. IEEE, 2013, pp. 1–6.

[14] E. J. Vergara Alonso, "Exploiting Energy Awareness in Mobile Communication," 2013.

[15] A. Azzara, S. Bocchino, P. Pagano, G. Pellerano, and M. Petracca, "Middleware solutions in WSN: The IoT oriented approach in the ICSI project," in *Software, Telecommunications and Computer Networks (SoftCOM), 2013 21st International Conference on*. IEEE, 2013, pp. 1–6.

[16] "Java Message Service documentation," http://docs.oracle.com/javaee/5/tutorial/doc/bncdq.html, accessed: 2014-04-19.

[17] "Data Distribution Service ver. 1.2 documentation," http://www.omg.org/spec/DDS/1.2, accessed: 2014-04-19.

[18] "Extensible Messaging and Presence Protocol documentation," http://xmpp.org/xsf/press/2004-10-04.shtml, accessed: 2014-04-19.

[19] "MQ Telemetry Transport (MQTT) documentation," http://mqtt.org/documentation, accessed: 2014-04-19.

[20] U. Hunkeler, H. L. Truong, and A. Stanford-Clark, "MQTT-SA publish/subscribe protocol for Wireless Sensor Networks," in *Communication Systems Software and Middleware and Workshops, 2008. COMSWARE 2008. 3rd International Conference on*. IEEE, 2008, pp. 791–798.

[21] "Mosquitto technology project," http://projects.eclipse.org/projects/ technology.mosquitto, accessed: 2014-04-19.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1065–1071
DOI: 10.15439/2014F246
ACSIS, Vol. 2

A low power Wireless Sensor Node with Vibration Sensing and Energy Harvesting capability

M. Zieliński, F. Mieyeville, D. Navarro
Ecole Centrale de Lyon
Université de Lyon,
Institut des Nanotechnologies de Lyon
Ecully, F-69134, France
Email: mateusz.zielinski@ec-lyon.fr

O. Bareille
Ecole Centrale de Lyon
Université de Lyon,
Laboratoire de Tribologie et Dynamique des Systemes
Ecully, F-69134, France

***Abstract*—This paper describes the design of the wireless sensor network node (WSN) for distributed active vibration control (AVC) system for the automotive application. The approach of the system is presented in details. A WSN node using one piezoelectric element provides several features (sensing, shunting and energy harvesting). Integration of the vibration sensing capability for active vibration control system with the energy harvesting capability is described here. Simulation results are compared with the prototype design.**

I. Introduction

WIRELESS Sensor Networks (WSNs) are made up of intelligent and autonomous nodes. Each node is able to work independently in distributed area, and is an energy aware and a low-power device. Nodes are able to establish autonomously an efficient wireless connection, which is used to send measured values. WSNs are commonly used to monitor the physical or environmental conditions like temperature, sound and vibration [1].

WSN nodes are commonly supplied from batteries. It is the last "wire" which should be cut, to provide long lifetime and maintenance-free systems. Nowadays, batteries need to be charged or replaced [2]. It introduces additional costs in the use of WSNs. Despite this inconvenience WSNs are applied in Structural Health Monitoring systems (SHMs), Environmental Monitoring systems which provide new features compared to existing wired solutions [3].

Active Vibration Control (AVC) systems become a very important issue in many engineering fields (SHM, automotive and industrial applications) which provide smart solutions for vibration and noise damping systems. Several solutions for AVC have been proposed and tested with promising results [4]. However those systems are centralised, wired and need a lot of computing power. Implementation of the AVC in a WSN is a challenge for designers. The entire vibration control system must be designed in order to pass very stringent requirements of the WSN and control law [5].

II. Research Assumptions

The usage of the active vibration control can reduce the weight of conventional passive methods, helping to push towards lighter, more fuel efficient vehicles [6]. Active methods include vibration sources which are driven by the control strategy algorithm to provide destructive interference of real vibrations. There are two possibilities to implement the control strategy: feed-forward (open-loop) and feedback (closed-loop). A variety of algorithms have been used to adapt the controller, most are based on adaptative filtering methods: least mean square (LMS) and filtered reference LMS (FxLMS) [7]. Conventional AVC systems based on wired networks (sensors and actuators) with high-power controller and fast data processing need a lot of energy. Replacing a big centralised (wired) system with a low power nodes can improve the AVC in the scope of functionality, maintenance costs and energy consumption.

III. System approach

WSNs have rather low transmission rates. Due to delays, an implementation of the real-time system which is necessary to provide data processing for centralised AVC systems is not possible [5]. In spite of, WSNs could be used to provide active control. A distributed approach used in place of the centralised one can be a solution. In our approach, intelligent nodes provide local action, which reduces the amount of the information to transfer, compared to the centralised approach. The feasibility of the distributed autonomous nodes with sensing feature coupled to semi-active vibration control dissipation is the aim of the work.

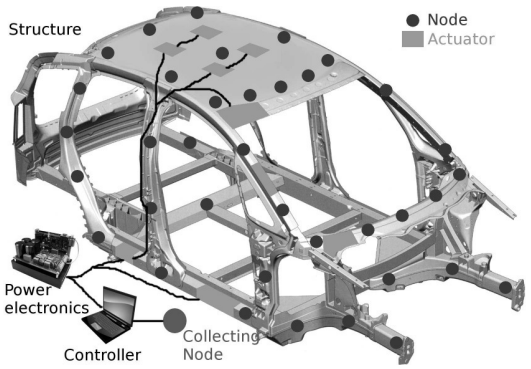


Fig. 1. System structure

978-83-60810-58-3/$25.00

Figure 1 presents the approach of the global system structure for distributed AVC system for automotive application. Wireless nodes are implemented in the body of the car. Nodes are in the star topology towards the collecting node. Each node in the system is powered from the harvested energy and provides mechanical damping. It is a first level of the AVC system. Nodes measure the value of vibrations and if necessary send data to the collecting node, which is responsible for second level of the AVC system (high power, wired actuators). If there are no vibrations in the system, the nodes do not have energy to work but there is no need to cancel the vibrations. When vibrations occur, the WSN nodes are initiated and provide first level of the mechanical damping. If necessary they send measurements to the collecting node, and ask for a second level of the AVC.

IV. WSN NODE DESIGN ASSUMPTIONS

The WSN node provides several hardware features: sensing, energy harvesting and local damping using one piezoelectric element. The following paragraphs describe these features and provide description of the chosen solutions.

A. Sensing vibrations

Sensing vibrations requires understanding two issues. The first one is the environment of vibrations; the second is a piezoelectric effect.

Various vibration noises have been already investigated in literature [7]. Body car vibrations are low frequency signals. Noises inside a car under operating conditions have low frequencies (less than 300Hz) and are mostly determined by acoustic resonances and body vibrations modes [8], [9].

The piezoelectric element can be used to track mechanical vibrations [10]. The output voltage of the piezoelectric element corresponds to the mechanical acceleration. It makes the piezoelectric element proper for sensing vibrations.

B. Energy Harvesting

The piezoelectric element can be described as a current source with an inbuilt capacitance. The existence of this capacitance suggests using the shunt inductor to achieve maximum power flow. However, the large value of the inductor, does not allow to use this solution in the real design. Several solutions for harvesting energy from the piezoelectric element have been already proposed. So-called "synchronised switch damping" (SSD) are semi-passive methods developed to address the problem of vibration damping. Techniques based on SSD method provide efficient energy harvesting by increasing the energy flow between the piezoelectric element and load [10].

C. Mechanical damping

Shunting the piezoelectric element (provided by the energy harvesting method) can be used for structural damping. The efficiency of dissipation energy in the shunt connected to the piezoelectric element is currently an issue of research. We distinguish passive and active methods. Passive methods use passive electrical elements to provide mechanical damping [11] whereas active methods use non-linear circuits like negative-capacitance [12] or switching methods [13].

D. Chosen solutions

The design of an autonomous and intelligent WSN node must consider requirements mentioned in previous paragraphs. Low frequency vibrations in the body of the car suggest the use of efficient energy but slow and low-computing, 8bit low-power microcontroller with internal analog-to-digital converter (ADC). It provides low power consumption. Moreover, the sensing vibrations capability needs the piezoelectric element in the open circuit.

The choice of the energy harvesting method requires to take under consideration several parameters: efficiency of the energy harvesting, capability of vibrations damping, efficiency in the power consumption, facility of implementation and sensing capability. The series switching over the inductance (Series SSHI) method is chosen for the design. It is semi-passive technique, based on non-linear processing on the piezoelectric voltage.

V. WSN NODE DESIGN

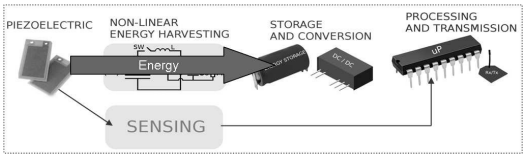


Fig. 2. WSN node schema

Figure 2 presents the schema block for WSN node. We distinguish two parallel systems. The first one for energy harvesting and damping vibrations, the second one for sensing vibrations. Both are connected to one piezoelectric patch transducer. Harvested energy is kept in the storage and used to supply the microprocessor and wireless transmission. The following paragraphs describe the design of the WSN node.

A. Piezoelectric element

Piezoelectric effect can be considered as a bidirectional energy conversion. A strain on the piezoelectric element generates the electrical tension and respectively the electrical tension over the piezoelectric element generates the mechanical strain.

To understand electrical properties of the piezoelectric patch transducer several measurements have been done. Figure 3 and figure 4 presents achieved results.

The current and voltage values are measured over the piezoelectric patch transducer in function of the resistive load for constant frequency (figure 3). The piezoelectric element is a real current source and for the optimal resistive load provides the maximal power (figure 4). It clearly shows that the energy harvesting circuit must be designed in accordance with the electrical properties of the piezoelectric patch (optimal load).

According to figures 3 and 4 we can observe that the high value of the resistive load reduces the amount of the energy

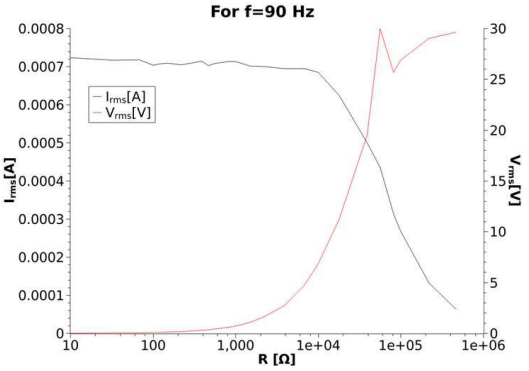


Fig. 3. Output electrical characteristics for piezoelectric patch transducer

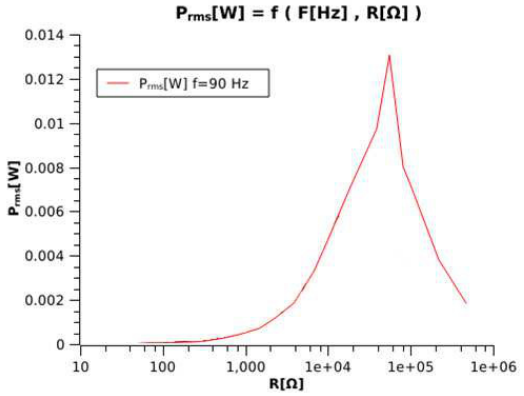


Fig. 4. Power received from the piezoelectric patch transducer in function of resistive load

received from the piezoelectric element. It proves the usage of the piezoelectric element with high resistive load for sensing. This solution provides a good resolution of measurements due to high voltage values and low energy leakage.

B. Sensing vibrations

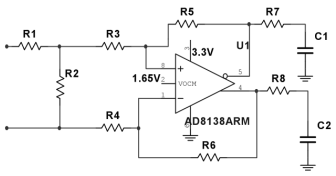


Fig. 5. Sensing circuit

The designed circuit for sensing vibrations and signal conditioning is presented on figure 5. It is composed of the voltage divider (R1 and R2) and the AD8138 low distortion differential analog-to-digital (ADC) driver from Analog Devices. The low pass filter is used to cut-off high frequencies over the output of the ADC driver (R7 with C1 and R8 with C2). The differential ADC driver provides also offset voltage (Pin 2 connected to the 1.65V). Hence, the negative and positive values are measured, it is necessary to provide active vibration control. In the designed circuit the piezoelectric element is not connected directly to the circuit ground. The low-power differential amplifier is supplied from single 3.3V, which simplifies the supply circuit. An internal ADC of the microcontroller is used. This solution provides low energy consumption since there is no additional ADC to supply.

C. Switching circuit for SSHI method

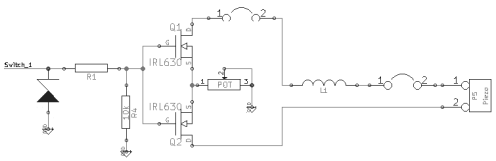


Fig. 6. Switching circuit for SSHI method

The series SSHI circuit is presented on the figure 6. It contains two IRL630 NMOS transistors driven by the microcontroller (the full-wave switch circuit with low current leakage and low short-circuit resistance). The usage of the logic-level transistor simplifies the circuit; the output of the microcontroller can be used to drive the switch. The zener diode D1 is used to protect the microcontroller pin, the R1 resistor sets the current value (it is correlated with the turn-on time). The R4 is used to reduce the turn-off time.

In the series SSHI method, the inbuilt piezoelectric capacitance and external inductance creates the series resonant circuit. The switch keep the circuit in the open-circuit. While the extremum of the mechanical displacement is detected, the switch is closed for half of the electrical resonant period. It causes inversion of the piezoelectric element voltage. The period of the mechanical displacement is much longer than the period of the electrical circuit.

VI. Simulations

Designed circuits: vibration sensing and energy harvesting are simulated in the NI Multisim Component Evaluator 13.0 from National Instruments [14]. Devices used in the design are implemented using SPICE models. The sensing circuit and the series SSHI switching circuit are connected in parallel with the model of the piezoelectric element.

Figure 7a shows simulated signal for sensing vibrations. The 1.65 V offset is used for signal conditioning, hence positive and negative parts of the signal can be measured in the ADC of the microcontroller.

Figure 7b shows simulation results for switching circuit while the switch is open. The sinusoidal signal is a voltage over the piezoelectric element. The second wave - small peaks represents the current in the circuit.

Figure 7c shows simulation results for switching circuit while the switch is closed. The sinusoidal signals presented on figure correspond to the voltage and current in the circuit. The shift phase between them proves existence of the capacity (inbuilt capacity of the piezoelectric element).

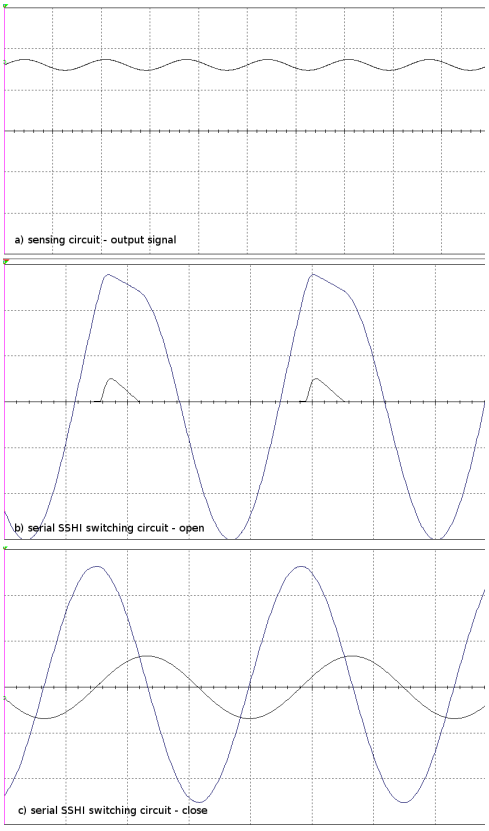


Fig. 7. Simulation results

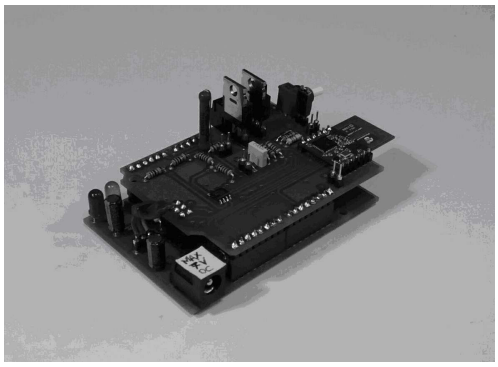


Fig. 8. Photo of the prototype WSN node

VII. Implementation

Figure 8 presents the photo of the real prototype device. It is composed of two PCB boards. The first board contains microcontroller and supply circuit, the second provides the SSHI circuit, sensing circuit and wireless communication. In designed prototype series SSHI circuit and sensing circuit are connected in parallel with the piezoelectric patch transducer which corresponds to the simulations.

For wireless communication we used the MRF24J40 radiofrequency transceiver from Microchip. It provides hardware support for IEEE 802.15.4 physical layer. The node is equipped with low-power microcontroller PIC16LF88 with internal 10bit ADC.

A. Measurement station

Fig. 9. Measurement station: metal plate with two piezoelectric patches

The measurement station used for tests is presented on figure 9. It contains two PI-876 piezoelectric patch transducers from PI Ceramic [14]. Vibrations are generated by an attached electric vibrator which is connected with the signal amplifier and function generator. Acceleration is measured by the external accelerometer connected directly to the plate next to the piezoelectric element (accelerometer sensitivity 10,43 mV/g).

B. Validation of sensing circuit

Figure 10 presents the validation of the sensing circuit. The WSN node is supplied; the transistor-switch is open (sensing capability). The first channel (CH1) presents the voltage over the ADC (with 1.65 V offset). It corresponds to simulations (figure 7a)

C. Validation of the switching circuit

Figure 11 presents waveforms of voltage and current while the transistor-switch is open. The value of the current is measured over the small series resistance added to the switching circuit. The CH1 channel presents the piezoelectric voltage, while the CH2 channel presents the current in the switching circuit. The values obtained from measurements are in keeping with its simulated counterparts. (figure 7b)

Figure 12 presents waveforms while the transistor-switch is closed. In this case the CH1 channel presents the acceleration waveform; while CH2 presents the current in the circuit. Figure 12 also presents the phase shift between voltage and current waveform (the capacitance character of the piezoelectric element). Obtained measured values correspond to simulations (figure 7c)

VIII. Discussion, perspectives

The sensing circuit and switching circuit have been designed separately to provide the best solution for each one. Both circuits have been simulated and implemented in the prototype

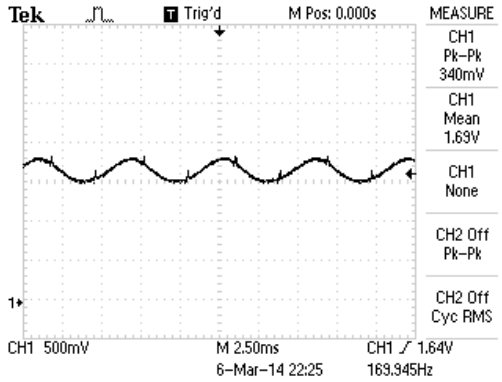


Fig. 10. Validation of the sensing circuit

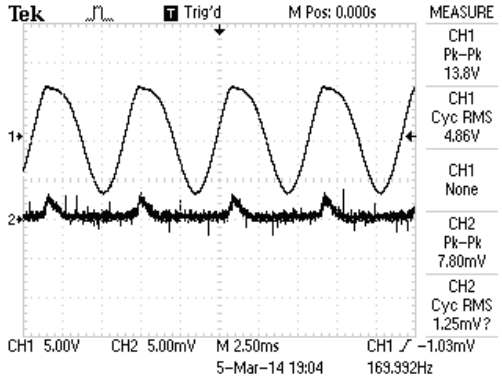


Fig. 11. Transistor-switch open

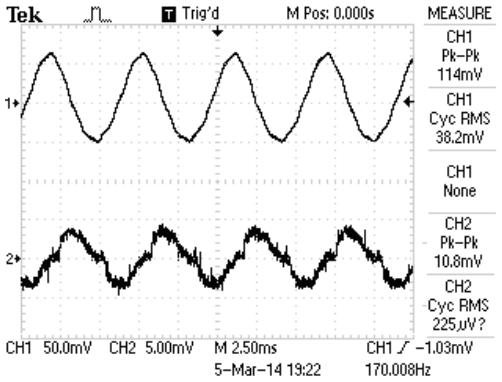


Fig. 12. Transistor-switch closed

design. Achieved measurements correspond to the simulations. According to figures 7b, 11, 7c and 12 the relative error between simulations and real measurements is 9.2 %.

The CH2 trace on figure 11 presents current in the switching circuit while the switch is open. It is default current caused by lack of the insulation between ground referenced switch, floating-source piezoelectric generator and differential-mode sensing circuit.

To provide galvanic separation (switch circuit not referenced to the ground) two solutions are considered: optical isolation [16], [18] and transformer [17], [18].

The optical isolation is recommended for its reliability and simple implementation. The disadvantage of this method is requirement of the second separated supply circuit to drive the MOSFET switch. It would add additional devices in the supply circuit. Since the WSN node is designed in order to verify the self-supply capability, the use of the additional devices which introduce energy losses should be avoided.

The second approach is to use the transformer to drive NMOS transistors. A pulse transformer is, in principle, a simple, reliable and highly noise-immune method of providing isolated gate drive. It can be advised for applications were the duty cycle is small. As a stand-alone component it can be used for duty cycles between 35% and 65% [19].

The series SSHI is a system with low duty cycle because the period of the mechanical displacement is much longer than period of the electrical circuit. It makes the pulse transformer a promising solution for our design.

Simulations have been realised to validate the circuit with the chosen solution (pulse transformer). The NI Multisim Component Evaluator 13.0 from National Instruments is used. The sensing vibrations circuit and the series SSHI circuit are in parallel with the piezoelectric element (design contains a switch not referenced to the ground). The entire simulation schema is presented on figure 13. The individual parts of the design are marked on the figure. We distinguish the switch driver composed of the voltage source V1 (corresponds to the microcontroller output signal in the real design). The series SSHI circuit contains two transistors and the inductor L2. Additionally, the bridge circuit (diodes: D1, D2, D3 and D4) is used to transform voltage from AC to DC. Finally, the load is simulated by the resistor R1 and the capacitor C1. The sensing circuit is connected in parallel with the SSHI circuit. Figure 14 compares the theoretical waveforms with the simulation results of the series SSHI method. Simulation results correspond to the theoretical waveforms but RLC and diode impacts are visible. The results justify the correctness of the integration of the sensing circuit with the energy harvesting circuit.

IX. Conclusion

Big centralised and wired systems for active vibration control are costly and use a large quantity of energy. A distributed solution based on an energy aware wireless sensor network has been proposed as a replacement for the centralised system. The autonomous WSN node needs to be designed to provide efficient wireless network for distributed active vibration control.

In this paper the global approach and the system assumptions are established and used as input data for the design. The proposed design of the WSN node is in accord with the prescribed requirements.

Designed node provides: vibration sensing, shunting the piezoelectric element and wireless communication. Furthermore, the series SSHI technique, chosen for the design, provides damping of the mechanical vibrations and the energy harvesting capability.

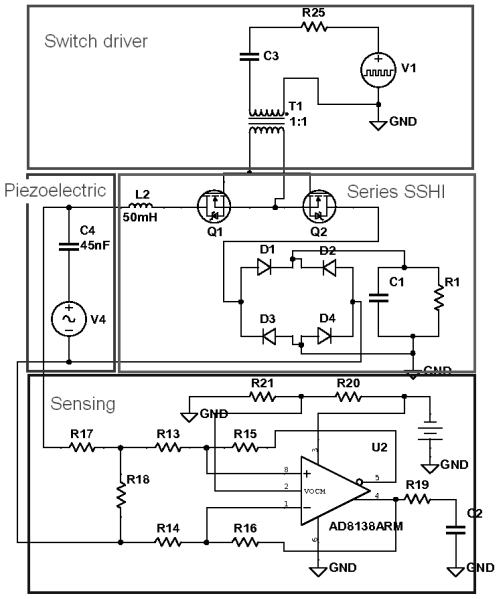


Fig. 13. Simulated circuit

Designed circuits for sensing vibrations and shunting the piezoelectric element are presented and described in details. The WSN node is modelled using the SPICE models. Achieved simulation results are consistent with the expected ones and validate the design.

The WSN node prototype has been constructed. The simulation results are compared with the measurements. The metal plate with the attached piezoelectric element is used as a measurement station. The measurement results correspond to the simulation results. The correctness of the simulations is proved with the measurements (relative error is 9.2 %).

Finally the separation problem for self-powered circuits has been indicated. Possible solutions are taken under consideration. The circuit with the chosen solution (pulse transformer) is simulated; obtained results correspond to the theoretical research (expected results).

The next step of the work would be modification of the prototype device according to the presented simulations. The galvanic separation for non-ground referenced switch will be used. The measurements would be compared with the simulation results. The following step would be validation of the energy harvesting and mechanical damping capability.

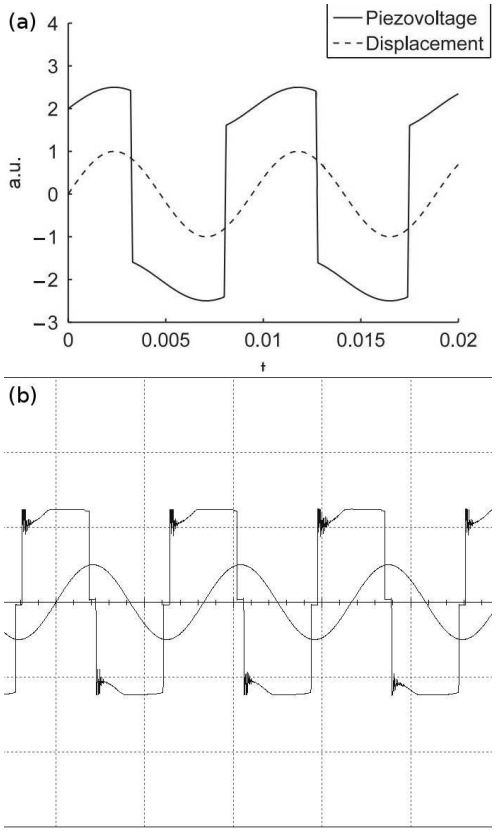


Fig. 14. (a) theoretical SSHI waveforms [10] (b) simulated SSHI waveforms

REFERENCES

[1] J. Yick, B. Mukherjee, D. Ghosal. August 2008. "Wireless sensor network survey" Computer Networks vol. 52 p.2292-2330, http://dx.doi.org/10.1016/j.comnet.2008.04.002

[2] J. Micek, J. Kapitulik. September 2012. "WSN sensor node for protected area monitoring" Federated Conference on Computer Science and Information Systems p.803-807, https://fedcsis.org/proceedings/2012/pliks/237.pdf

[3] J. P. Lynch. 2007. "An overview of wireless structural health monitoring for civil structures" Philosophical Transactions, Royal Society A, p. 345-372, http://dx.doi.org/10.1098/rsta.2006.1932

[4] F. Svaricek, T. Fueger, H. Karkosh, P. Marienfeld, C. Bohn. September 2010, Sciyo, Croatia. "Automotive Applications of Active Control", ISBN 978-953-307-117-6, pp. 380, http://cdn.intechopen.com/pdfs-wm/11899.pdf

[5] F. Mieyeville, M. Ichchou, G. Scorletti, D. Navarro, W. Du. 2012. "Wireless Sensor networks for active vibration control in automobile structures". Smart Mater. Struct 21, http://dx.doi.org/10.1088/0964-1726/21/7/075009

[6] S.J. Elliott. December 2008. "A review of Active Noise and Vibration Control in road vehicles", ISVR Tehcnical Memorandum No 981, http://eprints.soton.ac.uk/id/eprint/65371

[7] C.R. Fuller, A.H. Von Flotow. December 1995. "Active Control of Sound and Vibration" IEEE Control System, http://dx.doi.org/10.1109/37.476383

[8] L. Hermans and H. Van Der Auweraer. 1999. "Modal testing and analysis of structures under operational conditions: industrial applications" Mechanical Systems and Signal Processing 13(2), 193-216, http://dx.doi.org/10.1006/mssp.1998.1211

[9] S. H. Kim, J. M. Lee, M. H.Sung. 1999. "Structural-Acustic Modal Coupling Analysis and Application to Noise Reduction in a Vehicle Passenger Compartment". Journal of Sound and Vibration 255(5), p. 989-999, http://dx.doi.org/10.1006/jsvi.1999.2217

[10] E. Lefeuvre, A. Badel, C. Richard, L. Petit, D. Guyomar. 2006. "A comparison between several vibration-powered piezoelectric generators for standalone systems" Elsevier, Sensors and Actuators A 126 p. 405-416, http://dx.doi.org/10.1016/j.sna.2005.10.043

[11] Jin-Young Jeon. 2009. "Passive vibratin damping enhancement of piezoelectric shunt damping system using optimization approach", Journal of Mechanical Science and Technology, 23, http://dx.doi.org/10.1007/s12206-009-0402-8

[12] B de Marneffe, A Preumont. 2008. "Vibration damping with negative capacitance shunts: theory and experiment", Smart Materials and Structures 17, http://dx.doi.org/10.1088/0964-1726/17/3/035015

[13] Saber Mohammadi, Akram Khodayari. 2012. "Damping analyses of

structural vibratins and shunted piezoelectric transducers", Smart Materials Research, http://dx.doi.org/10.1155/2012/431790

[14] NI Multisim Component Evaluator White Paper. May 2012. Internet: http://www.ni.com/white-paper/9452/en/

[15] PI Piezo Technology, P-876 DuraAct Patch Transducer White Paper. 2014. Internet: http://piceramic.com/product-detail-page/p-876-101790.html

[16] E. L. Worthington, M. Zhu, P. Kirby. 2010. "Piezoelectric Energy Harvesting: Enhancing Power Output by Device Optimisation and Circuit Techniques". PhD Thesis, Cranfield University, https://dspace.lib.cranfield.ac.uk/

[17] L. Balogh. "Design And Application Guide for High Speed MOSFET Gate Drive Circuits", Texas Instruments, http://www.ti.com/lit/ml/slup169/slup169.pdf

[18] Fairchild Semicunductor. 2007. "Application Note AN-6069; Application review and comparative evaluation of low-side gate drivers", http://www.fairchildsemi.com/an/AN/AN-6069.pdf

[19] Vishay Siliconix. 2010. "Application Note AN-937;Gate Drive Characteristics and Requirements for HEXFET Power MOSFETs". Document Number: 91421, http://www.vishay.com/docs/91421/appnote9.pdf

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1073–1078
DOI: 10.15439/2014F436
ACSIS, Vol. 2

Carrier sense range effect on multipath routing performances in Wireless Sensor Networks

I. Bennis*‡, H. Fouchal‡, O. Zytoune*§, D. Aboutajdine*
*LRIT, unité associée au CNRST (URAC29), Université Mohammed V - Agdal, Rabat, Marocco
‡Université de Reims Champagne-Ardenne, France
§Université Ibn Tofail, Kénitra, Maroc
Email: i.bennis@fsr.um5a.ma, hacene.fouchal@univ-reims.fr, zytoune@univ-ibntofail.ac.ma, aboutaj@fsr.ac.ma

***Abstract*—In the recent years, Wireless Sensor Networks (WSNs) have become a great field of interest for scientific community. This kind of network provides a panoply of applications in different areas of human life. However WSNs must ensure the quality of service (QoS) to give the requested performance for the final user. Among different issues presented in the literature to provide high QoS, multipath routing is commonly used. But such a solution could be not enough efficient if the multipath routing design does not consider the phenomena of interference. Indeed the constructed paths can have an interference zone, mainly a shared carrier sense range. In this paper we show by analytical and experimental results, that using multipath routing can never overshoot the performance of single path when the interferences are not taken into account. Also, we show how the carrier sense range can influence the network performance.**

***Index Terms*—WSN, Carrier sense range, Interference, Multipath routing, QoS.**

I. Introduction

NOWADAYS Wireless Sensor Networks (WSNs) attract more and more researchers as being an interdisciplinary research of interest. This kind of networks offers countless applications like precise agriculture, system monitoring and many others. One of the most challenging issue in these networks is that the routing techniques must satisfy some QoS metrics. Especially in the case of multimedia applications that require the best performance in terms of energy consumption, delay, jitter, reliability and bandwidth.

There are many solutions that can be included in the protocol design to ensure QoS requirements as using multipath routing. This technique consists of giving a source node the possibility to use any of several paths to reach a particular destination at any time. According to [1], [2], the use of a multipath routing has many benefits such as aggregation of bandwidth by splitting data to the same destination into multiple streams. Also, the use of the multipath principle can reduce the end to end delay in case of route failure because there is no need to restart a new path discovery process. In addition, multipath routing has the ability to improve the reliability of the transmitted information by sending multiple copies of the same data on multiple paths, which increases the accuracy. Another interesting benefit of multipath techniques is the load balancing which allows a better use of available network resources in order to reduce traffic congestion.

To enhance the multipath routing scheme design, many works [3], [4] use the notion of disjoint paths. This notion reflects the independence of paths in terms of shared resources. The higher is the degree of independence, the more the multipath strategy promotes an adaptive use of the network resources. With such a feature, the multipath solution can avoid congestion by balancing the load among the multiple disjoint paths. In the literature, several techniques based on the degree of disjoint paths are used to classify the set of paths between a source node and a destination node. Among them we cite Link-disjoint, Node-disjoint, Maximally-disjoint and Radio-disjoint Multipath [1].

However multipath solution hides a serious drawback. In fact since a single channel is used in the wireless network, the sensors nodes share the medium of communication. And due to the broadcast nature of radio communication the level of interference is more pronounced [5], [6]. Also, when several paths are used simultaneously, even if the node disjoint priority is satisfied, it remains a significant risk of collisions that results in high packet loss rate. So the concurrent use of multiple paths constructed from source to destination results in intensive inter-path interference [7]. Therefore performances will decrease seriously. Authors in [8] concluded that transmitting data over multiple paths is not a synonym of result improvement unless the effects of the wireless communications are taken into account. This phenomena is also known as the route-coupling problem [9]. It occurs when simultaneous communications through multiple paths are ongoing, and these paths are located physically close enough in order to interfere with each other.

Another aspect that must be considered, is the effect of the carrier sense range on communication performances in sensor networks, mainly in routing protocol. Indeed, while the carrier sense range is usually more larger than the transmission range, there is more chance that interferences occur in this range, especially in the case of high density. We mean by carrier sense range effect the fact that a node cannot transmit as an other node in its carrier sense range is already in a transmitting phase.

In this work we study how the carrier sense range can influence the communication in a sensor network. We show by analytical and experimental results that using multipath routing can never overshoot the performances of a single path when

978-83-60810-58-3/$25.00

the carrier sense range effect is not considered in the routing design.

The remainder of this paper is organized as follows: in section II, some routing protocols with the aim of reducing the interference problem in the multipath case are presented. In section III, we describe by analytical model the effect of both wireless interferences and carrier sensing. Simulation results are shown in section IV. Finally in section V, we draw the conclusion and give some perspectives.

II. Related work

In the literature, there exists some solutions that aim to reduce the effects of interference such as directed antennas [10]. The authors try to find a zone-disjoint multipath to avoid collisions between paths. Another solution consists of using multi-channel transmission [11]. But these both methods cannot be easily used due to the resource-constrained propriety of the WSNs. An alternative technique is to calculate the degree of independence between a set of paths using the correlation factor or the coupling metrics. The correlation factor between two node-disjoint paths is defined as the total number of shared links of the paths [5]. It represents the chances that the transmission along the different paths could interfere with each other in a shared channel. The coupling between two paths (Pl and P2) is defined as the average number of nodes that are unable to receive data along P2 when a single node in P1 is transmitting [9]. The more the path has lower correlation factor or coupling effect, the more suitable is for multipath construction allowing better performances.

In the following, we will review some studies about routing protocols that aim reducing the interference problem in the multipath case.

- Energy Efficient Collision Aware Multipath Routing for WSN (EECA) [12]:
  The EECA is an on-demand routing protocol that constructs multiple paths using request/reply cycles. This protocol has two aims:
  - Reducing the flooding of route request messages by restricting it to the neighbors of nodes iteratively added to the route being discovered.
  - Saving energy by adjusting power needed to transmit the data and the control messages and so reducing the potential collision area of each node.

  The author makes assumption that each node can adjust the radio transmit power to vary its communication range from 0 to a specific transmit range. The EECA algorithm attempts to find two collision-free routes using the node position information. The source starts by checking if there are two groups in its neighbor list satisfying the following three conditions: 1) all these nodes are close to the destination; 2) The nodes of the two groups are opposite and separated by the source-destination line; 3) each node is distanced more than R/2 from the source destination line.
  However such restrictions limit the chance to find two paths far away from each other. So many nodes for the first constructed path remain in the carrier sense range of other nodes for the second path.
- Interference-Minimized Multipath Routing with Congestion Control in Wireless Sensor Network for High-Rate Streaming (I2MR) [13]:
  This protocol tries to increase the throughput by discovering zone-disjoint paths and adopting the load balancing scheme, while requiring minimal geographic information to reduce overheads. Localization support is only required at the source nodes, which are the most powerful sensor nodes equipped with Electro-Optic devices. The basic idea is to mark-out the interference zone of the nodes of the first path after it has been discovered. Then subsequent paths cannot be discovered within this interference zone. I2MR tries to construct zone-disjoint paths and distributes network traffic over the discovered paths by assuming a special network structure and the availability of particular hardware components. In I2MR, the source node tries to find three paths, but uses the two first paths for data transmission and keeps the third one as a backup path. However, this work needs a special network structure and particular hardware components making this protocol not applicable to all types of sensors. In addition, due to the high complexity of the introduced zone-marking mechanism, and the different type of packet control, the generated overhead is more pronounced.
- Maximally radio-disjoint multipath routing for wireless multimedia sensor networks (MR2) [14]
  The main objective for the MR2 protocol is to provide the required bandwidth for multimedia applications through non-interfering paths. MR2 utilizes an adaptive incremental technique to construct minimum-interfering paths. To do so, only one path is built for a given session. Additional paths are built when required, typically in case of congestion or bandwidth shortage. Interference awareness and energy saving are achieved by switching a subset of sensor nodes in a passive state in which they do not take part in the routing process. The passive state is represented by switching the sensor node to a sleep or an idle mode. Thus enabling increasing the network lifetime. However, MR2 is only suitable for query-driven applications. Also, the utilized flooding strategy for constructing non-interfering paths implies a high control overhead. And as the bepassive message is received only by the neighbors of each node forming the path, the constructed paths are spaced by a distance approximately equal to the transmission range. So the carrier sense range effect is not considered.

So for all discussed works the constructed paths are spaced by a distance equal to the transmission range or at most the interference range. Thus the carrier sense range has never been treated.

III. Analytical model

In this section, we will give a model for a wireless sensor network as a connectivity graph. After that by using the

protocol model of interference and the physical model of interference we will describe the effects of wireless interferences and of carrier sensing. And finally by a conflict graph we will present the different relations between the wireless links in the network.

A. Connectivity graph

We consider a wireless network with N nodes randomly located on a plane space. We denote our network by $G(V, E)$ where V is the vertex that represents a set of N nodes and E is the edge that represents set of directed links connecting the nodes in V. Let n_k where $1 \leq k \leq N$ denote the nodes in V, and l_{ij} and d_{ij} denote respectively the directed link and the distance between nodes n_i and n_j with $i, j \in [1, N]$. The notation l_{ij} means also that the node n_i sends packet to node n_j, so l_{ij} has not the same meaning than l_{ji} . Each node, $n_k \in V$, is equipped with a radio having three levels of range centred at the node n_k:

- The transmission range $R_{Tr}(k)$, is the range where a successful communication can be achieved.
- The interference range $R_I(k)$, is the range where every node that attempts to start a communication will cause collision at node n_k when it receives packet.
- The carrier sense range $R_{CS}(k)$, is the range where every node that attempts to start communication will prevent node n_k to transmit.

The relationship between the three ranges is $R_{Tr}(i) < R_I(i) < R_{CS}(i)$. We notice that the nodes are homogeneous, so each node has the same ranges than others. In our study we consider the case of a single wireless channel.

According to the protocol model of interference [15], there is a successful transmission between nodes n_i and n_j if the following conditions are satisfied:

- $d_{ij} \leq R_{Tr}(i)$.
- no transmitting nodes in the potential zone of interference of node n_j.
- no nodes in the $R_{CS}(i)$ that sends packet concurrently with the node n_i.

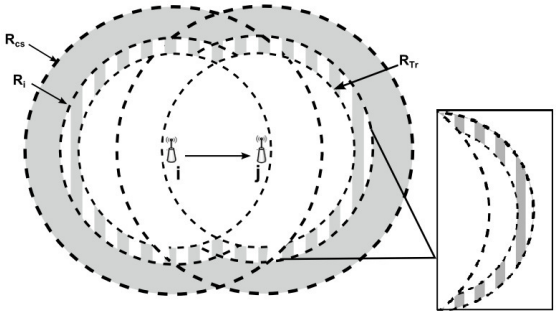


Fig. 1: Potential zone of interference

The potential zone of interference is shown in Fig.1. It is delimited by the intersection of carrier sense range of transmitter (node n_i) and the interference range and transmission range of the receiver (node n_j). The area of this zone depends on the distance between the transmitter and the receiver.

If we consider the physical model of interferences [15] a successful transmission arises only if the signal strength of the received frame at node n_j is stronger than RxThresh_. Otherwise if signal strength is less than RxThresh_ but greater than CSThresh_, the receiver will not be able to decode correctly the signal and the channel is considered as busy. In the case when multi-frames are received simultaneously by node n_j, it calculates the ratio of the strongest frame signal strength to the signal strength sum of other frames. If it is larger than CPThresh_, the frame will be received correctly and other frames are ignored. Otherwise, all frames are discarded. This case represents the interference that can be occurred at the receiver. The three values RxThresh_, CSThresh_ and CPThresh_ are specific thresholds for the wireless node. NS-2 [16] uses the above description to simulate the reception of signal.

B. Conflict graph

In order to show which wireless links interfere with each other in the network we consider a conflict graph $G'(V', E')$ where V' is the vertex that represents each link in the connectivity graph G, and E' is the edge which represents the set of all possible relations between the vertices in V'.

Based on the protocol interference model described above, an edge can be drawn between two vertices l_{ij} and l_{pq} in G' if the links l_{ij} and l_{pq} may be active simultaneously. Such a condition is achieved if one of the two following conditions are true:

- $i \in R_I(q)$ or $p \in R_I(j)$.
- $i \in R_{CS}(p)$ or $p \in R_{CS}(i)$.

C. Multipath case

Our aim in this subsection is to show how the multipath routing reacts under the three ranges defined above. Namely how each node in each path will interact with other nodes in the other paths. Let have two paths from the source *S* and the sink *(a-b-c-d)* and *(e-f-g)* as mentioned in Fig.2.
In the first path the node *b* can only forward its data if the nodes *a* and *c* are not in the transmitting phase. The same observation can be expected for the node *c* and the nodes *b* and *d*. We notice that the transmission is done by a broadcast way due to the nature of radio device.
At the second path, if the node *g* in $R_I(c)$ starts transmission to the sink (if *d* and *c* do not transmit), then the receiving data at the node *c* from the node *b* will be disrupted. Also, if the node *f* starts forwarding, it will be in a competition with nodes *b* and *c* at the same time as it is located in the carrier sense range of the two nodes; which will cause delays and packet losses. If the node *a* sends packets to node *b*, this one will be able to successfully receive the packets without any potential interference caused by the node *f*.

IV. Simulation and Experimental Results

The aim of this simulation is to show how the same routing protocol performs in single path case and in the multipath

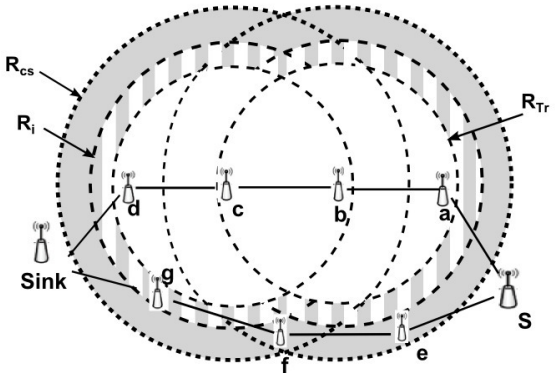


Fig. 2: Multipath study under the three ranges

case without considering the multipath effect. Also, we investigate how the carrier sense range can influence the routing results. We choose as routing protocol for this study our implementation [17] of the Two Phase geographical Greedy Forwarding (TPGF) as it has been one of the first protocol which introduces the concept of multi-paths in the field of WMSNs [18]. This algorithm focuses on the exploration and the establishment of a maximum number of best disjoint routes in terms of end-to-end delay. We evaluate the performance of the studied protocol under the delay metrics to measure the average end-to-end delay of successful received packets. And under the Packet Delivery Ratio (PDR) metrics to calculate the ratio between the number of correctly received packets at the destination and the number of packets sent by the source.

Working environment

Here we will describe our studied scenario. Each simulation scenario is presented as follows: **X** nodes are randomly located in an area of 1500*1500 m^2, where:

$$X=[\![100;150;200;250;300\ ]\!]$$

Data traffic is generated by a randomly source in the network to a sink. This one is located in the center of our experimentation area and has the last ID. The source node generates a constant bit rate sources (CBR) traffic with **Y** packets per second, where:

$$Y=[\![16;32;64;128;256\ ]\!]$$

The data packet size is 1000 bytes. The duration of communication is 40 seconds, an no mobility is supported in this scenario. For every value of **X** and **Y** 50 scenarios are generated and the average value of the results are calculated. Due to the lack of space we represent only the results in case of 64 packets per second.

Table I summarizes the parameters used for simulation.

Result analysis

From Fig. 3, we notice that for both multipath and single path cases, the average delay decreases as the network density grows, it is quite normal as we have only one source for each

TABLE I: Main configuration parametres

Parametres	value
link layer	LL
MAC layer	IEEE 802.11
radio propagation	two ray ground
interface queue	PriQueue
ifqlen	50
antenna	omni-antenna
$CPThreshold$ (Watt)	10
$CSThreshold$ (Watt)	$1.559 * 10^{-11}$
$RXThreshold$ (Watt)	$3.652 * 10^{-11}$
Pt (Watt)	0.2818

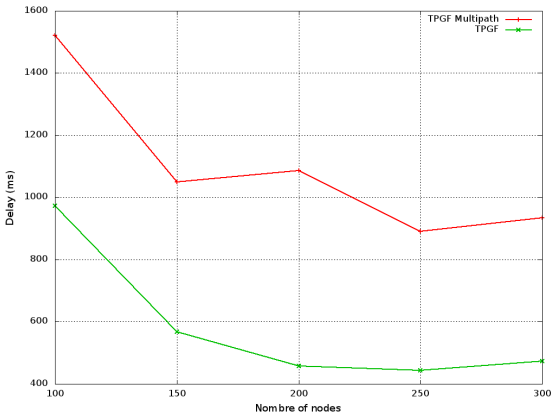


Fig. 3: Average delay vs network size

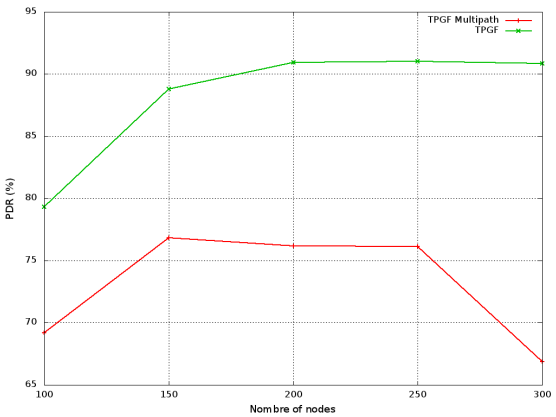


Fig. 4: Average PDR vs network size

scenario. Also, the main observation that Fig. 3 shows is the fact that the single path has the best average delay against the multipath case. The gain is approximately equal to 70%.

This result is expectable since the TPGF protocol does not take into account the interference problem. For Fig. 4 we can see also that the single path has the best average PDR against the multipath case. The gain is approximately equal to 20%.

This part of results confirms that using multipath routing

without taking into consideration interference and the carriers sense range effects makes the multipath solution less profitable than a single path one.

In order to show how the effect of the carrier sense range can influence performances of routing protocols, we discuss the following scenario. First of all, we modify the TPGF protocol in order to have two version. The first one represents the case where the different paths constructed from the source to the sink have at least a distance from each other equal to the transmission range. We denote this version by *RX-avoid*. The second version represents the case where the paths constructed from the source to the sink have at least a distance from each other equal to carrier sense range. We denote this version by *CS-avoid*. In this simulations, we have used a grid with 100 nodes spaced by a distance equal to 200 m. The area of the grid is 4000*2500 m^2. The source node is selected randomly and generatesa constant bit rate (CBR) traffic with **Y** packets per second, where:

$$Y=[\![16;32;64;80;128;256\ ]\!]$$

In this simulation we simply use two paths. We notice also that if there is no way to find the second path with the desired condition, we use only the first path. We repeat the simulations several times and we measure the average of delay sand PDRs.

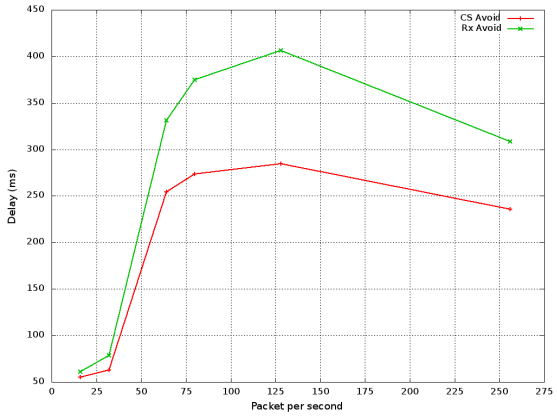


Fig. 5: Average delay vs Rate

From Fig. 5, we notice that for both *CS-avoid* and *RX-avoid* case the average delay increases as the rate grows espacialy from 16 to 128 packet per second. It is quite expected as more the rate is higher more the queue of intermediate nodes is filled. So each packet takes more time to reach destination. But the main observation that we can make from Fig. 5 is the fact that the *CS-avoid* case has the best average delay against the *RX-avoid* case.

As discussed in section III, the carrier sense range effect occurs when a node cannot transmit as an other node in its carrier sense range is already in transmitting phase. This is exactly the case here, in fact more the number of packets per second increases, more the need for channel access is higher. Therefore the nodes of the two paths deprive mutually the channel access since there is a competition between them. For

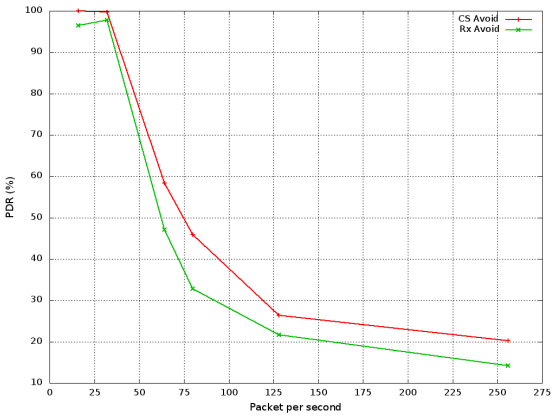


Fig. 6: Average PDR vs Rate

Fig. 6, we can see also that the multipath with the *CS-avoid* has the best average PDR against the *RX-avoid* case.

V. Conclusion and Future Work

Using the multipath routing in the WSN has many benefits such as aggregation of bandwidth, reducing the end to end delay, improving reliability. However, using the multipath routing scheme without considering the effect of the carrier sense range decreases the network performances instead of enhancing it.

We have presented in this paper how the carrier sense range can influence the performance of communication in wireless sensor networks. We show by analytical and experimental results that using multipath techniques can never overshoot the performances of a single path solution unless the carrier sense range effect is considered in the routing design.

We have performed a large number of simulations. Their results prove that multipath with the *CS-avoid* outperforms the multipath with *RX-avoid* for both metrics: delay and PDR. As a future work ,we intend to achieve much more simulations for large scale situations and will consider mobility feature. In addition, we need to experiment this protocol with realistic streams which model multimedia ones.

Acknowledgment

This work is partially supported by ”Projet de coopération Maroco-Française : Contribution à l’optimisation de la qualité de service dans les réseaux de capteurs sans fil : Application à la supervision agricole 2013-2014”.

This work is partially supported by the ”Projet SCOOP@F : Système de Coopération Pilote” funded by the EC 2014-2015.

References

[1] V. D. Mytri Jayashree A, G. S. Biradar. Review of multipath routing protocols in wireless multimedia sensor network -a survey. *International Journal of Scientific and Engineering Research*, 3(9):174–182, 2012.

[2] H. Zafar, D. Harle, I. Andonovic, and Y. Khawaja. Performance evaluation of shortest multipath source routing scheme. *Communications, IET*, 3(5):700–713, 2009.

[3] Zhongdong Wang and Qiushuang Wang. The research of a multiple disjoint paths routing protocol for ad hoc sensor networks. *Energy Procedia*, 17, Part A(0):499 – 505, 2012. International Conference on Future Electrical Power and Energy System.

[4] U.B. Mahadevaswamy and M.N. Shanmukhaswamy. Delay aware and load balanced multi-path routing in wireless sensor networks. *International Journal of Wireless Information Networks*, 19(3):278–285, 2012.

[5] Kui Wu and J. Harms. Performance study of a multipath routing method for wireless mobile ad hoc networks. In *Modeling, Analysis and Simulation of Computer and Telecommunication Systems, 2001. Proceedings. Ninth International Symposium on*, pages 99–107, 2001.

[6] Marjan Radi, Behnam Dezfouli, Kamalrulnizam Abu Bakar, Shukor Abd Razak, and Mohammad Ali Nematbakhsh. Interference-aware multipath routing protocol for qos improvement in event-driven wireless sensor networks. *Tsinghua Science & Technology*, 16(5):475 – 490, 2011.

[7] Marjan Radi, Behnam Dezfouli, Kamalrulnizam Abu Bakar, and Malrey Lee. Multipath routing in wireless sensor networks: Survey and research challenges. *Sensors*, 12(1):650–685, 2012.

[8] E.P.C. Jones, M. Karsten, and P.A.S. Ward. Multipath load balancing in multi-hop wireless networks. In *Wireless And Mobile Computing, Networking And Communications, 2005. (WiMob'2005), IEEE International Conference on*, volume 2, pages 158–166 Vol. 2, Aug 2005.

[9] M.R. Pearlman, Z.J. Haas, P. Sholander, and S.S. Tabrizi. On the impact of alternate path routing for load balancing in mobile ad hoc networks. In *Mobile and Ad Hoc Networking and Computing, 2000. MobiHOC. 2000 First Annual Workshop on*, pages 3–10, 2000.

[10] Dola Saha, Siuli Roy, Somprakash Bandyopadhyay, Somprakash B, Tetsuro Ueda, and Shinsuke Tanaka. An adaptive framework for multipath routing via maximally zone-disjoint shortest paths in ad hoc wireless networks with directional antenna. In *In IEEE Global Telecommunications Conference*, pages 226–230, 2003.

[11] Wai-Hong Tarn and Yu-Chee Tseng. Joint multi-channel link layer and multi-path routing design for wireless mesh networks. In *INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE*, pages 2081–2089, May 2007.

[12] Zijian Wang, E. Bulut, and B.K. Szymanski. Energy efficient collision aware multipath routing for wireless sensor networks. In *Communications, 2009. ICC '09. IEEE International Conference on*, pages 1–5, June 2009.

[13] Jenn-Yue Teo, Yajun Ha, and Chen-Khong Tham. Interference-minimized multipath routing with congestion control in wireless sensor network for high-rate streaming. *Mobile Computing, IEEE Transactions on*, 7(9):1124–1137, Sept 2008.

[14] Moufida Maimour. Maximally radio-disjoint multipath routing for wireless multimedia sensor networks. In *Proceedings of the 4th ACM Workshop on Wireless Multimedia Networking and Performance Modeling*, WMuNeP '08, pages 26–31, New York, NY, USA, 2008. ACM.

[15] Kamal Jain, Jitendra Padhye, VenkataN. Padmanabhan, and Lili Qiu. Impact of interference on multi-hop wireless network performance. *Wireless Networks*, 11(4):471–487, 2005.

[16] VINT. The network simulator ns-2.34, 2012.

[17] I Bennis, H. Fouchal, O. Zytoune, and D. Aboutajdine. An evaluation of the tpgf protocol implementation over ns-2. In *Communications (ICC), 2014 IEEE International Conference on*, pages 428–433, June 2014.

[18] Lei Shu, Yan Zhang, LaurenceT. Yang, Yu Wang, Manfred Hauswirth, and Naixue Xiong. Tpgf: geographic routing in wireless multimedia sensor networks. *Telecommunication Systems*, 44(1-2):79–95, 2010.

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 1079–1086

DOI: 10.15439/2014F82
ACSIS, Vol. 2

Switched-Beam Antenna for WSN Nodes Enabling Hardware-driven Power Saving

Luca Catarinucci, Sergio Guglielmi, Riccardo Colella, Luciano Tarricone
Department of Innovation Engineering, University of Salento
via per Monteroni, 73100, Lecce, Italy
Email: {luca.catarinucci, sergio.guglielmi, riccardo.colella, luciano.tarricone}@unisalento.it

***Abstract*—Energy saving is one of the most important issues in Wireless Sensor Network (WSN) context. Since the communication task is the most power-consuming operation, it is quite important to achieve an energy efficient communication in order to increase the lifetime of the devices through an intelligent use of the power transmission. In this context, the integration of WSN nodes with switched-beam antennas is becoming more and more appealing due to the possibility to extend sensor node lifetime by optimizing the transmitted power.**

In this work a switched-beam antenna for WSNs nodes in the ISM band (2.4-2.4835 GHz) is proposed. The radiating structure consists of four identical antennas, composed of an array of two L-shaped quarter-wavelength slot antenna elements arranged in a compact and symmetrical planar structure. Thanks to a properly designed switching circuit which controls the feeding of the antenna elements, one among eight possible different radiation patterns in the azimuth plane can be selected on the basis of specific needs. Simulations and experimental results, referred to a prototype realized on a FR-4 substrate, demonstrate the appropriateness of the proposed switched-beam antenna system as hardware element enabling new power saving strategies in WSN contexts.

I. Introduction

GUARANTEEING adequate energy efficiency in Wireless Sensor Networks (WSN) contexts is still an open issue. Indeed, WSN nodes, hereafter referred as motes, are frequently placed in unpractical or hardly accessible places and it could be difficult, time-consuming and expensive to replace batteries. Therefore, both WSN devices and communication protocols must be carefully designed to maximize the motes lifetime, in order to reduce maintenance costs and outages. In such a context, it is well known that the data communication through the RF frontend represents the most power consuming task [1-3]. Just to give an idea, the power required by the mote processor to process thousands of operations is comparable to that the RF transceiver needs to transmit a single bit. Therefore, in the last years, many techniques aimed at minimizing the energy consumption have been proposed [4-6]. Such techniques are often based on specific protocols optimizing the data transmission, for instance by periodically turning motes into sleep mode or opportunely controlling the radio transceiver activation/deactivation.

In order to further increase the energy efficiency, thereby extending even more the motes lifetime, the use of opportunely controlled directional or switched-beam antennas could be a winning approach. Indeed, as already stated, motes are usually equipped with almost omnidirectional antennas and, consequently, only the power portion transmitted toward the proper mote is effectively used, whilst most of the power is wasted elsewhere. Vice versa, a mote provided with a switched-beam antenna could smartly reconfigure the antenna radiation pattern so to convey the power only toward the destination mote.

In this work a full planar and really compact switched-beam antenna particularly suitable for WSN applications in the ISM band is proposed. When connected to the wireless module of a mote, it works in place of the common omnidirectional antenna and its radiation properties can be controlled through a digital interface. The proposed solution consists of four identical radiating structures, each one composed of two L-shaped quarter-wavelength slot antenna elements, arranged in a compact and symmetrical planar structure. Eight switchable radiation patterns with remarkable gain can be obtained, covering symmetrically 360 degrees in the azimuth plane, in order to reduce the transmission power and consequently extend the lifetime of the mote.

The paper is structured as follows: in Section II the state of the art on switched beam antennas in WSN context is discussed, whilst in Section III both working principle and design of the proposed antenna are described; later on in Section IV simulated and experimental results are shown and discussed. Finally, conclusions are drawn in Section V.

II. Related Works

Over the last few years, a strong research effort has been dedicated to the design of more and more performing switched-beam antennas [7-14]. Nevertheless, their integration in WSN motes to reduce energy consumption and to extend motes lifetime, has not been exhaustively explored yet. In particular, a smart switched-beam directional antenna is proposed in [7]. It is composed of four planar patch antennas arranged in a box-like structure. It can switch among four radiation beams with a uniform coverage of the azimuth plane and a good radiation gain in the main lobe direction, but its very large size is not compatible with the integration in WSN nodes. A pattern reconfigurable antenna composed of microstrip parasitic array elements is proposed in [8]. It is

978-83-60810-58-3/$25.00

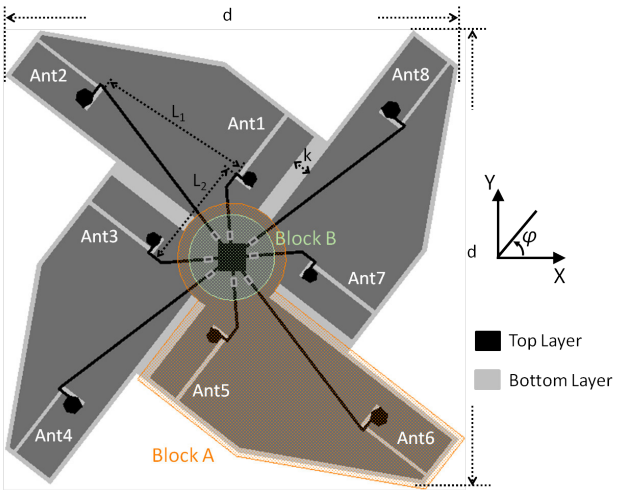


Fig 1. Structure of the proposed switched-beam antenna. The detailed structure of block A is shown in Fig. 2 and those of Block B in Fig. 3.

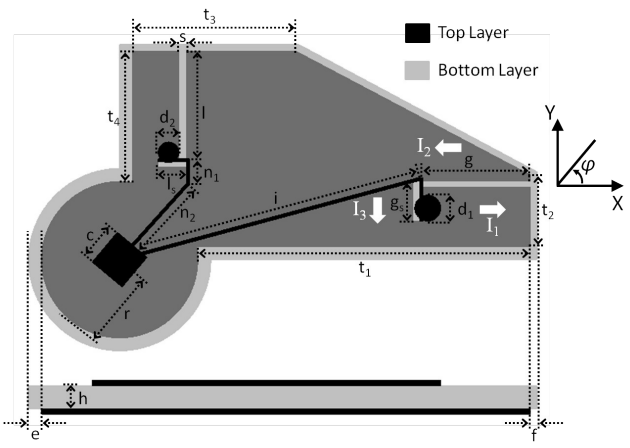


Fig 2. Geometry of the proposed array of two L-shaped quarter-wavelength slot antenna elements, corresponding to Block A in Fig. 1.

based on a simple and compact structure, but it does not ensure a 360-degree uniform coverage in the azimuth plane. Another interesting reconfigurable antenna for WSN sink nodes capable to switch between front-directional and conical beam pattern, is proposed in [9]. It is composed of an EBG ground plane and a power divider with a stub switched phase shifter. Despite the appreciable peak gain in the main lobe direction, it is rather cumbersome and presents no radiating beams in the azimuth plane. A reconfigurable angular diversity antenna using quad corners as reflector arrays and a switching control circuit is proposed in [10]. It presents a high radiation gain, but occupies a large volume and does not guarantee a uniform radiation patterns. On the contrary, a compact Switched-Beam Antenna composed of a four-element antenna array is presented in [11]. It shows eight switchable directional patterns and an omnidirectional one, thus ensuring an uniform coverage of the 360 degrees horizon. Moreover, it is both compact and inexpensive. Unfortunately, it exhibits an Half-Power Beam Width (HPBW) of a single beam of nearly 120 degrees, which causes a large overlapping area, thus not ensuring an optimized energy saving.

Finally, in our earlier work [12]-[14], a reconfigurable beam-steering antenna for WSN nodes is presented. It is composed of a vertical half-wavelength dipole antenna and eight microstrip antennas arranged in a 3D configuration. Thanks to a digital control circuit, it can switch among nine radiation patterns, one omnidirectional and eight directional with a HPBW of nearly 60 degrees in the azimuth plane. It guarantees a uniform coverage of the 360 degree horizon and a remarkable peak gain. The main issue is related to its integrability with real WSN nodes due to its 3D configuration.

In the following Sections, a new switched beam antenna controllable by and easily integrable into WSN nodes will be exhaustively presented along with simulated and measurement results.

III. Switched-Beam Antenna Design

The overall structure of the proposed switched-beam antenna is shown in Fig. 1. In order to explain the characteristics of the designed radiating structure, the geometry of the array of two L-shaped quarter-wavelength slot antenna elements is firstly introduced.

The geometry of the proposed array of two L-shaped quarter-wavelength slot antenna elements is shown in Fig. 2. As described in [11], [15]-[16], a conventional L-slot antenna is composed of two slots connected at their ends with an angle of 90 degrees, with a total slot length near to a quarter-wavelength at the reference frequency. The antenna is fed by a 50 Ω microstrip transmission line on the opposite side of the substrate (FR-4 in our case) with respect to the slot antenna. Such a feeding line configuration does not impact on weight and size of the entire system and it is suitable for the integration with electronic devices. As shown in Fig. 2, the total current on the ground plane can be decomposed in three parts; in particular, currents I_1 and I_2 have opposite directions and therefore their contribution vanishes, leaving only the contribution of current I_3. Therefore, the L-shaped slot antenna works as a small dipole oriented in the *y*-axis, with a bidirectional pattern in the *xy*-plane. Furthermore, due to the shape of the ground plane that has a larger area in the direction of the feed point, the antenna presents a directional pattern in *xy*-plane, with a main lobe direction oriented toward *x*-axis. Typically, the L-shaped slot antenna exhibits a gain in the maximum radiation direction similar to that of a dipole.

The proposed array consists of two L-shaped quarter-wavelength slot antenna elements, arranged at a distance of about a quarter of the wavelength at the reference frequency and perpendicular to each other. The resulting radiation pattern rotates of about 45 degrees from *x*-axis in the azimuth plane and the gain in the main lobe direction is about 5.09 dBi. The greater gain with respect to a dipole antenna gives the possibility to reduce the transmission power and hence extend the lifetime of a mote. The proposed array architec-

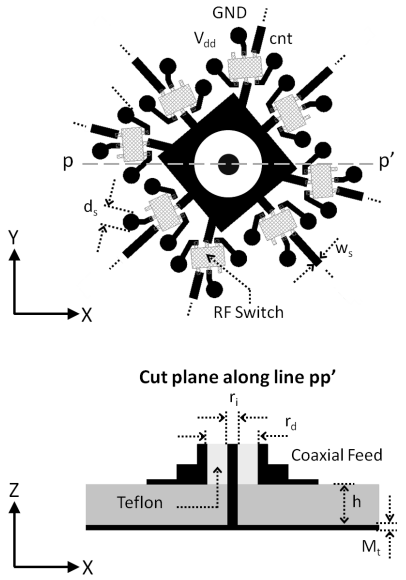


Fig 3. Detailed structure of the proposed switched-beam antenna, corresponding to Block B in Fig. 1.

ture is a good solution to increase energy efficiency or to reduce multipath effects, but it may deteriorate network performance if the signal comes from the directions of the nulls. For this reason, some arrays have been combined in order to have a uniform coverage of the 360 degrees in the azimuth plane.

The whole antenna structure is shown in Fig. 1 whilst the detailed parts, Block A and Block B, are depicted in Fig. 2 and 3 respectively. The antenna consists of four arrays of two L-shaped quarter-wavelength slot antenna elements arranged in a symmetrical planar structure. Basically, the array is replicated three times and rotated 90 degrees along the axis perpendicular to the *xy*-axis and passing through the center of the circumference with radius *r*. The radiating structure is fed by a 50 Ω SMA connector at the center of the antenna, so that the feeding point is sufficiently far from the radiating elements and does not interact with them. The eight microstrip transmission lines are interrupted by RF switches that control the feeding of the antenna elements. In Fig. 3, the structure of the feeding point and the digital circuit that controls the RF switches are shown. Powered through the lines V_{dd} and *GND*, the state of the RF switches is controlled by the signal *cnt*; when *cnt* is set to 3V, the microstrip line is connected and the relative antenna element radiates, otherwise when it is set to 0V the switch opens the line.

As described above, the array of two L-shaped quarter-wavelength slot antenna elements exhibits a directional radiation pattern with the main lobe direction oriented as the bisector of the angle identified by the two antenna elements. Thanks to the symmetrical replication of the array, the antenna can illuminate a specific direction that is $n\ \pi/2$ in the azimuth plane, where n = 0,...,3. Moreover, feeding two non-adjacent L-shaped elements (in particular Ant1-Ant3, Ant3-Ant5, Ant5-Ant-7 and Ant7-Ant1) other four patterns toward the directions $(1+2n)\ \pi/4$, where n = 0,...,3, can be obtained. Therefore, by carefully setting the *cnt* control line of the RF switches, the proposed antenna can switch among eight symmetrical radiation patterns spaced 45 degrees from each other.

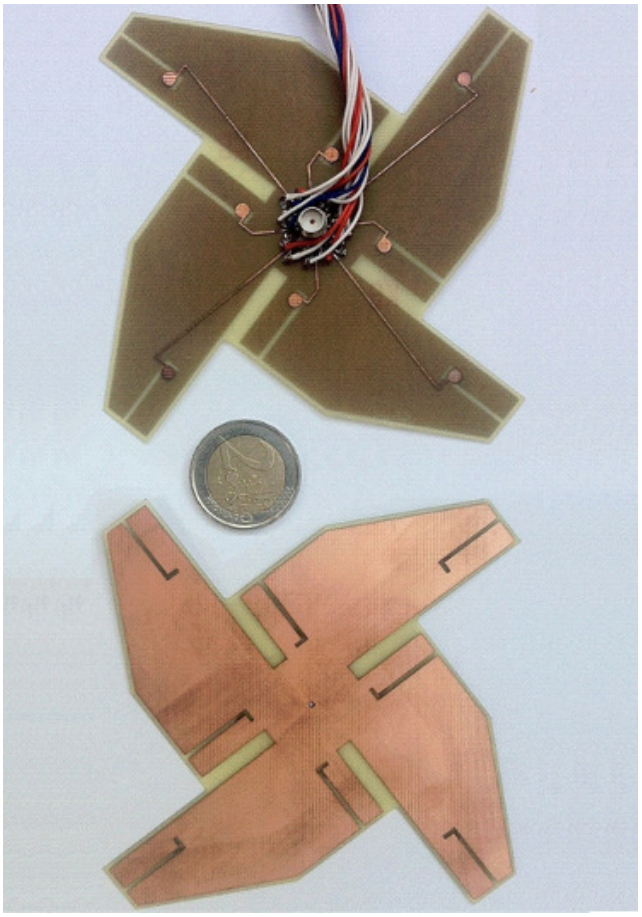


Fig 4. Proposed switched-beam antenna prototype (both sides).

IV. Simulated and Experimental Results

Several tests and measures have been performed in order to obtain an accurate characterization of the electromagnetic properties of the proposed antenna; for this purpose, a very cost-effective prototype of the switched-beam antenna with an overall size of 10 x 10 cm^2 has been fabricated through a technique described in [17]-[18], using an FR-4 substrate with thickness M_t = 0,8 *mm* and dielectric constant ε_r = 4,7 @ 2,45 *GHz* (see Fig. 4). The detailed design parameters of the proposed switched-beam antenna are listed in Table I. The used switches are the Peregrin PE4283 RF UltraCMOS switches with a single-pin CMOS logic control input, a 1.5 kV ESD tolerance, a low insertion loss of 0.65 dB at the reference frequency, an isolation of 33.5 dB between the output ports, a +3 V supply input and an operation band ranging from DC to 4GHz. The study of both radiation patterns and current consumption has been performed through the use of STM32W-EXT WSN boards with a 32 bit ARM micropro-

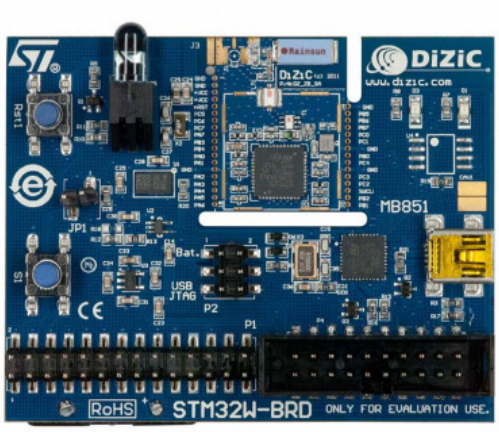


Fig 5. The STM32W-EXT WSN mote used for the experimental measurement.

cessor and an IEEE 802.15.4-compliant transceiver (see Fig. 5).

A. Antenna characterization

In order to characterize the antenna from an electromagnetic point of view, the reflection coefficient and the radiation pattern have been simulated and measured by varying the feeding configurations. In particular, the entire RF structure has been firstly modeled with the full-wave simulator CST- Microwave Studio [19] where the RF switches have been taken into account by means of proper equivalent circuits according to the datasheet. Return loss and radiation properties have been consequently calculated.

As described above, the proposed antenna presents eight switchable directional patterns, four of which, due to the feeding of Ant1-Ant2, Ant3-Ant4, Ant5-Ant6 and Ant7-Ant8 elements, are determined by the same feeding configuration (Conf 1) whereas the other four, due to the feeding of Ant1-Ant3, Ant3-Ant5, Ant5-Ant7 and Ant7-Ant1 elements, are determined by a different configuration (Conf 2). For such a reason, without loss of generality, only data referred to Conf 1 and Conf 2 will be discussed hereafter.

In Fig. 6 and Fig. 7, the measured return loss obtained through a Vector Network Analyzer (VNA) Agilent 3444/7 is compared with the simulated results showing a very good agreement. In particular, both Conf 1 and Conf 2 exhibit an appreciable impedance matching with an observed return loss smaller than -10 dB within the bandwidth of interest.

In Fig. 8 and Fig 9, the simulated radiation patterns in the azimuth ($\varphi = 0^o$), are reported in order to better appreciate the capability of the proposed antenna to correctly orientate the beam when varying the switch configuration. In particular, Fig. 8 is referred to the feeding of the L-shaped antenna element Ant1-Ant2 (belonging to Conf 1 scheme), and Fig. 9 to the feeding of Ant1-Ant3 (belonging to Conf 2 scheme), resulting in two adjacent radiation patterns separated by an angle of 30 degrees. The obtained main lobe magnitude for Conf 1 and Conf 2 are respectively 5.09 dBi and 4.85 dBi, the side lobe levels are respectively -15 dB and -14 dB and the cross-polarization levels are respectively -21 dB and -20 dB. Moreover the Half power beam width (HPBW) is nearly 70 degrees in the azimuth plane for both configurations, according to the proposed design specifications. As can be seen, the proper functioning of the proposed switched-beam antenna, in terms of overlapping area and beam width, are demonstrated.

TABLE I.
DETAILED PARAMETERS OF DESIGNED ANTENNA

Parameters	Value [mm]	Parameters	Value [mm]
ε_r	4,7	d	100
h	0,8	L_1	34
e	1,5	L_2	29
f	1	k	4
l	18	c	6
l_s	4,3	r	12
g	17,8	s	1
g_s	6,1	n_1	4
t_1	48	n_2	7,7
t_2	11	i	39
t_3	25	w_s	0,8
t_4	20	M_t	0,04
r_i	1,3	r_d	4

As further validation, in order to accurately characterize the radiation properties of the proposed antenna, several tests with WSN nodes operating in the ISM band have been performed. In particular, two STM32W108B-KEXT devices have been used, one connected to the proposed antenna and statically positioned in the middle of a 50x50 m^2 area, and the other, with a standard omnidirectional configuration, used to measure the number of packets received in different points of the same area (see Fig. 10). For each radiator, the diagram individuating the portion of the area where more than 95% of the sent packets is correctly received corresponds to the related actually covered area. As shown in Fig. 11 the measurement points are disposed on concentric circumferences with a minimum radius of 10 m (R1) and a maximum radius of 45m (Rn) with an angular distance of 10°. In particular, Fig. 12 and 13 show the obtained diagrams for the two feeding configurations previously introduced (blue curve) and for the canonical dipole (green curve); black dots represent measurements points (see Fig. 11 for the experimental setup). In each case, the lower emitted power has been considered. As expected, it can be observed the proper functioning of the proposed switched-beam antenna, that guarantees a longer working range than the half-wavelength dipole, which is, as expected, substantially omnidirectional in the azimuth plane. Alternatively, a lower power could be radiated to guarantee the same working distance. Moreover, as desired, coverage areas associated to the two feeding configurations guarantee a suitable overlapping area and a beam width compatible with the simulated one. In fact, despite conceptually different, the behaviour of the radiation patterns of Figs. 8 and 9 can be com-

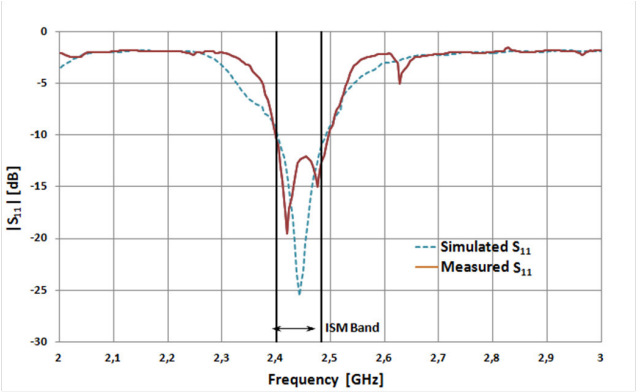


Fig 6. Simulated and measured reflection coefficient for the Conf 1 feeding configuration.

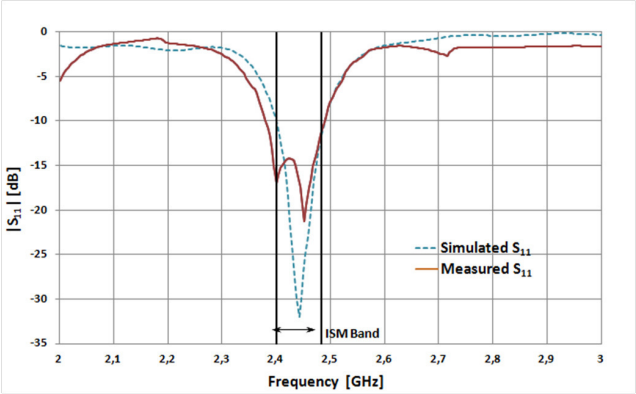


Fig 7. Simulated and measured reflection coefficient for the Conf 2 feeding configuration.

pared with those of the diagrams of Figs. 12 and 13. A substantial agreement can be observed for both feeding configurations previously introduced.

B. *Power consumption*

Several tests have been performed in order to accurately measure the power saving performance of the proposed antenna system. Similarly to the radiation patterns measurement, RSSI values estimated by STM32W-EXT WSN boards connected to the proposed antenna have been used also in this case. As previously explained, when the proposed antenna is used in place of the almost omnidirectional antenna of a mote, greater energy efficiency can be obtained. In order to estimate the power saving, the lower transmission power level P_{tl} to be set on the mote equipped with the proposed antenna, with respect to a canonical mote, must be calculated. For this purpose, a study on the received power varying the distance between the motes has been conducted. For such a purpose, it is considered that the most realistic way to correlate the RSSI to the distance is to use the Log-Normal Shadowing Model (LNSM) [20]-[22] which is able to predict the signal path loss in both indoor and outdoor conditions. This model is an extension particularly suitable for WSN context of the more general log-distance path loss

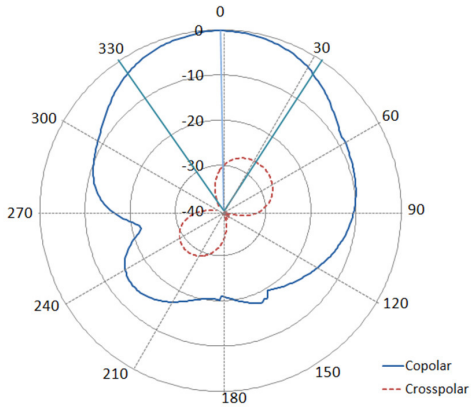


Fig 8. Simulated radiation patterns for the Conf 1 feeding configuration in the azimuth plane.

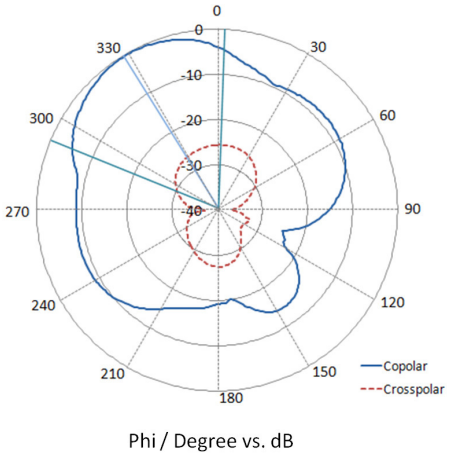


Fig 9. Simulated radiation patterns for the Conf 1 feeding configuration in the azimuth plane.

model described in [20]. Indeed, this last does not consider that the measured power values can be very different from the predicted average ones, as the reflections caused by irregularities of the surrounding environment result in a variance of the measured values. LNSM states that the path loss follows a log-normal distribution (normal in dB) due to the phenomenon called log-normal shadowing. Therefore the received power have a Gaussian distribution with zero mean and can be expressed as:

$$P_r(d) \sim N(\overline{P_r(d)}, \sigma^2)[dBm] \quad (1)$$

where $P_r(d)$ is the average received power and σ^2 is the variance related to the effects of reflections. According to LNSM the equation (1) can be rewritten as:

Fig 10. Measurement setup for the evaluation of the models and the diagrams of the average RSSI.

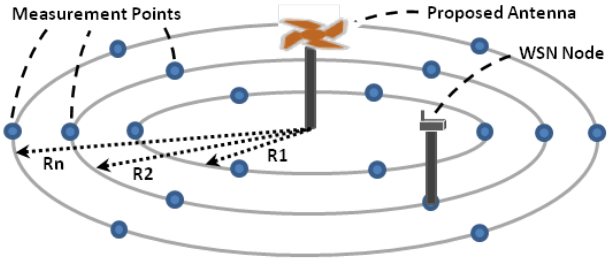


Fig 11. Experimental setup used for the evaluation of the diagrams of the coverage areas.

$$\overline{P_r(d)} = \overline{P_r(d_0)} - 10\, n\, \log_{10}\left(\frac{d}{d_0}\right) + X_\sigma \; [dBm] \quad (2)$$

where $P_r(d_0)$ is the power received at a reference distance d_0 that must be appropriate to the measurement context (in this case we selected a reference distance equal to 10 m), n is the path loss exponent that indicates how the received signal degrades in relation to the distance and depends on the electromagnetic characteristics of the surrounding environment, X_σ is a normal random variable with zero mean.

To obtain a relationship between the received power and the distance between motes, it is necessary to estimate n and X_σ; in practice, they are calculated from the measured power values at various distances, using the linear regression model:

$$\overline{P_r(d)} = \alpha + \beta \log_{10}(d) \; [dBm] \quad (3)$$

where $\alpha = P_r(10) + 10n + X_\sigma$ and $\beta = -10n$. The path loss model of equation (3) can be used to estimate the distance between the transmitting and receiving mote as follows:

$$d = 10^{\frac{\overline{P_r(d)} - \alpha}{\beta}} \; [dBm] \quad (4)$$

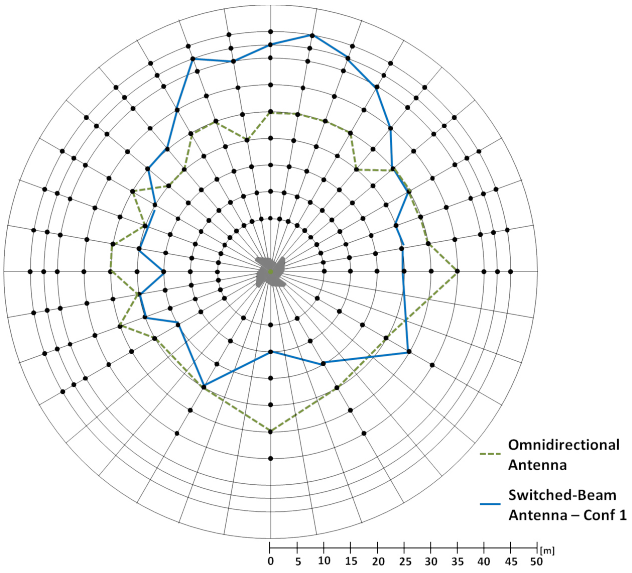


Fig 12. Diagram representing the regions where the WSN node in Conf 1 mode correctly receive more than 95% of packets.

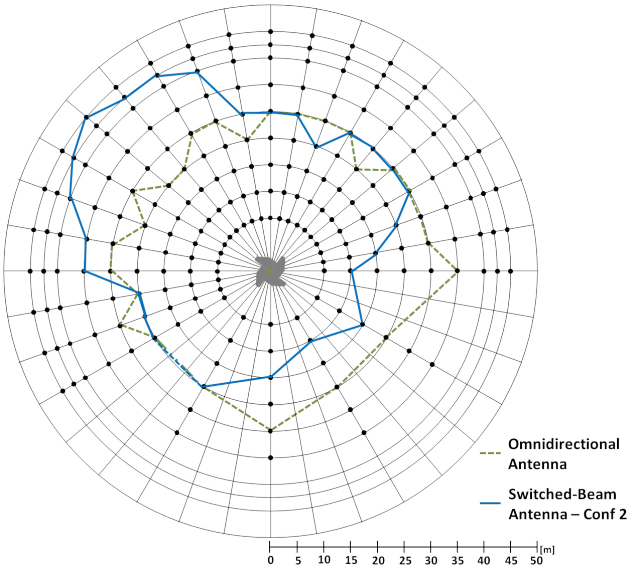


Fig 13. Diagram representing the regions where the WSN node in Conf 2 mode correctly receive more than 95% of packets.

In order to calculate the α and β parameters, and accordingly n and X_σ, several tests have been performed by varying both transmission power and distance and positioning the motes at an height of 1.5m. In particular, two kinds of tests have been performed in Line Of Sight (LOS) condition with surrounding obstacles. The former kind, named T1, has been performed using two motes equipped with a dipole antenna and setting the level P_{tl} to +3 dBm. Vice versa, the latter kind of tests, named T2, has been conducted using the proposed antenna and varying P_{tl}. In Fig. 14, the obtained results related to T1 and T2 tests, this last with P_{tl} to -3 dBm,

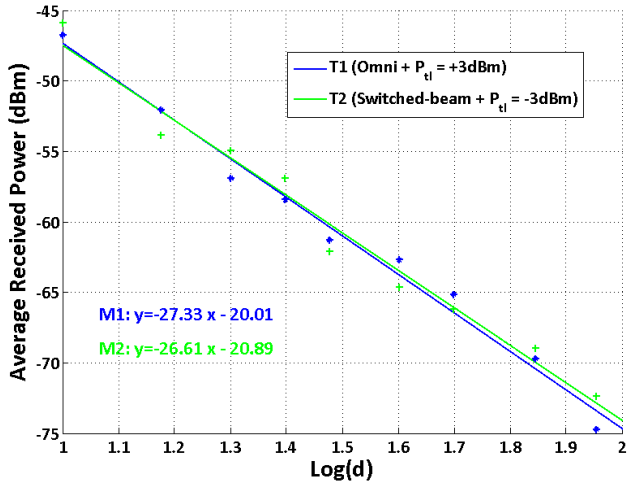


Fig 14. Average received power versus log(d) for the four tests performed.

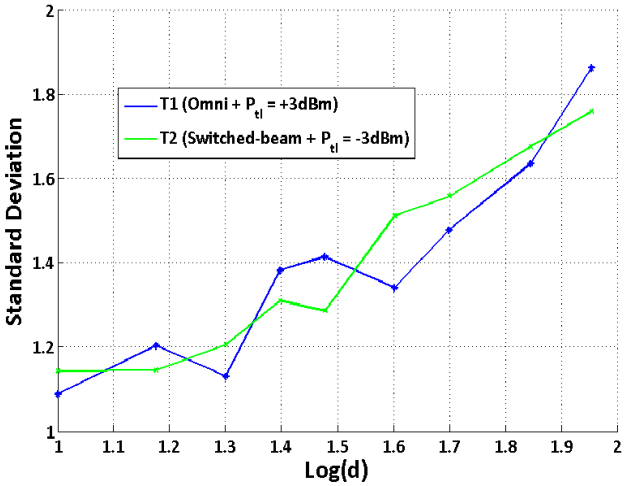


Fig 15. The standard deviations in RSSI measurements for T1 and T2

are reported. As can be seen, despite the strong difference in terms of emitted power, the two curves are in very good agreement; moreover, it is worth observing that measured RSSI values exhibits a linear dependence with log(d). By relating the linear regression model in equation (3) with the experimental measurements performed, the parameters n and X_σ have been determined and reported in Table II.

Fig. 15, shows the standard deviations in RSSI measurements at each distance for the T1 and T2 tests. As can be expected, the deviation in RSSI measurements increases with the distance between the nodes.

Once the parameters n and X_σ are known, α and β parameters can be straightforwardly be calculated and used in equation (4) to obtain the distance as a function of the received power. In Table 3 the estimated distance and the estimation error for the two tests considered are reported. In particular, Fig. 16 shows the comparison between the measured and estimated distances; the closeness of the curves proves the accuracy of the found models.

The presented results confirm the appropriateness of the proposed switched-beam antenna system as hardware element enabling new power-efficient WSNs. Indeed, as shown in Fig 14, the model M1 and model M2 are very similar. This is a crucial result because it allows a fair comparison in terms of power saving when the proposed antenna is used in place of the canonical dipole. Indeed, it is clear that, despite the different Ptl values, T1 and T2 configurations determine the same mote performance.

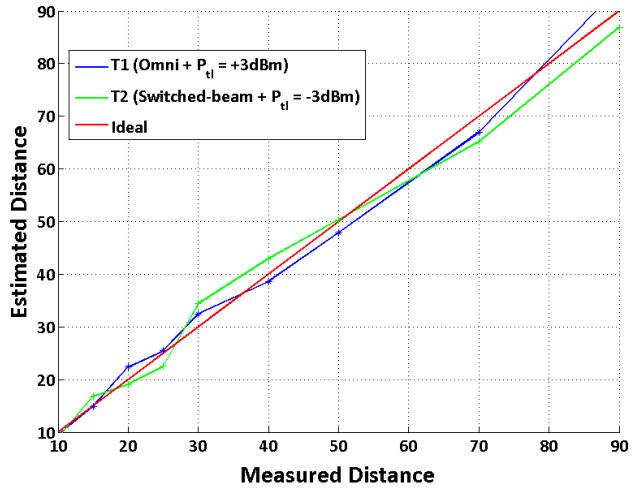


Fig 16. Measured versus estimated distance for the four models.

tests versus log(d).

V. Conclusions

This work proposes a switched-beam antenna particularly suitable for WSN applications. It is composed of four arrays of two L-shaped quarter-wavelength slot antenna elements arranged in a symmetrical planar structure. According to the context-dependent needs of the WSN, one among eight different radiation patterns can be selected by means of properly designed control lines. The antenna system has been rigorously characterized and has been then extensively tested in place of a canonical antenna mounted on a commercial WSN mote. The obtained and discussed results are really encouraging. Indeed, in spite of a very appreciable compactness and cost-effectiveness, the proposed switched-beam antenna guarantees a considerable power saving if compared with traditional and almost omnidirectional WSN mote antennas.

References

[1] G. Anastasi, M. Conti, M. Francesco, A. Passarella, "Energy conservation in wireless sensor networks: A survey," *Ad Hoc Networks*, vol. 7, no. 3, pp. 537-568, May 2009, http://dx.doi.org/10.1016/j.adhoc.2008.06.003

[2] G. Girban, M. Popa, "A glance on WSN lifetime and relevant factors for energy consumption," in *Computational Cybernetics and Technical Informatics (ICCC-CONTI), 2010 International Joint Conference on*, May 2010, pp. 523-528, http://dx.doi.org/10.1109/ICCCYB.2010.5491217

TABLE II
ERROR OF ESTIMATED DISTANCE

Measured Distance [m]	Estimated Distance (T1) [m]	Estimated Distance (T2) [m]	Error (T1) [m]	Error (T2) [m]
10	9,51	8,92	0,49	1,08
15	14,88	16,81	0,12	1,81
20	22,39	19,05	2,39	0,95
25	25,46	22,54	0,46	2,46
30	32,43	34,42	2,43	4,42
40	38,59	42,95	1,41	2,95
50	47,9	50,35	2,1	0,35
70	66,91	65,12	3,09	4,88
90	94,46	86,91	4,46	3,09

[3] Y. Chen, Q. Zhao, "On the Lifetime of Wireless Sensor Networks, " *IEEE Communications Letters*, Vol. 9, no. 11, pp. 976-978, Nov. 2005, http://dx.doi.org/10.1109/LCOMM.2005.11010

[4] L. Anchora, A. Capone, V. Mighali, L. Patrono, F. Simone, "A novel MAC scheduler to minimize the energy consumption in a Wireless Sensor Network," *Ad Hoc Networks*, vol. 16, pp. 88-104, May 2014, http://dx.doi.org/10.1016/j.adhoc.2013.12.002

[5] L. Catarinucci, R. Colella, G. Del Fiore, L. Mainetti, V. Mighali, L. Patrono, M.L. Stefanizzi, "A cross-layer approach to minimize the energy consumption in wireless sensor networks, " *International Journal of Distributed Sensor Networks*, vol. 2014, 2014, http://dx.doi.org/10.1155/2014/268284

[6] D. Alessandrelli, L. Mainetti, L. Patrono, G. Pellerano, M. Petracca, M.L. Stefanizzi, "Performance evaluation of an energy-efficient MAC scheduler by using a test bed approach," *Journal of Communications Software and Systems*, vol. 9, no. 1, pp. 84-96, Mar. 2013.

[7] G. Giorgetti, A. Cidronali, S.K.S. Gupta, G. Manes, "Exploiting Low-Cost Directional Antennas in 2.4 GHz IEEE 802.15.4 Wireless Sensor Networks", in *Proc. 10th European Conference on Wireless Technology*, Oct. 2007, pp. 217-220.

[8] S. Zhang, G. H. Huff, and J. T. Bernhard, "A pattern reconfigurable microstrip parasitic array," *IEEE Trans. Antennas Propag.*, vol. 52, no. 10, pp. 2773-2776, Oct. 2004, http://dx.doi.org/10.1109/ECWT.2007.4403985

[9] M.-I. Lai, T.-Y. Wu, J.-C. Hsieh, C.-H. Wang, S.-K. Jeng, "Pattern Reconfigurable Antenna for a Wireless Sensor Network Sink Node," in *Proc. Asia-Pacific Microwave Conference (APMC)*, Dec. 2010, pp. 2021-2024.

[10] D.-C. Chang, B.-H. Zeng, J.-C. Liu, "Reconfigurable angular diversity antenna with quad corner reflector arrays for 2.4 GHz applications," *IET Microw., Antennas Propag.*, vol. 3, no 3, pp. 522–528, Apr. http://dx.doi.org/2009, 10.1049/iet-map.2008.0119

[11] L. Ming-Iu, W. Tzung-Yu, H. Jung-Chin, W. Chun-Hsiung, J. Shyh-Kang, "Compact Switched-Beam Antenna Employing a Four-Element Slot Antenna Array for Digital Home Applications, " *IEEE Trans. Antennas Propag.*, vol. 56 , no.9, pp 2929-2936, Sept. 2008. http://dx.doi.org/10.1109/TAP.2008.928775

[12] L. Catarinucci, S. Guglielmi, L. Patrono, and L. Tarricone, "Switched-beam antenna for wireless sensor network nodes," *Progress in Electromagnetics Research C*, vol. 39, pp. 193–207, 2013, http://dx.doi.org/10.2528/PIERC13030707

[13] L. Catarinucci, S. Guglielmi, L. Mainetti, V. Mighali, L. Patrono, M.L. Stefanizzi, L. Tarricone, "An energy-efficient MAC scheduler based on a switched-beam antenna for wireless sensor networks," *Journal of Communications Software and Systems*, vol. 9, no. 2, pp. 117-127, Jun. 2013.

[14] L. Catarinucci, S. Guglielmi, R. Colella, L. Tarricone, "Compact switched-beam antenna enabling novel power-efficient wireless sensor network," *IEEE Sensors J.*, vol. PP, no 99, pp. 1-9, May 2014, http://dx.doi.org/10.1109/JSEN.2014.2326971

[15] S. I. Latif, L. Shafai, and S. K. Sharma, "Bandwidth enhancement and size reduction of microstrip slot antennas," *IEEE Trans. Antennas Propag.*, vol. 53, no. 3, pp. 994–1003, Mar. 2005, http://dx.doi.org/10.1109/TAP.2004.842674

[16] S. K. Sharma, L. Shafai, and N. Jacob, "Investigate of wide-band microstrip slot antenna," *IEEE Trans. Antennas Propag.*, vol. 52, no. 3, pp. 865–872, Mar. 2004, http://dx.doi.org/10.1109/TAP.2004.825191

[17] L. Catarinucci, R. Colella, and L. Tarricone, "Smart Prototyping Techniques for UHF RFID Tags: Electromagnetic Characterization and Comparison with Traditional Approaches," *Progress In Electromagnetics Research*, Vol. 132, pp 91-111, Sep. 2012. http://dx.doi.org/10.2528/PIER12080708

[18] L. Catarinucci, R. Colella, L. Tarricone, "Prototyping Flexible UHF RFID Tags Through Rapid and Effective Unconventional Techniques: Validation on Label-Type Sensor-Tag," *IEEE 2012 International Conference on RFID -Technologies and Applications (RFID - TA)*, p. 176-181, Nice, Nov. 2012, http://dx.doi.org/10.1109/RFID-TA.2012.6404506

[19] CST - Computer Simulation Technology. [Online], https://www.cst.com, (last accessed: March. 2014)

[20] T.S. Rappaport, *Wireless Communications: Principles and Practice, (2nd Edition)*, Prentice Hall, New York, Jan. 2007.

[21] R. Al Alawi, "RSSI based location estimation in wireless sensors networks", in *Proc. 17th IEEE International Conference on Networks (ICON)*, Dec. 2007, pp. 118-122, http://dx.doi.org/10.1109/ICON.2011.6168517

[22] P. Kumar, L. Reddy, and S. Varma, "Distance measurement and error estimation scheme for rssi based localization in wireless sensor networks," in *Proc IEEE Conference on Wireless Communication and Sensor Networks (WCSN)*, Dec. 2009, pp. 1-4, http://dx.doi.org/10.1109/WCSN.2009.5434802

Information Technology for Management, Business & Society

IT4MBS is a FedCSIS conference area aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to the disciplines of information technology and information systems. The IT4BMS area emphasizes the issues relevant to information technology and necessary for practical, everyday needs of business, other organizations and society at large. This area takes a sociotechnical view on information systems and relates also to ethical, social and political issues raised by information systems.

Events that constitute IT4BMS are:

- **ABICT'14** - 5th International Workshop on Advances in Business ICT
- **AITM'14** - 12th Conference on Advanced Information Technologies for Management
- **IT4L'14** - 3rd Workshop on Information Technologies for Logistics
- **KAM'14** - 20th Conference on Knowledge Acquisition and Management
- **SS4SI'14** - 1st International Symposium on Service Systems for Social Innovation

5th International Workshop on Advances in Business ICT

ABICT focuses on Advances in Business ICT approached from a multidisciplinary perspective. It will provide an international forum for scientists/experts from academia and industry to discuss and exchange current results, applications, new ideas of ongoing research and experience on all aspects of Business Intelligence. ABICT will be also an opportunity to demonstrate different ideas and tools for developing and supporting organizational creativity, as well as advances in decision support systems.

We kindly invite contributions originating from any area of computer science, information technology and computational solutions for different applications areas, data integration and organizational implementation of ABICT, as well as practical ABICT solutions.

Topics

Topics include (but are not limited to):

- Advanced technologies of data processing, content processing and information indexing
- Analytics as a service
- Big Data: benefits and challenges
- Business Analytics
- Business applications of social networks
- Business data mining and knowledge discovery
- Business Intelligence
- Business Rules
- Business-oriented time series data mining, analysis, and processing
- Cloud based Business Intelligence
- Creativity Support Tools
- Customer Relationship Management, social Customer Relationship Management
- Data driven marketing
- Data Warehousing
- Decision support
- Information forensics and security, information management, risk assessment and analysis
- ICT technologies in enterprise management
- Knowledge Management (for better Decision Support, Collaboration and Competitiveness)
- Legal text processing
- Semantic Web and Ontologies in Business ICT
- Virtual Enterprise
- Web 2.0 and Web 3.0 in fusing Business Intelligence systems and Decision Support Systems
- Web-Based Data Management Systems

Event Chairs

Mach-Król, Maria, Katowice University of Economics, Poland

Olszak, Celina M., University of Economics in Katowice, Poland

Pełech-Pilichowski, Tomasz, AGH University of Science and Technology, Poland

Program Committee

Abramowicz, Witold, Poznań University of Economics

Badica, Amelia, University of Craiova, Romania

Berio, Giuseppe, Universite de Bretagne Sud, France

Chiu, Dickson K. W., Dickson Computer Systems, Hong Kong S.A.R., China

Christozov, Dimitar, American University in Bulgaria, Bulgaria

Gaweł, Bartłomiej, AGH University of Science and Technology

Kacprzyk, Janusz, Institute of Computer Science, Polish Academy of Sciences, Poland

Khachidze, Manana, Tbilisi State University, Georgia

Konikowska, Beata, Institute of Computer Science, Poland

Koohang, Alex, Macon State Collage, United States

Korwin-Pawlowski, Michael L., Universite du Quebec en Outaouais, Canada

Kulczycki, Piotr, Systems Research Institute, Polish Academy of Sciences, Poland

Ligęza, Antoni, AGH University of Science and Technology, Poland

Loucopoulos, Peri, Harokopio University of Athens, Greece

Maamar, Zakaria, Zayed University, United Arab Emirates

Michalik, Krzysztof

Nycz, Malgorzata, Wroclaw University of Economics, Poland

Ogihara, Mitsunori, University of Miami, United States

Owoc, Mieczyslaw, Wroclaw University of Economics, Poland

Petryshyn, Lubomyr, AGH University of Science and Technology, Poland

Prasad, T. V., Visvodaya Technical Academy, India

Pulvermueller, Elke, University Osnabrueck, Germany

Reimer, Ulrich, University of Applied Sciences St. Gallen, Switzerland

Rossi, Gustavo, National University of La Plata, Argentina

Salem, Abdel-Badeeh M., Ain Shams University, Egypt

Sauer, Jurgen, University of Oldenburg, Germany

Szpyrka, Marcin, AGH University of Science and Technology, Poland

Teufel, Stephanie, University of Fribourg, Switzerland

Whatley, Janice, University of Salford, United Kingdom

Wrycza, Stanisław, University of Gdansk, Poland

Zadrozny, Slawomir, Systems Research Institute, Poland

Zurada, Jozef, University of Louisville, United States

Zurada, Jozef, College of Business University of Louisville, Louisville

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1091–1096
DOI: 10.15439/2014F290
ACSIS, Vol. 2

Selected Aspects of Temporal Knowledge Engineering

Maria Mach-Król
University of Economics
ul. Bogucicka 3,
40-226 Katowice Poland
Email: maria.mach-krol@ue.katowice.pl

Krzysztof Michalik
University of Economics
ul. Bogucicka 3,
40-226 Katowice Poland
Email:
krzysztof.michalik@ue.katowice.pl

***Abstract*—The paper presents some problems of logical coherence while reasoning temporally. It shows the importance of these problems in some application domains for temporal intelligent systems, e.g. in legal domain. It then presents Logos reasoning tool and its inference techniques, it also shows how Logos can handle temporal rules now, and it points out what should still be done in order to make the system resistant to temporal reasoning logical problems.**

I. Introduction

TIME indispensable representation in many artificial and temporal intelligence reasoning systems [1]. are It is so, because time is a basis for reasoning about change and actions. Many AI systems concern the currently changing economic environment (see e.g. [2]). Therefore, if the decisions taken on the basis of system's advice are to be correct, the system has to take into account the temporal dimension of information, the changes of information in time and has to be aware of the nature of those changes [1]. Therefore the need for representing temporal knowledge in artificial intelligence systems is nowadays obvious. Even intuitively, one can feel the need of capturing a temporal aspect of relationships between objects in AI systems.

Representation of knowledge changing in time and temporal reasoning have to be based on some formal foundations. One of such foundations is the language of logic, both the classical as the modal one (see e.g. [3], [4]). The shortest motivation for using the temporal logic may be found in [5], where it is concluded, that "in order to introduce temporal relationships (...) it is necessary to broaden the formal apparatus with the temporal logic" (p. 429).

The paper is organized as follows. In section 2 the general view of temporal knowledge is given. Section 3 contains a few examples of temporal knowledge engineering problems. Section 4 is devoted to the discussion on inference techniques already implemented in the Logos tool. The next section shows the importance of knowledge verification as important knowledge engineering process, and how it is handled by Logos. In section 6 we provide an example of temporal reasoning in the Logos system. The paper ends with summary and conclusions.

II. Knowledge Engineering Problems

The tasks for a temporal AI system encompass among others:

- maintaining temporal coherence,
- answering temporal queries,
- explanations,
- prediction, etc.

The most commonly known domains, in which there is a need for an explicit time notion, are: natural language processing (NLP), planning, robotics, image processing, medical diagnosis, and law [1], [2]. In all of the above mentioned domains, change has a primary meaning. Introducing time in an explicit way allows for reasoning about changing domains, also about the economic one. It also allows for a computer simulation of human reasoning process, because people reason about action and change [3]. In particular, there are described such notions, as change, causality or actions, therefore the proper representation of time and temporal reasoning are so important in (among others) artificial intelligence [1]. If an AI system is to simulate intelligent behavior, to adapt to changes in the environment, or to verify its beliefs, it has to be able not only to gain new knowledge, but also to keep its knowledge in an up to date state. Knowledge changes – due to two basic reasons. The first is simply the passage of time. The second reason is due to new information on objects, which possess temporal characteristics [4].

III. Knowledge Engineering Problems

There exist several problems which one has to overcome in some way to reason consistently about time and change. In the paper there will be mentioned three problematic issues: the TBS (Tossed Ball Scenario), the DIP (Divided Instant Problem), and finally the FP (Frame Problem).

Tossed Ball Scenario (TBS)

The problem called the Tossed Ball Scenario is connected with the question of temporal primitives in the ontology of time. As temporal primitives there can be chosen [1]:

- time points (also called instants) – as for example in McDermott's logic;
- time periods (also called intervals) – as for example in Allen's interval calculus;
- both primitives.

In logic systems using the notion of time points (instants), there is a question of modeling continuous change (e.g. in the environment). This is considered a problematic issue, because

978-83-60810-58-3/$25.00

to do it properly one must be able to describe fluents that hold at an instant, more precisely – an isolated instant.

Divided Instant Problem (DIP)

The problem is similar in its nucleus to the TBS. It consists of establishing logical value (truth-value) of a fluent f at an instant i, if f is true on period p1 and false on period p2, given that p1 ends at i and p2 begins at i. Of course we assume that we take instants and periods as temporal primitives. Let us discuss the problem using an example given by [5]. Consider fig. 1:

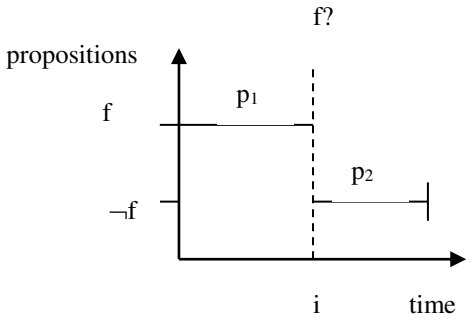


Fig. 1.: Divided Instant Problem [6]; "f" denotes fluent, "p1" and "p2" denote periods.

Vila gives an example concerning light. If fluent f means "the light is on", then on p_1 light is on and on p_2 light is off. Question: is the light on or off at i? More generally: what is a truth-value of the proposition at i?

Frame Problem (FP)

The frame problem is one of the central theoretical issues of artificial intelligence. The problem arises when logic is used to describe the effects of actions and events. The frame problem(s) appears in all approaches to reasoning about action and change [6].

In order to be historically correct, the first description of this problem has been done by McCarthy and Hayes in 1969 [7]. While working on situation calculus they noted that a problem occurs when there are several actions available, each of which changes certain features of the situation.

The classical examples of such problems are: 'Yale Shooting Problem' , 'Stolen Car Problem' and others [8]. Let's take a look on the first example. Given the sequence of events at times as denoted by the times of their assertion t(n):

t(0): I (now) pick up a loaded gun

t(1): I unload it

t(3): I point the gun at my head

t(4): I pull the trigger

The answers to the question: 'Am I alive?' should be negative as the representation did not contain the information that the gun was still unloaded (there is no information about the change of fact - gun is unloaded in t(3)).

There were already many attempts to solve the above enumerated logical problems, e.g. by [5] or by [9]. Especially the results obtained by Reiter are interesting, as he uses situation calculus. And this formalism or its mutation called Golog [10] are planned by us to be implemented in the intelligent system Logos. Therefore we hope Logos will be able not only to reason temporally, but to do it in a consistent way, that is the tool will be resistant to temporal logic inference problems. Some previous remarks on reasoning temporally about the legal domain in Logos were published in [11]. Logos now achieved the status of a research prototype and is treated as environment and reasoning engine for temporal knowledge bases experiments.

In the next section we will present currently implemented reasoning strategies in Logos, and we will show how the tool can now handle temporal rules.

IV. Inference in Logos

In present version of Logos system we successfully implemented four variants of inference:

1. Top-down inference using two-valued logic,
2. Top-down inference using Stanford Certainty Factor Algebra,
3. Bottom-up inference using two-valued logic,
4. Bottom-up inference using Stanford Certainty Factor Algebra.

Each of these methods of inference can operate in two modes: with askable or not askable conditions. The top down inference is goal-driven and is similar to that used in logic programs with backtracking mechanism which makes it possible to find all solutions to a given problem. Generally this method of inference is based on resolution principle developed by Robinson [12]. The whole process of inference starts with a given goal:

```
<- C1,..,Cm
```

In procedural interpretation each step of computation relies on matching of a given goal C_i with the head of any procedure in knowledge base. Procedure is Horn clause (rule) of the following form:

```
A <- B1,..,Bn
```

Where $\{C_1,..C_k,A, B_1,..,B_m\}$ are atomic formulas. In our implementation those atomic formulas take the form < O, A, V >, where O is identifier of object, A is identifier of attribute and V represents value. Value can be symbol (more generally any string), numeric constant or variable.

If the mentioned matching is not possible, we call it failure. If the matching of C_i is successful, then current goal is being reduced to the following form:

```
<- (C1,..,Ci-1, B1,..,Bn,Ci+1,..,Cm)Θ,
```

where Θ is called matching substitution, mainly concerned with assignment of temporary values to variables.

The inference process finishes when the goal is reduced to empty clause.

The bottom-up inference, as implemented in our system, starts from facts and using rules of knowledge base finishes with generation of set of conclusions.

While building inference module of Logos we assumed that in temporal knowledge base some parts of knowledge (facts or rules) can be to some degree uncertain, so we implemented

Stanford Certainty Factor Algebra. We chose this method because of the simplicity of its use and good practical verification. In this method for given rule R:

R: C if W with CF(R)

is assigned certainty factor CF(R) and its value is in the range <-1; 1>. The current CF for a conclusion of the rule R is dynamically computed in the following way:

$$CF(C) = CF(R) * CF(W).$$

If antecedent of a given rule consists of a set of conditions e.g. W_1 and W_2 joined by disjunction or conjunction its value is calculated as follows:

$$CF(W1 \wedge W2) = MIN \{ CF(W1), CF(W2) \}$$
$$CF(W1 \vee W2) = MAX \{ CF(W1), CF(W2) \}$$

During inference, especially using bottom-up (forward chaining) method it is possible that several rules add the same conclusion to the working memory of the knowledge base. In such situation the CF of the added conclusion have to be dynamically changed. Calculation of the new CF for two rules R_1 and R_2 is as follows:

$$CF(C) = CF(R1) + CF(R2) * (1 - CF(R1))$$
if $CF(R1) > 0$ and $CF(R2) > 0$

$$CF(C) = CF(R1) + CF(R2) * (1 + CF(R2))$$
if $CF(R1) < 0$ and $CF(R2) < 0$

$$CF(C) = (CF(R1) + CF(R2)) / (1 - MIN \{ |CF(R1)|, |CF(R2)| \}),$$

when the signs of $CF(R_1)$ and $CF(R_2)$ are different.

The next step of our empirical research will attempt to implement inference of temporal situation calculus method using reasoning schemes of already implemented methods. Problem of uncertainty will be important part of the temporal inference and knowledge representation language.

Logos inference engine is equipped with rich set of explanation facilities, among others:

- How explanations,
- Why explanations,
- Metaphors,
- What is explanations and
- Facts descriptions.

How explanations show the way the conclusion has been derived and can be used after the end of reasoning process. Why explanations explain the reason system asked a question during consultation. In this case Logos shows the current context of reasoning and shows in what way the answer will contribute to solving the problem. Metaphors enable knowledge engineer to attach more textual information about selected rules, what can be useful at initial stage of using the application. What is explanation provide more detailed, textual information about conclusions as well as some questions. It is also possible to attach some explanations to facts, in the form of facts descriptions, e.g.: source of information or availability of deeper/further How explanations (in the blackboard architecture) showing how the fact has been derived during consultation.

The mentioned explanations are equally important in usual knowledge bases and in the temporal ones.

V. Knowledge Verification as Important Knowledge Engineering Process

As it has been already mentioned, one of the main aims of our research is creation of temporal knowledge base using Logos system. To this end we started building special reasoning system called Logos which will be kind of experimental environment. One of our assumptions is that appropriate verification algorithms are necessary to provide solid foundation for temporal reasoning. At present we implemented broad range of knowledge base anomalies detection procedures, among others [13]:

1. Redundant rules,
2. Subsuming rules,
3. Contradictory rules,
4. Recursive rules (circular loop).

Ad. 1.

Two rules we call redundant if

$R_i \leftarrow C_{i1} \wedge \ldots \wedge C_{in}$ and $R_j \leftarrow C_{j1} \wedge \ldots \wedge C_{jn}$

where R_i and R_j are conclusions and C are conditions and $i \neq j$, holds: $\{ C_{i1}, \ldots, C_{in} \} = \{ C_{j1}, \ldots, C_{jn} \}$.

Ad. 2.

If for two different rules: $R_i \leftarrow C_{i1} \wedge \ldots \wedge C_{im}$ and $R_j \leftarrow C_{j1} \wedge \ldots \wedge C_{jn}$ and $i \neq j$,

holds $\{ C_{i1}, \ldots, C_{im} \} \subset \{ C_{j1}, \ldots, C_{jn} \}$, then we say that rule R_i subsumes rule R_j.

Ad. 3.

Two rules are regarded as contradictory if $R_i \leftarrow C_{i1} \wedge \ldots \wedge C_{in}$ and $R_j \leftarrow C_{j1} \wedge \ldots \wedge C_{jn}$ where $i \neq j$.

Ad. 4.

We distinguish in Logos two kinds of the recursion: direct and indirect. The notion of direct recursion is consistent with notion of circular rule set by Vermesan [14]: a rule set is circular iff the antecedents cannot be derived from any other rule except given rule consequent.

Typical situations are as follows:

$q \leftarrow q, p_1, .., p_n$

$q \leftarrow p_1, .., p_1, q, p_m, .., p_n$

$q \leftarrow p_1, .., p_n, q$

In practice the circular loop can be much more complicated and engage some set of rules (see fig. 2). Detection of such situation is more difficult but is necessary the knowledge base to function properly. Exemplification of this kind of indirect recursion can be as follows:

$q_1 \leftarrow p_{11} \wedge .. \wedge p_{1j} \wedge .. \wedge p_{1x}$

$q_i \leftarrow p_{i1} \wedge .. \wedge p_{il} \wedge .. \wedge p_{iy}$

$q_n \leftarrow p_{n1} \wedge .. \wedge p_{nm} \wedge .. \wedge p_{nz}$

where $p_{1j} = q_i$, $p_{il} = q_n$, $p_{nm} = q_1$.

where "=" has more general meaning and denotes beside equality/identity ability of instantiation and matching . Our system detects all levels of recursive interconnections between rules, what is important to guarantee correct behavior of knowledge base in practical use.

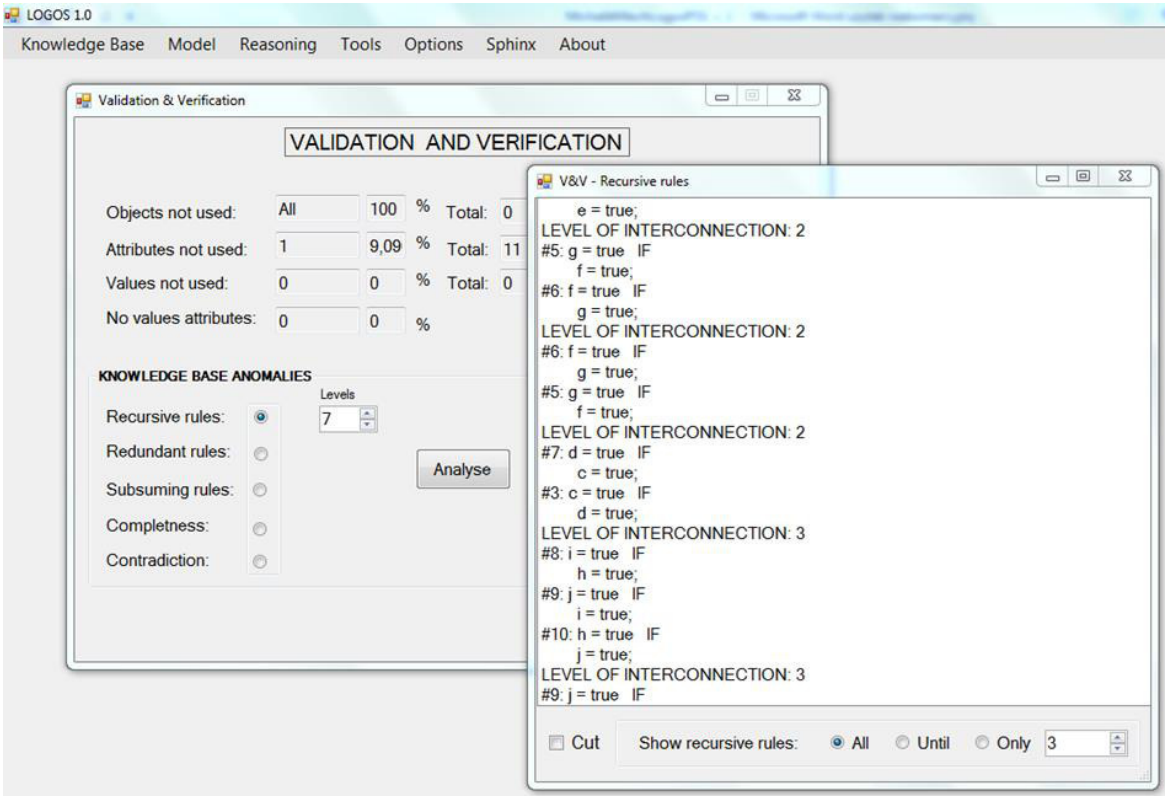


Fig. 2. Sample circular loop in Logos.

Source: own elaboration.

What is very important, we expect that full implementation of temporal knowledge base and temporal inference will require some specific verification algorithms in response to anomalies in temporal knowledge base. At this stage of our empirical research it is difficult to precisely point out these anomalies, it will be subject of further investigations.

VI. Temporal Reasoning in Logos

To show how Logos performs temporal reasoning, we have chosen an example arising from Polish act on economic activity. It concerns the obligation to obtain a license for certain economic activities: a certain enterprise wishes to perform an activity, that needs a license obtained from the state.

The act issued in 1988, that has been valid from January 1st 1989 till December 31st, 2000, stated, that "A license is given indefinitely"; while the act issued in 1999, valid from January 1st, 2001 till July 19th, 2004 stated: "A license is given for a specific period of time, not shorter than 2 years and not longer than 50 years". The actual act of July 2nd, 2004, which is valid from July 20th, 2004 till now, states that "A license is given for a specific period of time, not shorter than 5 years and not longer than 50 years".

While analyzing the above statements as a temporal legal knowledge, we can distinguish such temporal elements as events (an act of granting an enterprise a license), objects (a license), temporal relations (not shorter than, not longer than), temporal constants (2 years, 5 years, 50 years).

The basic temporal elements are points (e.g. the date when a license is granted) and intervals (e.g. the period for which a license is valid). We may also consider – in a broader perspective – an event consisting of "license withdrawal" which of course always happens before the period of license validity ends. Therefore while formalizing legal statements one should choose a formalism that keeps total linear order of time, with precedence relation. The formalism has to be a point-interval one, as we have both types of basic temporal elements to be taken into consideration. All the above conditions are fulfilled by the model of calendar time.

If we take into account the regulations coming from both acts, we immediately see two of the three temporal aspects of legal knowledge, discussed earlier. These are namely "law in time" (three periods of legal acts validity) and "time in law" (a period for which a license is granted) aspects. The third aspect – "transitional law" – will not be discussed here.

The above cited law articles of our example may be also written in the form of general rules:

If a license is issued, then it is valid indefinitely (Act of 1988)

If a license is issued, then it is valid for a period not shorter than 2 years and not longer than 50 years (Act of 1999)

If a license is issued, then it is valid for a period not shorter than 5 years and not longer than 50 years (Act of 2004)

The rules may be formalized e.g. in the Legal Temporal Representation (LTR) language [15], and they are as follows:

```
  Attributes(license,{who_issues,who_gets
})
  Attributes(is_issued, {what})
  Attributes(valid, {what},_,)
  Granularity(day)
```

Rules for the 1988 Act:
Rule 1:

```
  If TT1:license(issuing_authority,
enterprise)
     TT2: is_issued(TT1)
     Occurs(TT2)
  Then Occurs(valid(TT1), instant(TT2))
```

Rule 2:

```
  If TT2: valid(TT1)
     Occurs(TT2)
  Then Holds_on(valid(TT1), period(TT3))
    Period(TT3)   Equals   [instant(TT2),
+inf]
```

The above rules state, that if a certain enterprise has been granted a license issued by a certain authority and it happened in the point (day) stamped by token TT1, then the license is valid from this day, and for indefinite time (that is, over an interval from TT1 to infinity).

Rules for the 1999 Act:
Rule 3:

```
  If TT1:license(issuing_authority,
enterprise)
     TT2: is_issued(TT1)
     Occurs(TT2)
  Then Occurs(valid(TT1), instant(TT2))
```

Rule 4:

```
  If TT2: valid(TT1)
     Occurs(TT2)
  Then Holds_on(valid(TT1), period(TT3))
    Period(TT3) Equals [2y, 50y]
```

Rules for the 2004 Act:
Rule 5:

```
  If TT1:license(issuing_authority,
enterprise)
     TT2: is_issued(TT1)
     Occurs(TT2)
  Then Occurs(valid(TT1), instant(TT2))
```

Rule 6:

```
  If TT2: valid(TT1)
     Occurs(TT2)
  Then Holds_on(valid(TT1), period(TT3))
    Period(TT3) Equals [5y, 50y]
```

As it can be easily seen, rules 4 and 6 differ from rule 2 only in the length of a period over which a license is valid, while rules 1, 3 and 5 are identical. Example of the use of the discussed rules for the reasoning along with HOW explanations in the Logos system is shown in fig. 3.

VII. Summary and Conclusions

As it has been mentioned, in some areas, especially in the legal domain, temporal reasoning is natural part of problem solving. Therefore it should be taken into account while building real-life decision support systems in that domain. Our research is aimed to add temporality to knowledge representation of legal knowledge for building more adequate knowledge base. At present stage of our work we are trying to build temporal knowledge base using Logos reasoning system. Important part of any knowledge-based system should be module for automatic detection of all possible anomalies in the knowledge base, so we implemented wide range of special algorithms for that purpose. The next step it will be adding more temporal relations to our system, and to implement temporal reasoning in the Situation Calculus in order to overcome logical inference problems, pointed out in this paper. If we are succeed as far as building prototype temporal knowledge base, we will try to implement such application into practice.

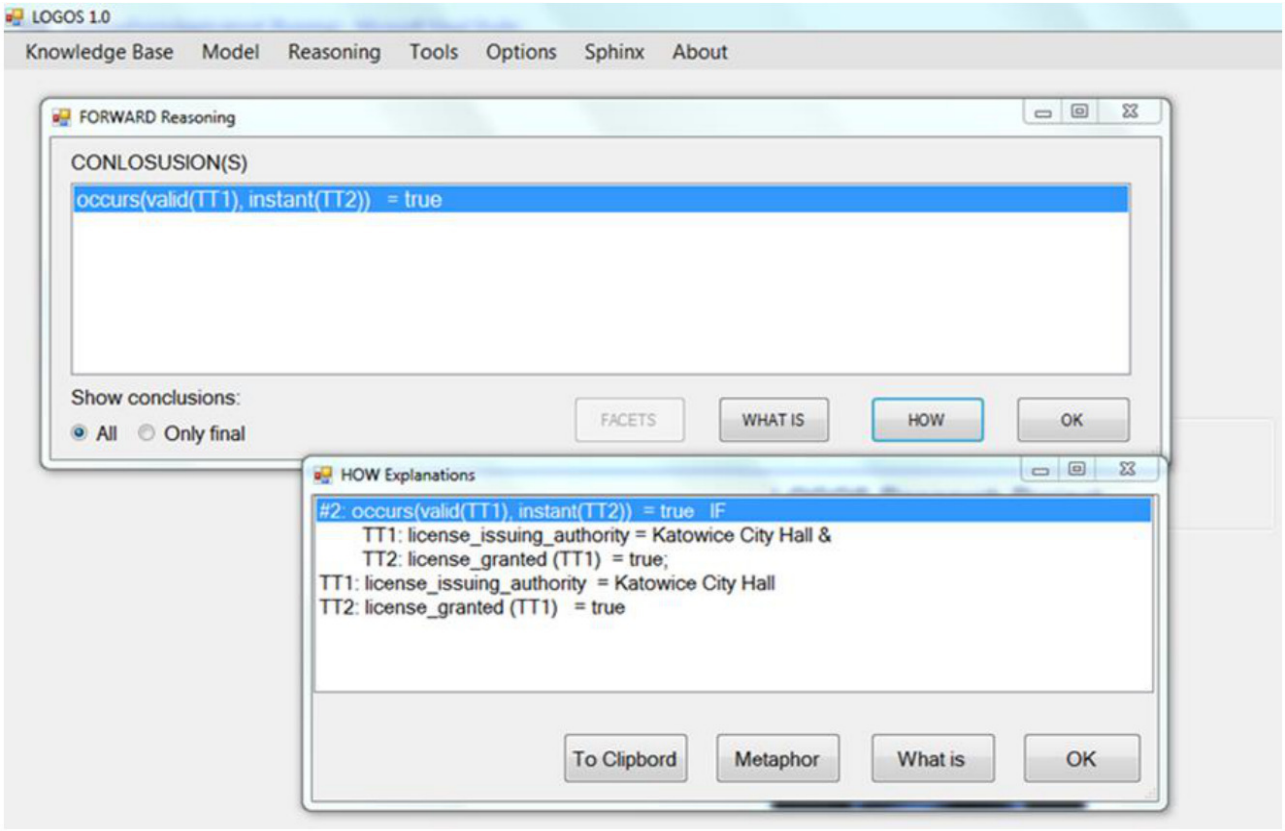


Fig. 3. Example of reasoning using temporal rulesSource: ownand HOW explanation facilities of Logos.
Source: own elaboration.

REFERENCES

[1] L. Vila, „A Survey on Temporal Reasoning in Artificial Intelligence," AI Communications, nr 7(1), pp. 4-28, 1994.

[2] T. Bench-Capon and F. Coenen, "The maintenance of legal knowledge based systems," *Artificial Intelligence Review,* vol. 6, no. 2, 1992.

[3] A. Galton, „Time and Change for AI," in Handbook of Logic in *Artificial Intelligence and Logic Programming. Vol. 4: Epistemic and Temporal Reasoning,* D. Gabbay, C. Hogger i J. Robinson, Eds., Oxford, Clarendon Press, 1995.

[4] J. Benthem van, „Temporal Logic," in Handbook of Logic in *Artificial Intelligence and Logic Programming. Volume 4: Epistemic and Temporal Reasoning,* D. M. Gabbay, C. J. Hogger, and J. A. Robinson, Eds., Oxford, Clarendon Press, 1995, pp. 241-350.

[5] L. Vila and E. Schwalb, "Revisiting Time and Temporal Incidence," in *Proc. of the AAAI'96 Workshop on Spatial and Temporal Reasoning,* 1996.

[6] M. Fisher, D. Gabbay and L. Vila, Eds., Handbook of Temporal Reasoning in Artificial Intelligence, Amsterdam: Elsevier, 2005.

[7] J. McCarthy and P. Hayes, "Some philosophical problems from the standpoint of artificial intelligence," *Machine Intelligence,* vol. 4, pp. 463-502, 1969.

[8] A. Zambak, „The Frame Problem," in *Philosophy and Theory of Artificial Intelligence,* V. Müller, Ed., Berlin Heidelberg, Springer, 2013, pp. 307-319.

[9] R. Reiter, „The frame problem in the situation calculus: A simple solution (sometimes) and a completeness result for goal regression," in *Artificial intelligence and mathematical theory of computation: papers in honor of John McCarthy,* V. Lifschitz, Ed., San Diego, Academic Press Professional, Inc., 1991, pp. 359-380.

[10] J. Claßen, „Planning and Verification in the Agent Language Golog," Aachen, 2013.

[11] M. Mach-Król and K. Michalik, "An Intelligent System with Temporal Extension for Reasoning About Legal Knowledge," in *Looking into the Future of Creativity and Decision Support Systems,* A. Skulimowski, Ed., Kraków, Progress & Business Publishers, 2013, pp. 360-369.

[12] J. Robinson, „A Machine-Oriented Logic Based on the Resolution Principle," *Journal of the ACM,* Vol. 12, No. 1, pp. 23-41 , Jan. 1965.

[13] K. Michalik, M. Kwiatkowska and K. Kielan, "Application of Knowledge-Engineering Methods in *Medical Knowledge Management," in Fuzziness and Medicine: Philosophical Reflections and Application Systems in Health Care, vol. III,* R. Seising and M. E. Tabacchi, Eds., Berlin Heidelberg, Springer, 2013, pp. 205-214.

[14] A. Vermesan, „Foundation and Application of Expert System Verification and Validation," w *The Handbook of Applied Expert Systems,* J. Liebowitz, Ed., Boca Raton, CRC Press LLC, 1998, pp. 5-1-5-32.

[15] L. Vila and H. Yoshino, "Time in automated legal reasoning," *Information and Communications Technology Law,* vol. 7, no. 3, 1998.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1097–1102
DOI: 10.15439/2014F185
ACSIS, Vol. 2

A Note on BPMN Analysis. Towards a Taxonomy of Selected Potential Anomalies

Anna Mroczek
Cracow University of Technology
ul. Warszawska 24, 31-155 Kraków, Poland
Email: amroczek@pk.edu.pl

Antoni Ligęza
AGH University of Science and Technology
al. Mickiewicza 30, 30-059 Kraków, Poland
Email: ligeza@agh.edu.pl

***Abstract*—Modeling based on a graphical notation understandable for different specialists has become very popular. Within the area of business processes, the most common one is the Business Process Modeling and Notation (BPMN). BPMN is aimed at all business users who design, analyze, manage and monitor business processes. Most papers in this area focus on making use of the possibilities that BPMN makes available, but there is lack of papers analyzing possible errors and ways of detecting and eliminating them. Specification of a BPMN diagram is relatively precise, but it is only a descriptive form presented at some abstract, graphical level. Hence, the main focus of this article is an attempt to analyze the topic of the anomalies which are likely to occur when modeling with use of BPMN.**

I. Introduction

CURRENTLY, the approach to modeling based on a graphical notations understandable for different specialists has become very popular. In the are of business processes the most common one is the Business Process Modeling and Notation (BPMN). BPMN is a business process modeling standard developed by Business Process Management Initiative. At present, BPMN is supported by the *Object Management Group* (OMG) because the two organizations merged in 2005.

In March 2011 the most recent specification of BPMN (BPMN 2.0) was released. The purpose of BPMN was to create a uniform notation of business processes that would be generally understandable — from professional process analysts, through managers to ordinary workers. According to [1], BPMN 'a standard Business Process Model and Notation (BPMN) will provide businesses with the capability of understanding their internal business procedures in a graphical notation and will give organizations the ability to communicate these procedures in a standard manner. Furthermore, the graphical notation will facilitate the understanding of the performance collaborations and business transactions between the organizations'. The main aims of BPMN include the following:

- process visualization which uses a graphical presentation of a business process. This form of visualization is much more effective than a textual representation;
- documentation through specification of process features;
- communication — provides a set of simple, commonly understandable notations.

BPMN is aimed at all business users, from the analysts, who create the initial process drafts, through the technical developers, whose responsibility it is to implement the technology performing those processes, and finally, to the business people, who will manage and monitor the aforementioned processes. The notation is clearly identified by various groups of experts, not only those connected with the IT industry. Yet, in spite of numerous endeavors, problems with unambiguous interpretation still exist. This fact stems from lack of a satisfactory BPMN interpreter. In fact as a consequence no semantics of BPMN components and connection is provided. Hence, various devices can interpret BPMN differently. The fact that there is no formal semantics may lead to misinterpretations and errors.

Most papers in the business process area focus on making use of the possibilities that BPMN brings, but there is lack of papers analyzing errors and ways of eliminating them. BPMN specification is precise but it is only a descriptive, graphical form. Hence, the subject of this article is an attempt to analyze the topic of the anomalies which are likely to occur in BPMN.

This paper is a kind of survey on anomalies in BPMN. An attempt has been made at presenting a taxonomy of possible problems, both of static, structural and dynamic nature. The research is based on literature analysis and some limited experience with BPMN models.

The article has been divided into five sections. The first one covers the basic elements of BPMN, the second presents an overview of the literature on anomalies. The consequent part touches potential misinterpretations and errors; it is based on examples. The last section contains the summary and conclusions.

II. ELEMENTS OF BPMN

BPMN model consists of simple diagrams made up of a limited set of graphical elements. Simplification of activity flows and processes is clearer for business users and developers. There are four main elements of BPMN, namely: Flow Objects, Connecting Objects, Swimlanes and Artifacts.

A. Flow Objects

Flow objects are the key elements describing BPMN. They consist of three core elements: events, activities and gateways [2].

An Event is represented by a circle and means something that happens (compared to an Activity, which is something

978-83-60810-58-3/$25.00

An Event is represented by a circle and means something that happens (compared to an Activity, which is something that has been done). The circular figures differ depending on the type of Event. Events may have an impact on a business process. An event can be an external or internal one. As long as they can influence the process being modeled, they should be modeled.

In general, there are three types of Events: Start, Intermediate and End. Start Event works like a trigger to a process. It is important for every process to have a Start Event to show the beginning of the business process. It allows readers to locate in the BPMN diagram where the process begins, and under what conditions. End event is used to indicate where the business process finishes. It presents the outcome of the process. Intermediate Event represents what happens in the gap between Start Event and End Event. It is responsible for driving a business flow based on the event it specifies.

An Activity is represented by a rounded-corner rectangle and describes a type of work that has to be completed within a business process. There are two kinds of Activities: Tasks and Sub-processes. Task means a single unit of work which is not or cannot by divided in the next stage of business processes specification; in certain sense a task is of an *atomic* nature. On the other hand, sub-process is used for complex work which can be divided into smaller units. It is applied in order to cover or uncover additional specification levels of business processes.

Gateways are elements used to monitor the way in which some business process flows interact with the others. A Gateway is represented by a diamond shape. Some of the typical types of gateways are the following ones:

- **Data-Based Exclusive Gateway** — it is used to control process flow based on given process data.
- **Inclusive Gateway** can be used to create parallel paths. (Fig. 1).

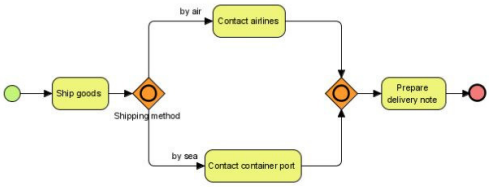


Fig. 1. Inclusive gateway

- **Parallel Gateway** — it is used to model the execution of parallel flows without the need of checking any conditions, all outgoing flows must be executed at the same time (Fig. 2).
- **Event-Based Gateway** — it is used to model alternative paths that are based on events (Fig. 3).

B. Connecting Objects

Flow objects are connected to each other using Connecting objects, which are of three types: sequences, messages, and associations [2]. Sequence Flow is used to show the order

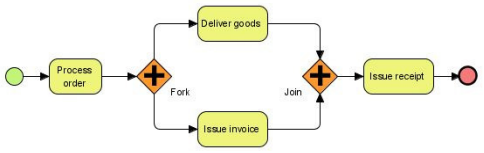


Fig. 2. Parallel gateway

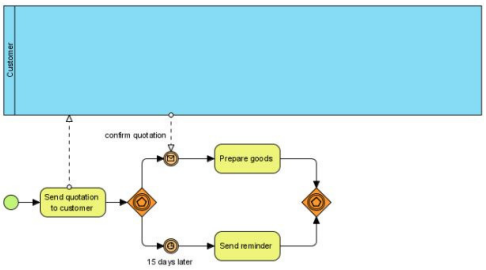


Fig. 3. Event-based gateway

in which particular activities will be performed in a process. Message Flow is used to show the flow of messages between two process participants entitled to send and receive them. Association is used to link information and artifacts to activities, events, gateways and flows.

C. Swimlanes

BPMN usually uses the concept of swimlanes in order to demonstrate what business function a particular activity is connected with or what system executes it. There are two types of swimlanes objects: lanes (sub-partition of pools) and pools (represent participants in a business process) [2].

Inside a pool, there are flow elements. It acts as a graphical container for partitioning a set of Activities from other Pools. Lanes can be used to represent specific objects or roles engaged in a process. They are used to organize and categorize activities in a pool, according to the function and role. They are represented by a rectangle extending either vertically or horizontally along the length of the pool. A lane contains flow objects, connecting objects and artifacts.

D. Artifacts

Artifacts are diagram elements used to display additional information relative to the process. They enable programmers to include more information in a model. In this way, the model becomes clearer. BPMN does not restrict the number of artifacts, though currently three have been defined [2]:

- *Data objects* are a mechanism whose aim is to show how data is prerequisite or result from activities. They are connected to activities though Associations.
- *Groups* can be used for analysis or documentation objectives but they do not affect the sequence flow.
- *Annotations* are a mechanism used in modeling to provide additional text information for the reader of BPD (Business Process Diagram).

III. THERMOSTAT: AN ILLUSTRATIVE EXAMPLE

[3].

In order to provide intuitions, the theoretical considerations will be illustrated with a simple example process. The process goal is to establish the so-called *set-point* temperature for a thermostat system [4]. The selection of the particular value depends on the season, whether it is a working day or not, and the time of the day. A BPMN diagram of the process is specified in Figure 4.

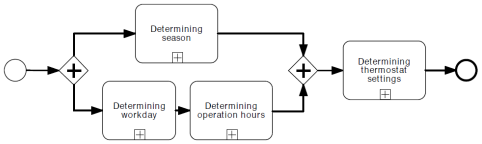


Fig. 4. An example BPMN diagram top-level specification of the thermostat system

After start, the process is split into two independent paths of activities. The upper path is aimed at determining the current season [1] *aSE*; it can take one of the values *sum, aut, win, spr*; the detailed specification is provided with rules 7-10. A visual specification of this activity with an appropriate set of rules is shown in Fig. 5.

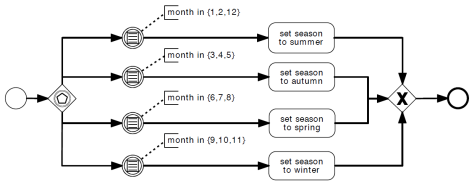


Fig. 5. An example BPMN diagram detailed specification a BPMN task

The lower path determines whether the day (*aDD*) is a workday (*aTD=wd*) or a weekend day (*aTD=wk*), both specifying the value of today (*aTD*); specification provided with rules 1 and 2, and then, taking into account the current time (*aTM*), whether the operation (*aOP*) is during business hours (*aOP=dbh*) or not (*aOP=ndbh*); the specification is provided with rules 3-6. This is illustrated with Fig. 5 and Fig. 6.

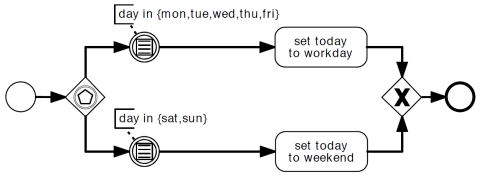


Fig. 6. An example BPMN diagram detailed specification of determining the day task

The whole process is formally specified with the following eighteen inference rules.

[1]For technical reasons all attribute names used in this example start with lower-case 'a'.

Rule 1:$\in \{,,,\} \longrightarrow = wd.$
Rule 2:$\in \{,\} \longrightarrow = wk.$
Rule 3:$= wd \wedge \in (9,17) \longrightarrow = dbh.$
Rule 4:$= wd \wedge \in (0,8) \longrightarrow = ndbh.$
Rule 5:$= wd \wedge \in (18,24) \longrightarrow = ndbh.$
Rule 6:$= wk \longrightarrow = ndbh.$
Rule 7:$\in \{,,\} \longrightarrow = sum.$
Rule 8:$\in \{,,\} \longrightarrow = aut.$
Rule 9:$\in \{,,\} \longrightarrow = win.$
Rule 10:$\in \{,,\} \longrightarrow = spr.$
Rule 11:$= spr \wedge = dbh \longrightarrow = 20.$
Rule 12:$= spr \wedge = ndbh \longrightarrow = 15.$
Rule 13:$= sum \wedge = dbh \longrightarrow = 24.$
Rule 14:$= sum \wedge = ndbh \longrightarrow = 17.$
Rule 15:$= aut \wedge = dbh \longrightarrow = 20.$
Rule 16:$= aut \wedge = ndbh \longrightarrow = 16.$
Rule 17:$= win \wedge = dbh \longrightarrow = 18.$
Rule 18:$= win \wedge = ndbh \longrightarrow = 14.$

Let us briefly explain these rules. The first two rules define if we have today (*aTD*) a workday (*wd*) or a weekend day (*wk*). Rules 3-6 define if the operation hours (*aOP*) are during business hours (*dbh*) or not during business hours (*ndbh*); they take into account the workday/weekend condition and the current time (*hour*). Rules 7-10 define the season (*aSE*) is summer (*sum*), autumn (*aut*), winter (*win*) or spring (*spr*).

Finally, the results are merged together, and the final activity consists in determining the thermostat settings (*aTHS*) for particular season (*aSE*) and time (*aTM*) (the specification is provided with rules 11-18).

This is illustrated with Fig. 7.

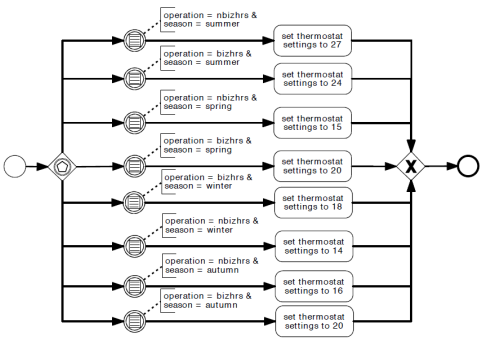


Fig. 7. An example BPMN diagram — detailed specification of the final thermostat setting task

Even in this simple example, answers to the following important questions are not obvious:

1) *Data flow correctness*: Is any of the four tasks/activities specified in a correct way? Will each task end with producing desired output for any admissible input data?
2) *Split consistency*: Will the workflow possibly explore all the paths after a split? Will it always explore at least one?
3) *Merge consistency*: Will it be always possible to merge knowledge coming from different sources at the merge node?
4) *Termination/completeness*: Does the specification assure that the system will always terminate producing some

temperature specification for *any* admissible input data?
5) *Determinism*: Will the output setting be determined in a unique way?

Note that we do not ask about *correctness* of the result; in fact, the rules embedded into a BPMN diagram provide a kind of *executable specification*, so there is no reference point to claim that final output is correct or not.

IV. Related work

There is a possibility of defining incoherent business logic specification and its interpretation. Even in basic processes anomalies are observed [5]. An improvement is required in the mechanism which provides cohesion in detecting anomalies in business processes [6]. Anomalies have been defined in numerous papers, yet a uniform definition was presented in [7] IEEE standard classification for Software Anomalies and it says: *"Each condition different from the expected is an anomaly"*.

In business logic an anomaly can be considered as every negative influence on modeling and models. There is a special kind of anomaly — a defect, which blocks the correct and efficient flow of objects completely.

A taxonomy of anomalies was created on the basis of literature. It concerns the flow control, bases and verification rules of data as well as flow accuracy. The taxonomy can make up a base for classification and research on anomaly possibilities.

The anomaly problem in BPMN is based on searching business logic for particular patterns. In [8], typical controls for anti-patterns are searched for by using a query language for BPMN. It is confirmed by deadlocks or livelock patterns which are used improperly. A similar thing happens in [9] where typical gateway constellations leading to problematic situations in the flow work diagram are presented. A comparable situation occurs in [10] as well, where an 'anomaly pattern' is used. This approach is based on detecting anti-patterns in the data flow. The whole thing is based on time logic using a real model control. By making use of different tools, position [11] is focused on various anomalies which stem from formalism or inadequacy of the tools. Yet another approach is a conception based on UML diagrams in development stages [12].

Control flow anomalies concern problems connected with flow control and gateways conditions [12]. In [13] , a problem was presented of control over many semantically identical connections between two work flow elements. This multiplicity complicates changes in the work flow, which is not desired.

Another element of flow control are gateways placed in the modeling center. It was stated in [14] that XOR-gateways with undefined gateway conditions can cause practical problems or even be a reason of an error. A similar thing happens when XOR-gateways conditions do not exclude each other and partly or fully overlap. What happens in flow control in case of lack of synchronization is multiple flow execution. For example, branches and some loop instruction cause such an anomaly [15].

Another situation is a flow deadlock. It is a situation in which the work flow is stopped in the current position of the path and cannot be accomplished. Another lock of flow is known as livelock. In [8] it is called an 'infinite loop'. Flow livelock keeps the operating work flow system in an infinite loop. The reason are bad modeling conditions, which prevent leaving the loop. Both cases — deadlock and livelock are described in [8], [15].

Rule-based anomalies are described in numerous papers [16], [17], [18], [19]. They involve mainly two problems connected with base rules. First, Rule-base Consistency are anomalies concerning coherence. Problems result from the set of rules, which have determined conditions but different outcomes at the same time. Rule-base livelocks, also called *circular rules* [17]. Rule-base livelocks and rule-base deadlocks describe a problem with creation rules, which are dependent on one another although they should not. This type of anomaly suggests that rule-base does not encompass the basic context in which it is used. Coverage anomalies concern the rules in which conditions can be fulfilled by the base context but conclusions are modeled in such a way that no effect will ever be seen. Another type of data flow anomaly is based on [20]. Such anomalies are influenced by those data elements which can be processed by workflow activities.

V. Anomalies in business processes

There are two kinds of business process anomalies which can occur while process modeling, namely [19]: Syntactic anomalies and Structural anomalies.

A. Syntactic Anomalies in Business Process

Analysis of Syntactic Business Process anomalies is important while designing a business process model. In this section examples of syntactic anomalies in business processes will be presented. A division into three groups has been made: **Incorrect usage of Flow Objects**, **Incorrect usage of Connecting Object** and **Incorrect usage of Swimlanes**.

B. Incorrect usage of Flow Object

The anomaly of this type result from improper use of the Event, Activity and Gateway.

Incorrect usage of Activities: Invalid use of Start Event or End Event. The BPMN specification defines the start and end events as optional. However, their usage is highly recommended, since each process starts and ends somewhere. Without explicitly using start and end events, a regular BPMN process might look the process in Fig. 8. This modeling

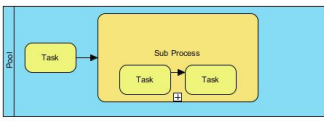


Fig. 8. Implicit process events

approach is undesirable and could lead to misinterpretations.

Depending on application, three anomalies can be distinguished. These are: Activities without Activation, Activities without Termination and Invalid use of Receive Task. Activities without Termination and Invalid use of Receive Task.

- **Activities without activation** If an activity is situated on a path that has no start, then this is an activity without activation. Even if a start of an event is used.
- **Activities without Termination**. An activity without termination happens when the activity cannot be brought to an end. Even if End Event is used.
- **Invalid use of receive task**. Receive Task Element is designed to wait for incoming messages from outside users in a business process.

Invalid use of Gateway. There are two groups of anomalies: invalid use of Data-Based XOR Gateway and invalid use of Event-Based XOR Gateway.

- **Invalid use of Data-Based XOR Gateway**. A data-based XOR Gateway relies on the arrival of a data token that has traversed the Process Flow. Data-based XOR Gateway must be date-based objects.
- **Event-Based XOR Gateway**. According to BPMN, the event-based gateway cannot be used as a merge gateway. It can only be used as a decision type gateway (multiple outputs).

Incorrect usage of Connecting Object. Anomalies concerning connecting objects stem from incorrect usage of their elements, that is message flow and sequence flow. As far as incorrect utilization of connecting objects is concerned, a few anomalies can be differentiated: the ones concerning incoming sequence flows, outgoing sequence flows, invalid use of conditional sequence flow. In this case there are two possible irregularities regarding the invalid use of a pool or lane [26].

C. Invalid use of Pool

When modeling multiple pools, a common mistake is that activities in a Pool are not connected with sequence flows. It is incorrect to use multiple pools as a single process and incorrectly interprets messages flows as way of indicating a sequence of activities (Fig. 9).

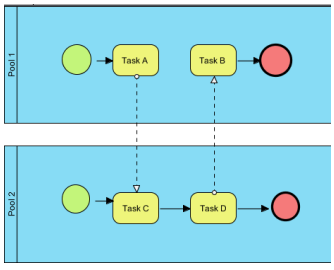


Fig. 9. Missing sequence flow

Another common problem when modeling multiple pools is the use of a set of pools as a single pool with multiple lanes. The end result will be an incorrect model (Fig. 10) that represents a single process that spreads over the boundaries of the pool.

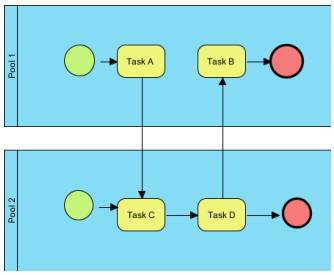


Fig. 10. A Sequence flow May not cross pools boundaries

D. Invalid use of Lane

Improper use of lane as a pool, thereby representing individual processes within separated lanes. This is wrong, because a lane is just a activity-classifying mechanism (Fig. 11).

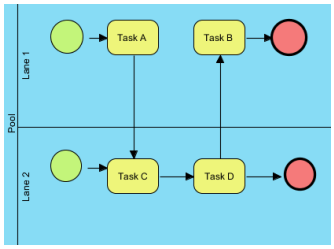


Fig. 11. Two Lances are used as two Pools

E. Structural Anomalies

Structural anomalies have been described in the literature [21], [22], [23], [24]; they are classified as four types: **Deadlock**, **Lack of synchronization**, **Dead Activity** and **Infinite Loop**.

Note that in fact all the above anomalies correspond to *wrong dynamic behavior*; all of them occur during execution of the process.

A process is sound [25] if and only if it is free of two control-flow errors: the deadlock and the lack of synchronization. First, deadlocks are blockings in the process model, which occur when gateways are used incorrectly. In this case, the links in the process where gateways were installed should be checked. Deadlocks occur when an exclusive gateway was picked for linking and this linkage was combined again with a parallel gateway. They may arise from added intermediate events or multiple exclusive start events, which should be checked again. There are two types of deadlocks: deterministic deadlock (Fig. 12) and non-deterministic deadlock (Fig. 13).

A deadlock is a reachable state of the process that contains a token on some Sequence Flow that cannot be removed in any possible future. A lack of synchronization (Fig. 14) is a reachable state of the process where there is more than one token on some Sequence Flow. To characterize the lack of synchronization, we follow the intuition that potentially

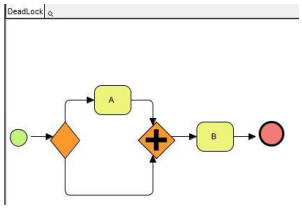


Fig. 12. Deterministic deadlock

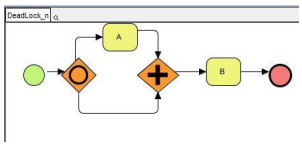


Fig. 13. Non-deterministic deadlock

concurrent paths, paths starting with an IOR-split or an AND-split, should not be joined by XOR-join. In the following, we formalize this characterization and show that such structure always leads to lack of synchronization in deadlocks free acyclic workflow graphs [24]. While *Dead Activities* are activities which will never be executed. A last type of anomaly is Infinite Loop [27] , also called *closed loop*. A closed loop is a cycle without any split. Tokens that enter a closed loop are forever lost to the rest of the workflow. In our model, this leads to a deadlock, because each token entering the closed loop will have a synchronization copy of itself placed on the incoming edge of the initial join that loops back from the cycle. It is hard to imagine a sensible real-world example that contains a closed loop (the BPMN standard document admits this). Banning closed loops from workflows is thus not a serious restriction, especially since infinitely looping cycles are still possible as long as they are not closed [25].

VI. Conclusion

BPMN is a popular business process modeling language. The ability of using it is very important in the modeling stage. Yet, despite its advantages, the problem of effective anomaly detection still remains. There is a lack of a proper tool that would automate the process of detecting anomalies in business process modeling.

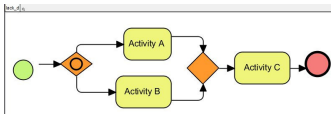


Fig. 14. Lack of synchronization

References

[1] The BPMN page: http://www.bpmn.org
[2] The BPMN documentation page: [2]http://www.omg.org/bpmn/Documents/Introduction_to_BPMN.pdf
[3] A. Ligęza,'BPMN a logical model and property analysis. Decision Making in Manufacturing and Services', vol. 5 2011, pp. 57-67, http://dx.doi.org/10.1002/widm.11
[4] M. Negnevitsky, 'Artificial Intelligence. A Guide to Intelligent Systems', Pearson Education Paperback, 2002.
[5] J. Mendling, E. Verbee, B.van Dongen, W.van der Aalst and G.Neumann, 'Detection and Prediction of Errors in EPCs of the SAP Reference Model', *Data & Knowledge Engineering* vol. 64(1), 2008, pp. 312-329.
[6] A. Hallerbach, T. Bauer, R. Manfred, 'Capturing Variability in Business Process Models: The Provop Approach', *Journal of Software Maintenance and Evolution: Research and Practice*, in BPM 2008, vol. 22, 2010, pp. 519-546.
[7] The IEEE Computer Society, 'IEEE Standard Classification for Software Anomalies", 2010 https://standards.ieee.org/findstds/standard/1044-2009.html
[8] R. Laue, A. Awad, 'Visualization of Business Process Modeling Anti Patterns, Visual Formalisms for Patterns', *Electronic Communications of the EASST*, vol. 25, 2009.
[9] S. Kuhne, H. Kern, V. Gruhn, R. Laue, 'Business process modeling with continuous validation', *Journal of Software Maintenance and Evolution: Research and Practice*, BPM 2008 Vol. 22, 2010, pp. 547-566.
[10] N. Sidorova, N. Trcka, W. M. P. van der Aalst, 'Data-Flow Anti-patterns: Discovering Data-Flow Errors in Workflows', Springer, 2009, pp.425-439.
[11] N. Lohmann, K. Wolf, 'How to Implement a Theory of Correctness in the Area of Business Processes and Services', BPM 2010, Springer, 2010, pp. 61-77, dx.doi.org/10.1007/978-3-642-15618-2_7
[12] A. Stephen, 'Process Modeling Notations and Workflow Patterns', White, IBM Corp., United States, Nov. 2009. The BPMN documentation page: http://www.omg.org/bpmn/Documents/Notations_and_Workflow_Patterns.pdf
[13] L. Olkhovich, 'Semi-Automatic Business Process Performance Optimization Based On Redundant Control Flow Detection', AICT-ICIW '06,2006, pp. 146-146.
[14] The OMG documentation page: 'Documents Associated With Business Process Model And Notation (BPMN) Version 2.0', http://www.omg.org/spec/BPMN/2.0/, Jan. 2011.
[15] R. Liu, A. Kumar, 'An Analysis and Taxonomy of Unstructured Workflow', BPM 2005, 2005, pp. 268-284.
[16] X. Desheng, X. Kejian, Z. Dezheng, Z. Huangsheng, 'Model Checking the Inconsistency and Circularity in Rule-Based Expert Systems', *Computer and Information Science*, 2009, pp. 12-17.
[17] A. K. Zaidi, A. H. Levis, 'Validation and verification of decision making rules', *Automatica*, vol. 33, 1997, pp. 155-169.
[18] M. Do ohring, S. Heublein, 'Anomalies in Rule-Adapted Workflows - A Taxonomy and Solutions for vBPMN', *Software Maintenance and Reengineering* (CSMR), 2012, pp. 117 - 126, http://doi.ieeecomputersociety.org/10.1109/CSMR.2012.22
[19] A. Ligęza, G. J. Nalepa, 'A study of methodological issues in design and development of rule-based systems: proposal of a new approach', *Data Mining and Knowledge Discovery*, 2011, pp. 117-137, http://dx.doi.org/10.1002/widm.11
[20] A. Awad, G. Decker, N. Lohmann, 'Diagnosing and Repairing Data Anomalies in Process Models', BPM 2009, Springer, 2009, pp. 5-16, http://dx.doi.org/10.1007/978-3-642-12186-9_2
[21] G. Kim, J. H. Lee, J. H. Son, 'Classification and Analyses of Business Process Anomalies', *Communication Software and Networks*, ICCSN, 2009, pp. 433-437.
[22] H. Lin, Z. Zhao, Z. Chen, 'A novel graph reduction algorithm to identify structural conflicts', HICSS-35 Proceedings of the 35th Hawaii International Conference on System Sciences, 2002, pp. 289.
[23] H. Ling, Z. J. Bo, 'Research on workflow process structure verification', *e-Business Engineering*, ICEBE, 2005, pp. 158-165.
[24] W. M. P. van der Aalst, A. Hirnschall, H. Verbeek, 'An Alternative Way to Analyze Workflow Graphs', 14th International Conference, CAiSE 2002, Springer, 2002, pp. 535-552, http://dx.doi.org/10.1007/3-540-47961-9_37
[25] E. Boerger, O. Soerensen, B. Thalheim, 'On defining the behavior of OR-joins in business process models', *Journal of Universal Computer Science*, vol. 14, no. 1, 2009, pp. 3-32.
[26] The page: http://blog.goodelearning.com/
[27] The BPMI page, 'Business Process Modeling Notation Specification', http://www.omg.org/technology/documents/speccatalog.htm, 2006.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1103–1110
DOI: 10.15439/2014F68
ACSIS, Vol. 2

Towards an Understanding Business Intelligence. A Dynamic Capability-Based Framework for Business Intelligence

Celina M. Olszak
Katowice University of Economics
ul. Bogucicka 3b,
40-287 Katowice, Poland
Email: celina.olszak@ue.katowice.pl

Abstract—Although Business Intelligence (BI) is one of the most essential technologies to be purchased, the implementation of many BI applications fails. The reasons for this failure are not clear and still not well investigated. Resource-based View (RBV) and dynamic capability theory could help to overcome this gap and to provide an appropriate theoretical basis for future research in BI area. It is considered that BI capabilities may be critical functionalities that help organizations to improve their performance and adopt to environmental change. The research objectives for this study are: (1) conceptualization and discussion on BI dynamic capability, (2) building the comprehensive framework of BI capabilities. In order to address these objectives, the remainder of the paper is structured as follows: The first sections provide the theoretical foundations of BI, RBV and dynamic capability theory. Next, the BI capability was conceptualized. Finally, a model of BI as a dynamic capability, was proposed. The study was based mainly on: (1) a critical analysis of literature, (2) an observation of different BI initiatives undertaken in various organizations, as well as on (3) interviews with managers and experts in BI. The results of this study can be used by IT and business leaders as they plan and develop BI capabilities in their organizations.

I. Introduction

In order to gain competitive advantage, many organizations decide to use Business Intelligence (BI) systems. It is believed that BI enables organizations to better understand not only internal business processes, but also the competitive environment through the systematic acquisition, collation, analysis, interpretation and exploitation of information. BI allows for the identification of the opportunities and threats, which may occur on the market, while cooperating with customers, suppliers and competitors [1], [2], [3], [4], [5], [6], [7], [8].

It is worth mentioning, that in 2010, BI topped the list of the most important application and technology development in an annual survey of IT executives [9]. According to Gartner research and Forrester the BI market will grow from $8,5 billion in 2008 to $12 billion in 2014 [10]. Although BI is one of the most essential technologies to be purchased, many BI applications fail or the organizations do not achieve the appropriate benefits [11], [12], [13], [14], [15]. The reasons for this failure are not clear and still not well investigated. Resource-based View (RBV) and dynamic capability theory could help to overcome this gap and to provide an appropriate theoretical basis for future research in BI area.

This paper seeks to throw more light on the concept of BI by using a dynamic capabilities perspective. I consider that BI capabilities may be critical functionalities that help organizations to improve their performance and adopt to environmental change.

The research question I ask in this paper is: what new light contribute RBV and dynamics capabilities to BI area. Consequently, the research objectives for this study are: (1) conceptualization and discussion on BI dynamic capability (2) building a comprehensive framework of dynamic capabilities for BI.

In order to address these objectives, the remainder of the paper is structured as follows: The first sections provide the theoretical foundations of BI, RBV and dynamic capability theory. Next, the BI capability was conceptualized. Finally, a comprehensive framework of BI as a dynamic capability was proposed. The study was based mainly on: (1) a critical analysis of literature, (2) an observation of different BI initiatives undertaken in various organizations, as well as on (3) interviews with managers and experts in BI. The results of this study can be used by IT and business leaders as they plan and develop BI capabilities in their organizations.

II. Background on Business Intelligence and Dynamic Capabilities

A. Business Intelligence

Business Intelligence has become the significant research area in the domain of management information systems in the last years. The roots of BI originate from decision support systems, which first emerged in the early 1970s when managers used computer applications to model business decisions. Over the years, other applications, such as executive information systems (EIS), online analytical processing (OLAP), data warehousing, and data mining became important [5],[6], [7]. Today BI is compared to "an umbrella" that is commonly used to describe the technologies, applications, and processes for gathering, storing, accessing and analyzing data to help users to make better decisions [1], [16].

BI is comprised of both technical and organizational elements [17], [18], [19], [20], [21]. From technical point of view BI is an integrated set of tools, technologies and software products that are used to collect heterogenic data

978-83-60810-58-3/$25.00

from dispersed sources and then to integrate and analyze data to make them commonly available. The key BI technologies include: data warehousing, data mining and OLAP [22]. They are often called BI.1.0.

In the last years, new techniques, such as: web mining, opinion mining techniques, mobile mining techniques and semantic processing are applied in building BI systems They are focused on processing of semi-structured or un-structured data that originate mainly from Internet and social media. BI addressed for acquiring and processing data from web resources are named BI 2.0. In turn, BI 3.0 are responsible for collecting and analyzing data from various mobile devices [7], [24].

From organizational perspective, BI means a holistic and sophisticated approach to cross-organizational decision support [8], [11], [25]. Negash and Gray [3] argue that BI is responsible for transcription of data into information and knowledge. Also, it creates some environment for effective decision-making, business processes, strategic thinking, acting in organizations and taking the competitive advantage [26], [27], [28], [29], [30]. Many authors highlight that BI is predisposed to support decision-making on all levels of management [1], [3], [8], [31], [32]. On the strategic level, with the help of BI it is possible to set objectives precisely and follow the realization of such established objectives. BI allows for performing different comparative reports, e.g. on historical results, profitability of particular offers, effectiveness of distribution channels or forecasting future results on the basis of some assumptions. On the tactical level BI may provide some basis for decision-making within marketing, sales, finance, capital management etc. BI allows for optimizing future actions and modifying organizational, financial or technological aspects of company performance appropriately in order to help enterprises to realize their strategic objectives more effectively. In turn, on the operational level, BI systems are used to perform ad hoc analyses and answer questions related to departments' ongoing operations, up-to-date financial standing, sales and co-operation with suppliers, customers [22].

It is indicated that BI facilitates the realization of business objectives through reporting of data to analyse trends, creating predictive models for forecasting and optimizing process for enhanced performance. The value of BI systems for business is predominantly expressed in the fact that such systems cast some light on information that may serve as the basis for carrying out fundamental changes in a particular enterprise. It is stated that BI has become the critical component for the success of the contemporary organization [2], [15], [33], [34]. Wells [18] argues that BI is the "capability of an organization to explain, plan, predict, solve problems, think in an abstract way, understand, invent, and learn in order to increase organizational knowledge, provide information for the decision-making process, enable effective actions, and support establishing and achieving business goals".

It should be pointed that although, BI applications have become the most essential technologies to be purchased in the last years, the BI success is still questionable. It is reported that the practical benefits from BI are often unclear and some organizations fail completely in their BI approach or they do not achieve the appropriate benefits [11], [12], [13], [14], [15]. It is said that about 60 to 70% of business intelligence applications fail due to the technology, organizational, cultural and infrastructure issues [35], [36], [37], [38]. It is reported that the most important elements that decide on BI success in the organizations include: quality of data and used technologies, skills, sponsorship, alignment between BI and business, and BI use [35]. Other elements concern: organizational culture, information requirements, and politics. According to Olszak and Ziemba [38] the biggest barriers that the organizations encounter during the implementation of BI systems have a business and organizational character. Among the business barriers, the most frequently mentioned are: the lack of well defined business problem, not determining the expectation of BI and the lack of relations between business and BI vision system. Whereas as the key organizational barriers the enterprises enumerate: the lack of manager's supporting, the lack of knowledge about the BI system and its capabilities, exceeded the BI implementation budget, ineffective BI project management and complicated BI project, the lack of user training and support.

B. Resource-based View

RBV argues that about the success of organization's strategy decide the configuration of its resources and capabilities that are the basis to build key competences. Acquiring, configuration, reconfiguration and developing of available resources is critical factor for taking the competitive advantage and creating the value [39], [40], [41].

RBV was put forward by Wernerfelt [42] and subsequently popularized by Barney's work [39]. Many authors made significant contribution to its conceptual development [43], [44], [45].

According to RBV in order to provide sustainable competitive advantage, resources should be (VRIN): Valuable (enable an organization to implement a value-creating strategy), Rare (are in short supply), Inimitable (cannot be perfectly duplicated by rivals) and Non-substitutable (cannot be countered by a competitor with a substitute). In an extended approach of RBV resources imply intangible categories including organizational, human and networks [46]. This knowledge-based resource approach of RBV encourages organizations to obtain, access, and maintain intangible endowments because these resources are the ways in which firms combine and transform tangible input resources and assets [47]. It is reported that BI technology, as well others ICT, do not satisfy the VRIN criteria [48]. However, they may be synergistically combined with existing organizational resources, to form other VRIN resources [41], [49].

C. Dynamic capabilities theory

The concept of dynamic capabilities is rooted in the RBV of competitive advantage. RBV defines capability as the ability of a bundle of resources to perform an activity. It is a

way of combining assets, people and processes to transform inputs into output [50].

Teece et al. [50] define capabilities as "the key role of strategic management in appropriately adapting, integrating, and reconfiguring internal and external organizational skills, resources, and functional competences to match the requirements of a changing environment". Many authors, explaining the topic of capabilities, highlight some differences between competency, capability and capacity [51]. Competence is the quality or state of being functionally adequate or having sufficient knowledge, strength and skill. While capability is a feature, faculty or process that can be developed or improved. Capability is a collaborative process that can be deployed and through which individual competences can be applied and exploited. Capacity is the power to hold, receive or accommodate.

Hamel and Prahalad [52] coined the term core competence to distinguish those capabilities fundamental to a firm's performance and strategy. Core competencies are the activities that the firm performs especially well compared to competitors and through which the firm adds value to its goods and services over a long period of time. They emerge over time through an organizational process of accumulating and learning how to deploy organizational resources and capabilities.

The RBV conceptualizes organizational resources as static, neglecting changes due to turbulent environments. A stable resource configuration can not guarantee long-term competitive advantage as organizations have to adopt this configuration to the market environment [50]. This argument is even stronger in dynamic market environments where there is "rapid change in technology and market forces and feedback effects on firms [53]. Dynamic capabilities were conceptualized in response to this criticism [41], [44].

Teece et al. [50] identify dynamic capabilities as "the firm's ability to integrate, build, and reconfigure internal and external competences to address rapidly changing environments". The notion of dynamic capabilities was subsequently refined and expanded [44], [45], [54]. Zollo and Winter [45] also distinguish dynamic capabilities from operational or ordinary capabilities. Operational capabilities enable firms to perform their every day living, "and while dynamic (as all processes are), they are used to maintain the status quo" [54]. By contrast, dynamic capabilities are those that enable a firm to constantly renew its operational capabilities and therefore achieve long-term competitive advantage.

It is worth noting, that RBV has been used extensively in IS (Information Systems) research to explain how IT (Information Technology) assets provide value and sustainable competitive advantage to organizations [41]. Some studies found a direct link between IT assets and value but most found that IS capabilities and the interaction of IT assets with other organizational resources, lead to business value [40], [41]. IS capabilities are created through combining IT assets with other resources including people, routines and processes. IS capabilities develop and mature over time as organizational learn [43]. Dynamic capabilities are the high-order capabilities and thus can be disaggregated into different capacities, such as the capacity for improving quality, the capacity for managing human resources and the capacity for utilizing technologies [55].

D. Conceptualization of dynamic Business Intelligence capabilities

Drawn from the concept of dynamic capabilities, BI capability may be defined as IT-enabled, analytical dynamic capability for improving decision making and firm's performance [55]. It is a specific and important type of IS capabilities. Different organizational characteristics and strategic goals may also require using different BI capabilities. According to Gartner Group BI capabilities relate to information access and analysis to decision-making style within an organization [11]. Isik, Jones and Sidorova [11] delineate information access and analysis capabilities and relate them to the overall BI success. Davenport and Harris [6] state that analytical capability is a key element of strategy for the business. Wixom, Watson and Werner [5] argue that BI capability is "a journey over long periods of time during which foundational competencies are developed".

According to Teece et al. [50] dynamic capabilities can be distinguished into three classes of activities including sensing, seizing, and transformation. In the context of Business Process Management [53] and also of BI, sensing refers mainly to identification of the need to change an organization's business processes, relations with customers and suppliers. Seizing means the exploration and selection of opportunities for change. Transformation concerns socio-technically implementation of changed business processes in the organization. Some authors argue that BI capabilities are critical functionalities of BI that help an organization to improve its adoption to change as well as to improve its performance [5], [11].

Organizations may develop two activities in order to build BI capability. The former concerns the widely understood data exploration, the latter, data exploitation [56]. Data exploration enables organization to overcome the bounder of actual knowledge and its capabilities. This may refer to new technical capabilities, market experiences and new relations with the environment. Also, the exploration is a conscious searching of new knowledge sources, enriching of existing resources, adoption of new behavioral orientations and acquisition of new competencies. It can be achieved through: advances data mining, text mining, web mining, intelligent agents, and search based application. In turn, data exploitation concerns the using of existing knowledge bases. It is limited to actual resources and refers to their detail analysis.

Davenport and Harris [6] distinguish five stages of analytical capability called: analytically impaired, localized analytics, analytical aspiration, analytical companies, and analytical competitors. The first stage means that "organizations have some desire to become more analytical, but thus far they lack both the will and the skill to do so". They face some substantial barriers – both human and

technical. They may also lack the hardware, software and skills to do substantial analysis. The second stage "localized analytics " is characterized by reporting with pockets of analytical activity. The organizations undertake the first analytical activities, but they have no intention of competing on it. BI activities produce economic benefits but not enough to affect the company competitive strategy. The third stage called "analytical aspirations" is triggered when BI activities gain executive sponsorship. The organizations build the plan of using BI. The primary focus in "analytical companies" stage is building word-class analytical capabilities at the enterprise level. The organizations implement the plan developed in previous stage, making considerable progress toward building the sponsorship, culture, skills, strategic insights, data and technology needed for analytical competition. At the last stage, analytics moves from being a very important capability for an organizations to the key to its strategy and competitive advantage. Executive managers trust in BI and all users are highly educated in BI.

For the purpose of this paper it is assumed that dynamic BI capability is the ability of an organization to integrate, build and reconfigure the information resources, as well as business processes to address rapidly changing environments.

III. RESEARCH METHODOLOGY

An interpretative philosophy and an inductive qualitative approach were applied to build a comprehensive, dynamic BI capabilities framework. The theories (from IS and management literature) and studies developed mainly by Davenport and Harris [6], Wixom, Watson, and Werner [5], Cosic, Shankes, and Maynard [41] were adopted and used to create the dynamic BI capabilities framework.

BI is regarded as an applied discipline and therefore practitioner, viewpoints and opinions were considered of high importance. Therefore, I have used the results from the survey that was conducted in 2012 among 20 purposefully selected firms (in Poland) that are considered to be advanced in BI [24]. They represented the service sector: telecommunications, consulting, banking, insurance, and marketing agencies. Interviews were held with executives, senior members of staff and ICT specialists. Interviewees were selected on their involvement in BI or on their ability to offer an insight based on experience in BI and related decision support systems. The research was of qualitative nature and used as a research technique of an in-depth interview. Types of core interviews questions relevant to this paper have included among others: (1) Does your organization have a defined BI strategy?, (2) Does your organization have defined business processes?, (3) Are you skilled enough in order to take advantage of BI systems?, (4) Are you motivated to use BI (how)?, (5) Do you use BI for analyzing customers, suppliers, competitors and other business partners?, (6) What kind of BI software do you use?, (7) Describe some successes/failures from using BI. This methodology is appropriate for the explorative objectives of this research as it aimed to build dynamic BI capabilities framework.

IV. FINDINGS AND DISCUSSION

Figure 1 provides an initial framework for dynamic BI capabilities. It includes six capabilities areas like: governance, culture, technology, people, processes, and change management & creativity. So far, these areas were presented separately and were used for different aims and tasks. In this study, I integrate them into one comprehensive model for dynamic BI capabilities. Below, I present the arguments for adopting them to create a dynamic BI capability.

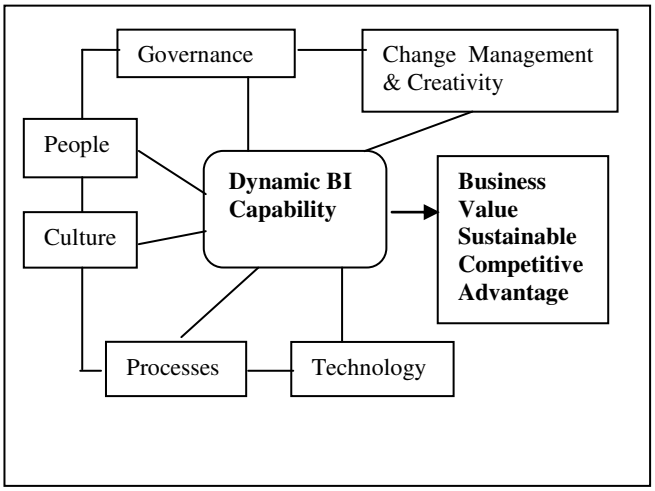


Fig. 1 Framework for BI capabilities

The governance is "the mechanism for managing the use of BI resources within an organizational and the assignment of BI initiatives with organizational objectives. It also involves continuously renewing BI resources and organizational capabilities in order to respond to changes in dynamic environments and mitigating resistance to change" [41].

Culture is often described as "personality of the organization" and comprises the assumptions, values, norms, and behavioral signs of organization/s members. They form over time and lead to systematic ways of gathering, analyzing and disseminating data. It influences the way decisions are made [41].

People refer to "all those individuals within organization who use BI as part of their job function. BI initiatives are considered to be knowledge intensive and require technical, business, managerial and entrepreneurial skills and knowledge" [41].

Technology refers "to the development and use of hardware, software and data within BI activities. It includes the management of an integrated and high quality data resources, the seamless integration of BI systems with other organizational information systems, the conversion of data into information through reporting and visualization systems and to use of more advanced statistical analysis tools to discover patterns, predict trends and optimize business process" [41].

Process constitutes of activities to gather, select, aggregate, analyze, and distribute information. Some of these activities are the responsibility of the BI staff, while others are the joint responsibility of the BI staff and the business units. Processes may be divided into categories: internal and external processes. The first group relates mainly to accounting, finance, manufacturing, and human resources. The second group concerns managing and responding to customer demand and supplier relationships [6].

Change management & creativity are organization's abilities to meet the requirements of dynamic environments. Organizations face rapid change like never before. Therefore, the ability to manage and adapt to organizational change is an essential ability required in the workplace today. Change management is an approach to transitioning individuals, teams, and organizations to a desired future state. BI requires permanent development and adaptation to new challenges and expectations of an organizations. While an organizational creativity is the firm's ability to generate new and useful ideas to address rapidly changing opportunities and threats by making timely and market-oriented decisions, and to frame breaking changes in its resource base.

The essentials analysis of the literature and the conducted interviews with various BI experts and managers allowed me to identify the detailed capabilities for each BI area (table 1). The number of the organizations that declared the possession of various BI capabilities and competences is presented in the last column of the table.

Table I.
BI CAPABILITIES AREA

BI Capabilities Area	Detailed BI capabilities	Number of organizations
Governance	Business vision and plan	10
	Business analysis planning and monitoring	11
	Strategic alignment BI and business strategy	5
	Decision rights (operational, tactical, strategic)	16
	BI solution assessment and validation	7
Culture	Executive leadership and support	6
	Flexibility and agility	8
	Establishing a fact-based and learning culture	7
Technology	Data management	16
	Systems integration and interaction with other systems	17
	Flexibility	17
	Reporting and visualization technology	20
	Advanced BI technology (OLAP, data warehousing, data mining, predictive analysis)	16
People	Securing and building technology skills	7
	Mathematical and statistical skills	5
	Organizational skills	7
	Organizational knowledge, knowledge sharing	5
	Managing analytical people	6
	Business interpersonal communication	12
	Entrepreneurship and innovation	5
	Trustworthiness	6
Process	Holistic overview business process/ knowledge processes	14/7
	Business process/knowledge/ modeling and orchestration	16/6
	Process redesign and integration	16
Change & Creativity	Monitoring of competitors, customers and current trends in the marketplace	9
	Introducing new business models oriented on change management, knowledge management and customer relationship management	7
	Generation of new and useful products, services, ideas, procedures, and processes	7

In the next step of my research, five detailed BI capabilities areas were mapped onto Davenport and Harris model. As a the result, a BI capabilities maturity matrix was created (table 2).

The analysis of the literature and the conducted survey allow me to state that the dynamic BI capabilities do not go hand in hand with the possibilities offered by BI technologies. Most organizations need to raise their "analytical erudition." Managers do not always know how such sources can be used in making decisions. The most of the organizations do not think creatively about the potential of data sources. They have a relatively high level of the technical competences. Unfortunately, they do not correspond with another BI capabilities (e.g., strategic alignment BI and business strategy, establishing a fact-based and learning culture, entrepreneurship and innovation, change management, and creativity).

BI is still treated as a technology or tool to acquire and analyze data and not as a trigger for making more effective decisions, improving business processes and business performance, as well as doing new business or creating new ideas and procedures. The organizations still underestimate the soft competences and skills needed for BI (e.g., culture-based on facts and knowledge, trust, human resources management, managing analytical/creativity people). Worried, that BI and business strategy are not aligned in many organizations.

Table II.
BI CAPABILITIES MATURITY MATRIX

BI capabilities area	Analytically impaired	Localized analytics	Analytical aspiration	Analytical companies	Analytical competitors
Governance	Lack of vision and plan	Businesses plans for limited departments	Integrated business strategy	Have an enterprise BI strategy	BI strategy oriented on customers, suppliers etc.
Culture	No flexibility and agility	Low support from senior executives	Users are encouraged to collect, process analyze and share information	Establishing a fact-based and learning culture, skill training in BI	Learning from customers, suppliers, communities of practice, social media
Technology	Missing/poor data, Unintegrated systems	Missing important data, Isolated BI efforts	Proliferation of BI tools	High- quality of data, integrated knowledge repositories	Enterprise-wide BI architecture largely implemented
People	Users do not know their own data requirements or how to use them	The users take the first BI initiatives	Users try to optimize the efficiency of individual departments by BI	Users have high BI capabilities, but often not aligned with right role	Users have capabilities and time to use BI
Processes	Users do not know business processes	Identification of basic business processes	Standardization of business processes, and building best practices in BI	Business process management based on facts	Broadly supported, process-oriented culture based on facts
Change & Creativity	Fear of change, no creativity	Risk management for selected business process, poor and limited creativity	Building the best practices for change management, individual and team creativity	Integrated risk management, team and organizational creativity	Cooperation with competition, organizational creativity, creative environment

In order to reach a comprehensive, dynamic BI capability, organizations should simultaneously build and developed a whole bundle of various BI capabilities. Undoubtedly, it is a long journey and developed over long periods of time. I think, that organizations should not start from building technical competences, structures (data bases, data warehouses etc.), without prior the implementation of knowledge-based organization, change management, and organizational creativity.

Concluding, I consider that organizations should simultaneously develop different BI capabilities in order to achieve high BI maturity. These capabilities may be focused on data exploration and data exploitation. As mentioned earlier, data exploration enables organization to overcome the bounder of actual knowledge and its capabilities. In contrast, data exploitation concerns the using of existing knowledge bases. It is limited to actual resources and refers to their detail analysis. The adequate linking capabilities concerning exploration and exploitation of the knowledge are useful solution for organizations. This results from the rapid obsolescence of knowledge, shortening life cycle of many products and services. Therefore, it is important for the survival and success of the organization to maintain some balance between these activities.

V. CONCLUSIONS

The research propose of this study was to investigate how Resource-based View and dynamic capability theory may be adopted and used in BI area. It was illustrated that they through more light on our BI understanding. I have proposed a comprehensive, dynamic BI capabilities framework that reflects six BI capabilities areas: governance, culture, technology, people, processes and change & creativity. This dynamic capability framework suggests that the abilities needed depend highly on the dynamics of the environment.

The conducted survey has shown that BI is still treated as a technology or tool to acquire and analyze data and not as a trigger for making more effective decisions, improving business processes and business performance, as well as doing new business or creating new ideas and procedures. The organizations still underestimate the soft competences and skills needed for BI (e.g., culture-based on facts and knowledge, trust, human resources management, managing analytical/creativity people). Worried, that BI and business strategy are not aligned in many organizations.

I consider that the father development of BI in organizations will depend on how they will focus on strategic alignment BI and business strategy, establishing a fact-based and learning culture, entrepreneurship and innovation, change management, and creativity. The time of technical BI competences is over. Organizations should build a whole bundle of more soft BI capabilities.

Future research might take some of the following directions. It would be valuable to build holistic approach for building the dynamic BI capabilities. Further research might explore the detailed BI capabilities areas. Some empirical investigations and precise validations would be useful to explore the associations between BI capabilities and strategic orientations of the organizations.

REFERENCES

[1] T. H. Davenport, J. G. Harris, and R. Morison, Analytics at Work: Smarter Decisions, Better Results, Harvard Business Press, Cambridge, 2010.

[2] B.H. Wixom, and H.J. Watson, "The BI-based organization", International Journal of Business Intelligence Research, Vol. 1, No. 1, 2010, pp 13-28.

[3] S. Negash, and P. Gray, "Business Intelligence", in F. Burstein, and C.W. Holsapple (ed), Decision Support Systems, Springer, Berlin, 2008, pp 175-193.

[4] B. Liautaud, and M. Hammond, E-Business Intelligence. Turning Information into Knowledge into Profit, McGraw-Hill, New York, 2002.

[5] B.H. Wixom, H.J. Watson, and T. Werner, "Developing an Enterprise Business Intelligence Capability: the Norfolk Southern Journey", MIS Quarterly Executive, Vol. 10, No.2, 2011, pp 61-71.

[6] T. H. Davenport, and J. G. Harris, Competing on Analytics. The New Science on Winning, Harvard Business School Press, Boston Massachusetts, 2007.

[7] H. Chen, R.H.L. Chiang, and V.C. Storey, "Business Intelligence and *analytics: from Big data to big impact",* MIS Quarterly, Vol. 36, No. 4, 2012, pp. 1-24.

[8] L. Moss, and S. Atre, Business Intelligence Roadmap: The Complete Lifecycle for Decision-Support Applications, Addison-Wesley, Boston, 2003.

[9] J. Lufman, and T. Ben-Zvit, "Key issues for IT executives 2009: difficult economy's impact on IT, MIS Quarterly Executive, Vol. 9, No 1, 2010, pp 203-213.

[10] Gartner, Gartner's 2011 CIO survey results, http://www.gartner.com/it/page.jsp?id=1526414.

[11] O. Isik, M. C. Jones, and A. Sidorova, "Business Intelligence (BI) Success and the Role of BI Capabilities", Intelligent Systems in Accounting, Finance and Management, Vol. 18, 2011, pp 161-176.

[12] H.J. Watson, B. Wixom, "Enterprise agility and mature BI capabilities", Business Intelligence Journal, Vol. 12, No. 3, 2007, pp 4-6.

[13] S. Chaudhary, "Management factors for strategic BI success", in , M.S. Raisinghani (ed.), Business Intelligence in digital economy. Opportunities, limitations and risks, Hershey: IGI Global, 2004, pp 191-206.

[14] A. Schick, M. Frolick, and T. Ariyachandra, "Competing with BI and Analytics at Monster Worldwide", in Proceedings of the 44th Hawaii International Conference on System Sciences, 2011.

[15] C. Howson, Successful Business Intelligence: Secrets to Making BI a Killer Application, McGraw-Hill, New York, 2008.

[16] H.J. Watson, SME performance: Separating myth from reality, Cheltenham: Edward Elgar Publishing, 2010.

[17] A. Alter, A work system view of DSS in its fourth decade, Decision Support System, Vol. 38, No. 3, 2004, pp 319-327.

[18] D. Wells, "Business analytics – getting the point", [online], http://b-eye-network.com/view/7133, 2008.

[19] Z. Jourdan, R. K. Rainer, and T. Marschall, "Business Intelligence: An Analysis of the Literature*", Information Systems Management*, Vol. 25, No. 2, 2007, pp. 121-131.

[20] W.W. Eckerson, The keys to enterprise Business Intelligence: critical success factors. The Data Warehousing Institute, 2005, Retrieved October 02 2011 from http://download.101com.com/pub/TDWI/Files/TDWIMonograph2-BO.pdf.

[21] C.M. Olszak, and E. Ziemba, "Business Intelligence as a key to management of an enterprise", in E. Cohen, & E. Boyd (ed.), Proceedings of *Informing Science and IT Education InSITE'2003*, Santa Rosa, The Informing Science Institute, 2003.

[22] C.M. Olszak, and E. Ziemba, E. "Business Intelligence systems in the holistic infrastructure development supporting decision-making in organizations", Interdisciplinary Journal of Information, Knowledge and Management, Vol. 1, 2006, pp 47-58.

[23] C.M. Olszak, and E. Ziemba, "Business Intelligence systems as a new generation of Decision Support Systems", in J.V. Carrasquero (ed.), Proceedings of PISTA 2004, International Conference on Politics and Information Systems: Technologies and Applications. Orlando: The International Institute of Informatics and Systemics, 2004.

[24] C. M. Olszak, "Assessment of Business Intelligence Maturity in the Selected Organizations", in: M. Ganzha, L. Maciaszek, M. Paprzycki (ed.), Annals for Computer Science and Information Systems, Vol. 1, 2013, pp. 951–958, https://fedcsis.org/proceedings/2013/index.html.

[25] C. M. Olszak, "Business Intelligence as a key for the success of the organization", in M. Tvrdíková, J. Ministr (ed.), ICT for Practice, Ekonomicka Fakulta VSB-TU Ostrava, 2013, pp. 31-40.

[26] F. Albescu, I. Pugna, and D. Paraschiv, "Business Intelligence & Knowledge Management – Technological Support for Strategic Management in the Knowledge Based Economy", Revista *Informatica Economică*, Vol. 4, No. 48, 2008, pp. 5-12.

[27] H. Baaras, and H.G. Kemper, "Management support with structured and unstructured data – an integrated Business Intelligence framework", Information Systems Management, Vol. 25, No. 2, 2008, pp. 132-148.

[28] W. Chung, H. Chen, and J.F. Nunamaker, "A visual framework for knowledge discovery on the web: An empirical study of business intelligence exploration", Journal of Management Information Systems Vol. 21, No. 4, 2005, pp. 57-84.

[29] P. Venter, and D. Tustin, "The availability and use of competitive and business intelligence in South African business organizations", South African Business Review, Vol. 13, No 2, 2009, pp. 88-115.

[30] C. M. Olszak, "The Business intelligence-based Organization- new chances and Possibilities", in V. Ribiere and L. Worasinchai (ed.), Proceedings of the International Conference on Management, Leadership and Governance, Bangkok University, Published by Academic Conferences and Publishing International Limited Reading, 2013, pp. 241-249.

[31] R. T. Herschel, and N.E. Jones, "Knowledge management and business intelligence: the importance of integration", Journal of Knowledge Management, Vol. 9, No. 4, 2005, pp. 45-54.

[32] J.J. McGonagle, and C.M. Vella, Bottom Line Competitive Intelligence, Quorum Books , Westport, CT, 2002.

[33] A. Weiss, "A brief guide to competitive intelligence", Business Information Review ,Vol. 19, No 2, 2002.

[34] S. Williams, N. Williams, The Profit Impact of Business Intelligence. Morgan Kaufmann, San Francisco, 2007.

[35] P. R. Clavier, H. Lotriet, and J. Loggerenberger, "Business Intelligence Challenges in the Context of Goods-and Service-Domain Logic", in 45th Hawaii International Conference on System Science, IEEE Computer Society, 2012, pp. 4138-4147.

[36] M. Hannula, and V.Pirttimaki, "Business intelligence empirical study on the top 50 Finnish companies", Journal of American Academy of Business, Vol. 2, No. 2, 2003, pp. 593-599.

[37] A.J. Karim, "The value of Competitive Business Intelligence System (CBIS) to Stimulate Competitiveness in Global Market", International Journal of Business and Social Science, Special Issue, Vol. 2, No. 19, 2011, pp. 196-203.

[38] C. M. Olszak, and E. Ziemba, "Critical Success Factors for Implementing Business Intelligence Systems in Small and Medium Enterprises on the Example of Upper Silesia, Poland", Interdisciplinary Journal of Information, Knowledge, and Management, Vol. 7, 2012, pp.129-150. Informing Science Press, (http://www.ijikm.org/Volume7/IJIKMv7p129-150Olszak634.pdf)

[39] J. Barney, "Firm Resources and Sustained Competitive Advantage", Journal of Management, Vol. 17, No. 1, 1991, pp 99-120.

[40] M. Wade, and J. Hulland, "Review: The Resource-Based View and Information Systems Research: Review, Extension, and Suggestions for Future Research", MIS Quarterly, Vol. 28, No. 1, 2004, pp 1-25.

[41] R. Cosic, G. Shankes, and S. Maynard, "Towards a Business Analytical Capability Model", in 23rd Australian Conference on Information Systems, Geelong, 2012.

[42] B. Wernfelt, "A Resource-based View of the Firm", Strategic Management Journal, Vol. 5, 1984, pp 171-180.

[43] J. Barney, M. Wright, and D.J. Kitchen, "The resource-based view of the firm: ten years after 1991", Journal of Management, Vol. 27, 1991, pp 625-641.

[44] K. M. Eisenhardt, and J.A. Martin, Dynamic Capabilities: What Are They? Strategic Management Journal, (2000), Vol. 21, No 10/11.

[45] M. Zollo, and S.G. Winter, "Deliberate Learning And The Evolution Of Dynamic Capabilities". Organization Science, 13(3), 2002, pp 339-351.

[46] M.J. Ahn, and A.S. York, "Resource-based and institution-based approaches to biotechnology industry development in Malaysia". Asia Pacific Journal of Management, Vol. 28, No. 2, 2011, pp 257-275.

[47] J. Wiklundi, D. Shepherd, "Knowledge-based resources, entrepreneurial orientation, and the performance of small and medium-sized businesses". Strategic Management Journal, Vol. 24, 2003, pp 1307-1314.

[48] Gartner, "Magic Quadrant for Business Intelligence Platforms", Gartner Group, report G00210036.

[49] S. Nevo, and M. Wade, "The formation and Value of It-Enabled Resources: Antecedents and Consequences of Synergistic Relationship", MIS Quarterly, Vol. 34, No. 1, pp 163-183.

[50] D.J. Teece, G. Pisano, and A. Shuen, "Dynamic capabilities and strategic management", Strategic Management Journal, Vol. 18, No. 7, 1997, pp 509-533.

[51] L. Vincent, "Differentiating Competence, Capability and Capacity", Innovating Perspective, Vol. 16, No. 3, 2008, 460-1313, http://www.innovationsthatwork.com/images/pdf/June08newsltr.pdf

[52] C.K. Prahaland, and G. Hamel, The Core Competence of the Corporation, Harvard Business Review, May-June 1990.

[53] K. Ortbach, R. Plattfaut, J. Poppelbuss, B. Niehaves, "A Dynamic Capability-based Framework for Business Process Management: Theorizing and Empirical Application", in 45th Hawaii International Conference on System Sciences, IEEE, 2012, pp 4287-4296.

[54] C.E. Helfat, S. Finkelstein, W. Mitchell, M.A. Peteraf, H. Sing, D.J. Teece, and S.G. Winter, Dynamic Capabilities Understanding Strategic Change in Organisations, Carlton: Blackwell, 2007.

[55] B.K. Chae, D.L. Olson, "Business Analytics for Supply Chain: A Dynamic-Capabilities Framework", International Journal of Information Technology & Decision Making, Vol. 12, No. 1, 2013, pp 9-26.

[56] D. Lavie, U. Stettner, and M.L. Tushman (2010). "Exploration and Exploitation Within and Across Organizations", The Academy of Management Annals, Vo. 4, No. 1, 2010, pp 109-155.

Acknowledgment

This paper has been supported by a grant: „*Methodology for Computer Supported Organizational Creativity"* from National Science Centre in Poland, 2013/09B/HS4/00473.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1111–1116
DOI: 10.15439/2014F123
ACSIS, Vol. 2

Using parameter optimization to calibrate a model of user interaction

Tommy Baumann*, Bernd Pfitzinger‡§ , Dragan Macos†, Thomas Jestädt‡
*Andato GmbH & Co. KG, Ehrenbergstraße 11, 98693 Ilmenau, Germany. tommy.baumann@andato.com
‡Toll Collect GmbH, Linkstraße 4, 10785 Berlin, Germany. {bernd.pfitzinger|thomas.jestaedt}@toll-collect.de
§FOM Hochschule für Oekonomie & Management, Zeltnerstraße 19, 90443 Nürnberg, Germany.
†Beuth Hochschule für Technik Berlin, Luxemburger Str. 10, 13353 Berlin, Germany. dmacos@beuth-hochschule.de

***Abstract*—Simulation models of real-world distributed systems depend both on the accuracy of the underlying model and the interaction between user and system. The user interaction is typically modeled as stochastic process depending on parameters and distributions describing the actual usage. Accurate data is often not available and (manual) assumptions are necessary. Taking an existing large-scale simulation model of the German tolling system we discuss the use of a genetic optimization algorithm for calibrating the simulation model.**

I. Introduction

DISTRIBUTED software-intensive systems become a part of everyday life. The engineering and operations of these systems is not yet well established: Most techniques in use focus on standalone systems [1] and even there successful implementations are not guaranteed [2]. Instead of the engineering aspects one can rather argue [3] that the integration of subsystems into a functioning system-of-systems becomes a core strategic business capability. Whether it is the engineering, integration or the subsequent operation of a highly automated software-intensive system – many dynamic aspects depend on its usage and often unknown user behavior.

Taking the example of the German automatic toll system for heavy goods vehicles (HGVs) – a typical example of a modern toll system based on global navigation satellite systems (GNSS) [4] – we complement the system operations and system design through simulations [5]. Having a detailed, realistic simulation model at hand it is possible to predict the upcoming operational behavior (e.g. for fleet-wide updates) even for systems under design e.g. when a different system architecture is proposed. In both scenarios simulation results yield numerical results to support decisions and to reduce the risk inherent in any software development process. Particularly in the latter case simulation models help to explore different solutions and to create exact specifications right from the start of a software development project.

"*Good models are essential for communication among project teams and to assure architectural soundness*" [6]. Yet the emphasis on communication (even in more formal methods as UML [6]) creates a source for misunderstandings and errors through the inexact verbalization of the requirements. To reduce the ambiguity we use executable models, i.e. implement the requirements in a model that can compiled and executed [7]. From the very beginning this allows verifying the system requirements by comparing the simulation results with the expected behavior. As the development of the system progresses the simulation models can progress as well (to give an accurate model at the level of abstraction available at that time) or remain at a reasonable level of detail sufficient e.g. to simulate the overall system dynamics. In the case of the German toll system we have established a simulation model of the automatic tolling process and used it both to predict the operational behavior of the existing system [8] and to aid the software development process to better scope proposed changes (e.g. [5]).

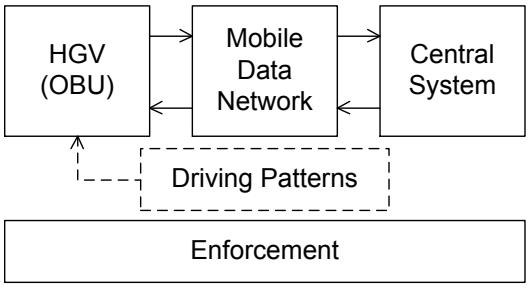


Fig. 1. High-level system design of a GNSS-based electronic tolling system and its dependency on the user interaction (driving patterns).

In the next section we introduce the simulation model of the German toll system. To get an executable system we need two models: A model of the technical system and a model of the user interaction. At present not much is known about the user interaction (due to technical restrictions and data protection regulation) and we started with a simple model that can be parametrized to fit the observed dynamic behavior of the toll system. Section IV uses a genetic algorithm (introduced as a separate model in section III) to find a set of parameters that best reproduces the actual system dynamics. The initial results are given in section V followed by a brief summary.

II. Simulation model of the German toll system

Starting with an existing model of the German toll system and a simple model of the user interaction [8] we take data observed in the real-world system to measure the accuracy of the simulation model. This section gives a brief overview of two models necessary to reproduce the dynamic behavior of the German toll system. In addition we need to address the

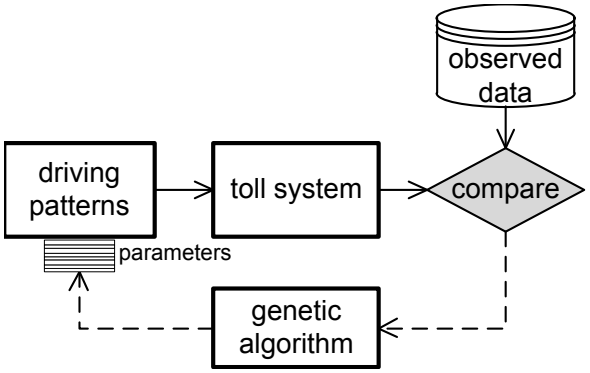


Fig. 2. The existing models for the driving patterns and the toll system are complemented by a dataset of the known dynamic behavior of the toll system. Adding a fitness functions allows comparing the simulation results with the observed data and the use of a genetic algorithm (again as a model) to choose the right parametrization.

term "accuracy" to define the appropriate in our context more clearly.

A. System model

The main model from the perspective of system operations or engineering is of course the model describing the (technical) toll system (Fig. 2). It is a discrete event simulation model of the German automatic toll system encompassing a fleet of almost 800 000 HGVs, each with an on-board unit and access to the central system via a mobile data network. The major processes at time scales of 1 second and above are included in the model (and some at considerably shorter time scales) including the network connection with their respective bandwidths and latencies (but not modeled on the level of the TCP/IP protocol). Simulating at a scale of 1:1 we do not introduce ambiguities due to scaling (especially since the system under consideration is in parts highly non-linear) but have achieve a high simulation speed. Looking at fleet-wide updates taking more than a month the model itself works with typical time scales of one second. All in all the typical simulation performance after a number of performance improvements [9] gives execution times of less than 10 hours for the simulating the whole fleet over half a year.

The model has been verified through software inspection [10] and validated through the comparison with the data observed in the real-world system. At present the biggest source of uncertainty is introduced by the driving patterns (our model of the user interaction).

B. "Driving patterns": A model of the user interaction

The model of the toll system depends on the external stimulus of the user interaction. With an emphasis on the fleet-wide propagation of updates we started with a simple model describing only the temporal behavior (fig. 3): The points in time when an HGV is powered on (or off) and when a toll event is created. Since we do not yet include any geographic information the toll events correspond to the HGV driving

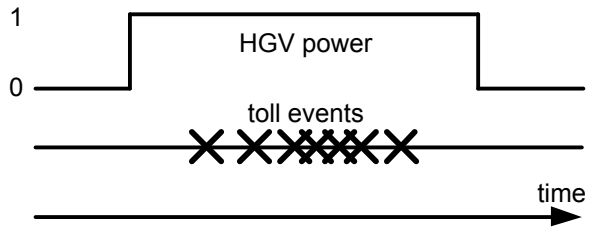


Fig. 3. The user interaction is modeled as the points in time where power-cycles or toll events are registered for a given HGV.

distance of 4.2 km (the average length of a toll segment) at an average speed of $\approx$ 80 km/h.

To achieve a realistic update behavior (taking several weeks to reach the whole fleet) we start with a probability distribution of "active weeks", i.e. the probability that a given HGV is powered-on (at least once) in N of M weeks and take data observed in the actual system and existing HGV-fleets used for testing (typically covering several hundred to a few thousand HGVs). This is in turn followed by probability distributions determining the number of "active days" within a week, the number and duration of power cycles per day and the time during the day when power-cycles take place (for details see [8]). The driving patterns are (manually) calibrated to reproduce the update behavior observed in the real-world system – i.e. looking at time scales of several weeks taking one sample per day.

However, in reality not much is known about the user interaction. Apart from small test fleets no data is available on the power-cycles of the HGVs (even the average speed is not known): Most often the data is not collected and even if data is available data protection regulation often prohibits its use [11]. Looking at the simulation results in more detail it is very difficult to manually adjust the parametrization as to reproduce the intra-day dynamic system behavior (see fig. 4). To make matters worse, some processes are deliberately made strongly non-linear (e.g. to favor updates during the night-time or to protect the central system when it is operating close to the specification limit).

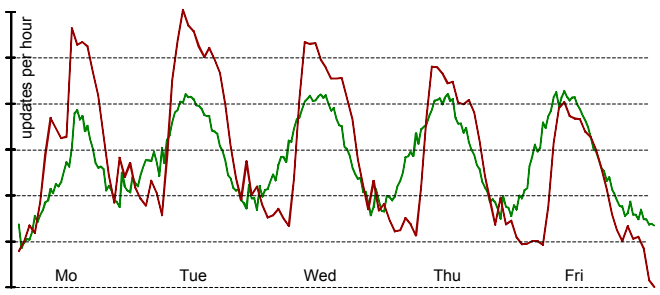


Fig. 4. Example of the simulation results (green line) in comparison to the observed data (red line) of the hourly update rate.

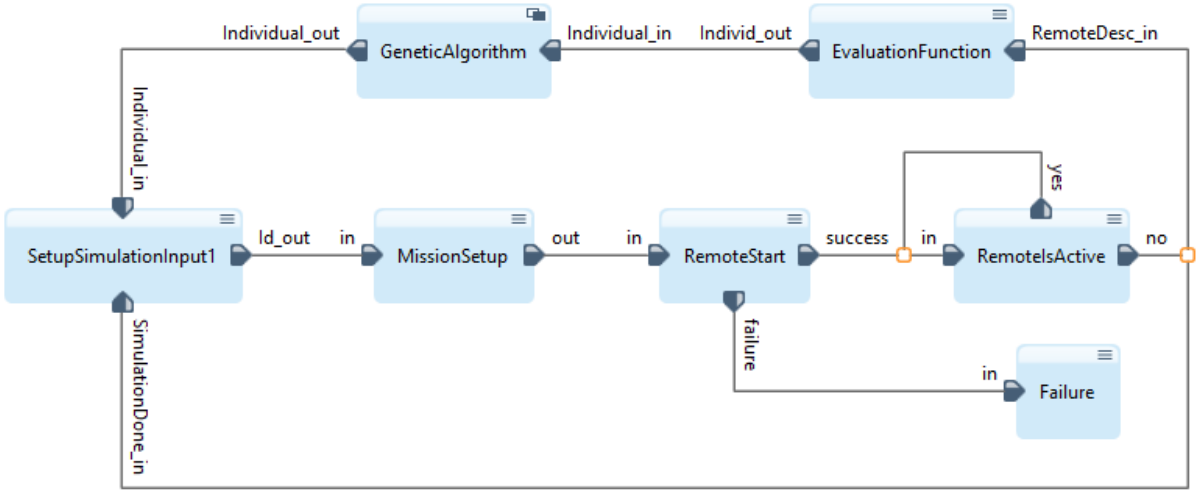


Fig. 5. MSArchitect model using the genetic algorithm to optimize the user interaction (driving patterns) model.

C. Measuring accuracy

To overcome these difficulties we started to let an optimization algorithm adjust the parameters to improve the simulation results. Keeping in mind the mantra "good enough" [12] we need to look at the context in which the simulation results are used:

- *Updates* take several weeks to propagate across the whole fleet. A comparison of the daily data is sufficient to measure the quality of the simulation results.
- *Utilization* of servers or data networks changes rapidly and a time resolution of an hour or less needs to be taken into account.

With the emphasis on fleet-wide updates we compute the fitness as rms-deviation using one data point per day. Depending on the use case this may include the one or several components that are updated: geo and tariff data and the OBU software. For the purpose of this article we take the OBU software as the only input to the fitness function and compute the rms-deviation only for the first weeks after the start of a software update. Only the first weeks of an update are influenced by the system simulation model (and its parametrization of the update rate), very soon the HGV activity is the limiting factor (i.e. a substantial part of the HGV fleet connects only rarely to the central system either because the HGV is constantly powered off or in a foreign country).

III. Simulation Model to Optimize the Driving Patterns

The limited knowledge of the actual user interaction (as mentioned above) combined with the difficulty of setting the parameters of the user interaction manually necessitates the use of an automatic optimization algorithm to calibrate the model with the data available. Consequently the model of the German toll system is connected to an optimization algorithm (see Fig. 2). To separate these tasks we introduce a second optimization model responsible for the optimization of the driving patterns (user interaction). Using a genetic algorithm the model iterates automatically over different parametrization improve the simulation results for the model of the German toll system.

To that extent the existing model of the German toll system needed only minor modifications: Any parameters intended for optimization were implemented as explicit parameters at the top-level of the model and exported to the optimization model. The existing models generating the driving patterns and simulating the tolling system are now integrated into one model and followed by a *fitness function* evaluating the deviation between the simulation results and real-world data (typically of the progress of fleet-wide updates computed as the rms of the daily version status).

The idea of using a separate model controlling the optimization process, including structural and parameter modifications as well as evaluation of the model to be optimized, is a basic concept of Simulation-Driven Design [7]: There it is called *Executable System Design Process*, defined as an automated series of design steps, which alter the Executable System Specifications (in our case the model to be optimized) in a formal, consistent, and self-contained manner to enable processing [13]. Three base types of components are differentiated:

- *Execution components* are responsible for the execution of the whole, of the parts or of abstractions of the embedded executable system specification as well as execution of associated systems.
- *Control components* implement the evaluation of constraints, rules and objective functions to control the execution of process components.
- *Generator components* generate, transform and extend executable models to comply with different purposes, abstraction levels, parameterizations and structural architectures.

Looking at the simulation model used for optimizing the driving patterns (Fig. 5) we recognize all three types of compo-

nents: Starting with the block instance `GeneticAlgorithm` we recognize a generator component responsible for generation of individuals with different genomes. The genome represents a parametrization of the toll system in the form of a parameter vector and is sufficient to execute the model of the toll system for the given individual. The value of each parameter is bound to a defined range and granularity (i.e. a bit representation of the value) – at present the parameters are the number of active weeks of domestic and foreign trucks (see section II). In the next step `GeneticAlgorithm` sends the representation of the individual via its output port and connection to the block instance `SetupSimulation` responsible for the preparation of the toll system model. `SetupSimulation` (obviously a generator component) creates the necessary environment e.g. the directory structure for simulation in- and output including the parameters and any necessary configuration files. In addition `SetupSimulation` checks the available resources and chooses the number of simulations to be executed in parallel (depending on the number of CPU cores available and the population size). However, the final setup of a simulation run is delegated to another generator component (`MissionSetup`). It generates a mission descriptor data structure, containing the command line parameters to setup and execute a simulation run. This data structure is in turn sent to the first execution component (`RemoteStart`) to execute the model on a specific CPU core. While the simulation run is ongoing two further execution components (`RemoteIsActive` and `Failure`) observe the simulation run and inform the subsequent block instances `SetupSimulation` and `EvaluationFunction` when the simulation run has finished: `SetupSimulation` prepares a new simulation run if necessary an `EvaluationFunction` evaluates the simulation results with respect to the fitness function (see section IV). Since block instance `EvaluationFunction` decides about continuing the optimization loop or not it is a control component.

IV. Applied Optimization Algorithm

To solve our optimization problem we decided to use a genetic algorithm. The straight-forward implementation of a parallelized genetic algorithm was the main reason to choose this optimization algorithm. A genetic algorithm is a search algorithm for optimization purposes based on the mechanics of natural selection and natural genetics. Genetic algorithms are able to avoid getting stuck in a local optimum in the search space, can be used in high-dimensional search spaces and are trivially parallelized ("embarrassingly parallel", [14]). They belong to the group of so called meta heuristics – search methods for approximate solutions [15].

Fig. 6 gives the generic flow chart of the genetic algorithm used: At the beginning an initial population of a fixed size is generated either randomly or using previously available individuals. To limit the search space we choose the parameters to be within pre-defined intervals. Next the fitness of all individuals in the population is computed (step 2 in Fig. 6) to yield the 'parent population'. In the third step the algorithm

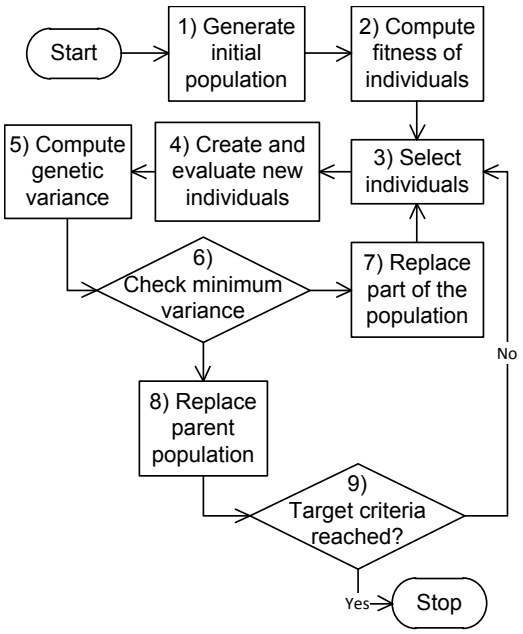


Fig. 6. General sequence of action of the genetic algorithm used (based on [16, 17]).

randomly selects two sets of parents in a *tournament selection* to choose those two parent individuals with the best fitness value.

Once two parent individuals are selected one child individual is created using uniform crossover without mutation (step 4 in Fig. 6). For crossover we select one part of the genome of the first parent individual and the complementary part from the second parent individual. In order to do this a crossover point is chosen randomly. At present we are not using mutation where parts of the genome of an individual are changed randomly to increase population diversity. From the perspective of optimization theory this method is used to overcome local optima [18] – which is implemented in our case by enforcing a minimum variance within the population (step 6 in Fig. 6).

When the child population is fully populated the optimization model starts to evaluate the fitness function by executing simulation runs of the toll system model (as described above). When the evaluation is finished the algorithm checks the genetic variance [19] inherent within the child population (step 5 in Fig. 6). If the variance becomes too small a part of the population is replaced by new randomly generated individuals (step 7 in Fig. 6) otherwise the child population replaces the parent population (step 8 in Fig. 6) and the optimization run continues until the target criteria are met (step 9 in Fig. 6). In addition we use step 6 to check whether the optimization run finds better solutions and again replace part of the population if the results did not improve within 6 generations.

In our case we use fairly small populations with 765 individuals i.e. at a scale of 1:1000 and a genome of 32 parameters each expressing the probability for an HGV of

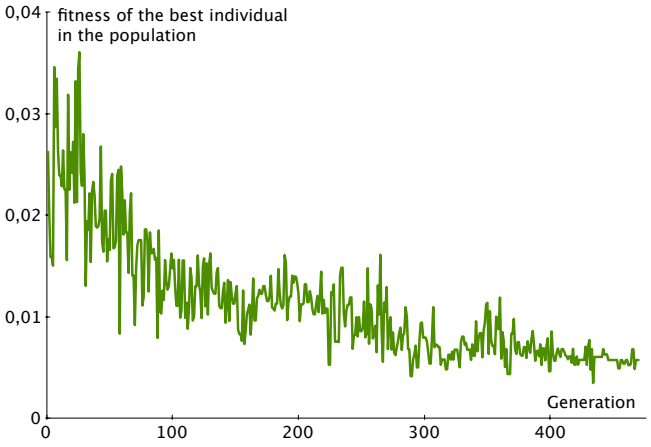


Fig. 7. Fitness of the best individual per generation over the optimization run.

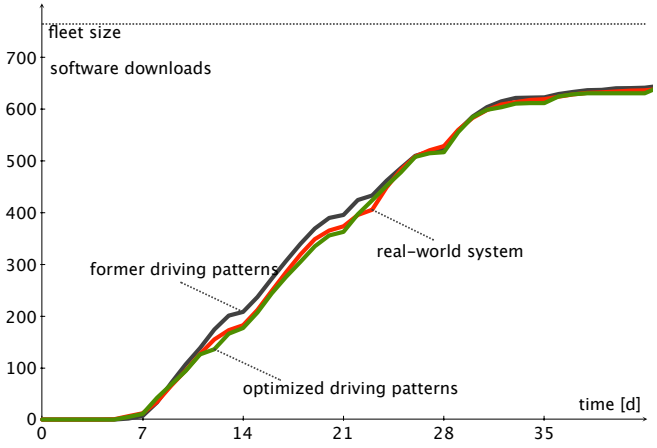


Fig. 8. Simulation results for a fleet-wide software update before and after optimization (red line: data observed in the real-world system, black line before optimization and green line after optimization).

being active (with at least one power cycle) a certain number of weeks within a 15 week period (once for German HGVs and foreign ones). The optimization algorithm starts with a given probability distribution based on historical data from the real-world system (or a previous optimization run). In the next step the fitness function (also called target or evaluation function) is evaluated to quantify the fitness of each individual. Since the evaluation of each individual is independent the algorithm is trivially parallel and we send each computation via the `Remote` component to a different CPU node (see section III) for execution, i.e. the model of the German toll system with the parameterization given by the individual is executed for each individual in parallel.

As an example we take a fleet-wide software update that was rolled out in spring 2012. Using fleet-wide timing parameters the roll-out was configured to spread over 6 weeks where a single update needed less than 10 minutes to download under optimal conditions. To achieve a reasonable number of function evaluations in the optimization we run the simulation model at a scale of 1:1000 resulting in an execution time of less than one minute for the time period of interest – 4 weeks before the start of the update and the first 10 weeks of the update. The simulation model itself is not modified from previous versions [5] and each instance works within its own subdirectory to read and write intermediate results as necessary. At the end of each simulation run the fitness function is evaluated expressing the quality of each individual as the square deviation between the simulated and the real world update roll-out curve (see section V).

V. Optimization Results

For the purpose of this discussion we choose a software update in 2012. Without access to the optimization algorithm the simulation model was parametrized using statistical data from the real-world system and subsequent minor manual adjustments. In comparison we give the results after performing an optimization run with the simulation model at a scale of 1:1000, a population size of 64 and some 400 generations.

Fig. 7 gives the evolution of the fittest individual per generation. The fitness gradually improves over run-time but from generation to generation it can give worse results since the algorithm creates a completely new population for each generation without keeping the best individual around.

Looking at the results (Fig. 8) the optimized driving patterns perform considerably better during the main update phase. The real-world system is configured to give an almost constant rate of updates during the first weeks (red line in Fig. 8) where the OBUs decide randomly when to download the update according to fleet-wide timing parameters. After a few weeks the update rate is determined mostly by those OBUs that are rarely active within the German mobile data networks and no longer depends on the download parameters. So far the previously existing user interaction model typically produces too many updates during the 2nd and 3rd week (black line in Fig. 8) even though the model uses statistical data gathered in the real-world system on Toll Collect test fleets.

To emphasize the time period where the algorithms of the toll system determine the download rate rather than rarely visiting HGVs we compute the fitness function only for the initial 6 weeks. This results in a marked improvement of the simulation results (green line in Fig. 8) for the time period shown. However, since the long-term activity pattern was in this case not subjected to optimization the optimized driving patterns give somewhat worse results for longer time periods (not shown). Taking the deviation from the data observed (Fig. 9) the improvement during the first two weeks of the software update are obvious.

Optimizing the probability distribution for the weekly activity pattern quickly improved the simulation results. However, deviations are still visible even when using a coarse time resolution of one day : Typically at the end of the workweek the difference is biggest and changes its sign with the coming week. This suggests that at least further parameters need

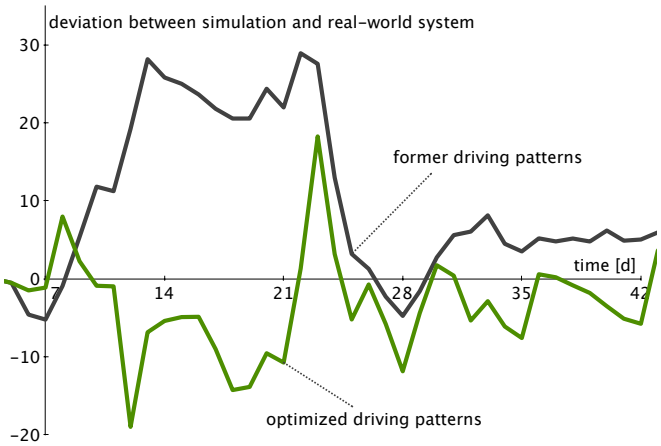


Fig. 9. Difference (in cumulated downloads) between the data observed in the real-world system and both simulation runs (black line before optimization, green line after optimization).

systematic optimization or even a different underlying model to create the driving patterns.

VI. Summary

Taking the example of the German automatic toll system we have discussed the challenge to model the user interaction. Even with a simple model many parameters (e.g. probability distributions) need to be adjusted so as to achieve "good enough" simulation results. The use of a genetic algorithm simplifies the optimization i.e. adjusts the parameters as good as possible starting from the limited data available. The sheer number of parameters available poses a significant challenge even to a parallelized genetic algorithm. To us this suggests that the model of the user interaction is not yet expressed in the right way. In addition, "good enough" models depend on the context. In our example, a better model of the user interaction is needed to reproduce the intra-day behavior – e.g. to use the simulation to monitor everyday operations of the toll system.

Looking back at fig. 2 this article discussed the recently added model of a genetic algorithm. In future work the model of the user interaction should be split in two parts: A generic, domain-independent model of stochastic processes and its domain-specific application to generate driving patterns.

References

[1] Barry W. Boehm. A view of 20th and 21st century software engineering. In *Proceedings of the 28th international conference on Software engineering*, pages 12–29. ACM, 2006. doi: 10.1145/1134285.1134288.

[2] Robert L Glass. The standish report: does it really describe a software crisis? *Communications of the ACM*, 49(8):15–16, 2006. doi: 10.1145/1145287.1145301.

[3] Michael Hobday, Andrew Davies, and Andrea Prencipe. Systems integration: a core capability of the modern corporation. *Industrial and corporate change*, 14(6):1109–1143, 2005. doi: 10.1093/icc/dth080.

[4] Julia Numrich, Sascha Ruja, and Stefan Voß. Global navigation satellite system based tolling: state-of-the-art. *NETNOMICS: Economic Research and Electronic Networking*, 13(2):93–123, 2012. doi: 10.1007/s11066-013-9073-9.

[5] Bernd Pfitzinger, Tommy Baumann, Dragan Macos, and Thomas Jestädt. Using simulations to study the efficiency of update control protocols. *47th Hawaii International Conference on System Sciences (HICSS)*, pages 5154–5161, 2014. doi: 10.1109/HICSS.2014.634.

[6] ISO. ISO/IEC 19501:2005 unified modeling language specification, 2005.

[7] Tommy Baumann. Simulation-driven design of distributed systems. *SAE Technical Paper*, (2011-01-0458), 2011. doi: 10.4271/2011-01-0458.

[8] Bernd Pfitzinger, Thomas Jestädt, and Tommy Baumann. Simulating the German toll system: Choosing "good enough" driving patterns. In *Proceedings of the mobil.TUM 2013 – International conference on mobility and transport*, 2013. URL http://www.mobil-tum2013.vt.bgu.tum.de/media/contributions/.

[9] Tommy Baumann, Bernd Pfitzinger, and Thomas Jestädt. Simulation driven design of the German toll system – profiling simulation performance. In *2013 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 923–926. IEEE, 2013. ISBN 978-1-4673-4471-5.

[10] M. E. Fagan. Design and code inspections to reduce errors in program development. *IBM Systems Journal*, 15(3):182–211, 1976. doi: 10.1147/sj.153.0182.

[11] Reinhard Fraenkel and Volker Hammer. Keine Mautdaten für Ermittlungsverfahren. *Datenschutz und Datensicherheit – DuD*, 30(8):497–500, 2006. doi: 10.1007/s11623-006-0259-2.

[12] Dave Thomas and Andy Hunt. *The Pragmatic Programmer: From Journeyman to Master*. Addison-Wesley Professional, 1999. ISBN 978-0201616224.

[13] Tommy Baumann. *Automatisierung der frühen Entwurfsphasen verteilter Systeme*. Südwestdeutscher Verlag für Hochschulschriften, Saarbrücken, Germany, 2009. ISBN 978-3-8381-1266-4.

[14] Barry Wilkinson and Michael Allen. *Parallel programming*. Prentice Hall New Jersey, 2nd edition, 2004. ISBN 978-0131405639.

[15] Ibrahim H Osman and James P Kelly. *Meta-heuristics: theory and applications*. Springer, Berlin Heidelberg, 1996. ISBN 978-1-4613-1361-8.

[16] Gabriel Alvarez. Can we make genetic algorithms work in high-dimensionality problems. In *Report 112*, Stanford Exploration Project, pages 195–212. November 2002.

[17] Hartmut Pohlheim. *Evolutionäre Algorithmen – Verfahren, Operatoren und Hinweise für die Praxis*. VDI-Buch. Springer, Berlin, Heidelberg, 2000. ISBN 978-3-540-66413-0.

[18] Ingrid Gerdes, Frank Klawonn, and Rudolf Kruse. *Evolutionare Algorithmen: Genetische Algorithmen – Strategien und Optimierungsverfahren – Beispielanwendungen*. Springer-Vieweg, Wiesbaden, 2004. ISBN 978-3-322-86839-8.

[19] Ronald W. Morrison and Kenneth A. De Jong. Measurement of population diversity. In Pierre Collet, Cyril Fonlupt, Jin-Kao Hao, Evelyne Lutton, and Marc Schoenauer, editors, *Artificial Evolution*, volume 2310 of *Lecture Notes in Computer Science*, pages 31–41. Springer Berlin Heidelberg, 2002. ISBN 978-3-540-43544-0. doi: 10.1007/3-540-46033-0_3.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1117–1122
DOI: 10.15439/2014F347
ACSIS, Vol. 2

Hybrid framework for investment project portfolio selection

Bogdan Rębiasz
AGH University of Science and Technology
Krakow, Poland
Email: brebiasz@zarz.agh.edu.pl

Bartłomiej Gaweł
AGH University of Science and Technology
Krakow, Poland
Email: bgawel@zarz.agh.edu.pl

Iwona Skalna
AGH University of Science and Technology
Krakow, Poland
Email: skalna@agh.edu.pl

***Abstract*—Project selection is a complex multi-criteria decision making process that is influenced by multiple and often conflicting objectives. The complexity of the project selection problem is mainly due to the high number of projects from which an appropriate collection (an effective portfolio) of investment projects must be selected. This paper presents a new conception of a hybrid framework for construction of an effective portfolio of investment projects. The parameters of the considered model are described using both probability distributions and fuzzy numbers (possibility distributions). The proposed framework enables to take into account stochastic dependencies between model parameters and economic dependencies between projects. As a result, a set of Pareto optimal solutions is obtained. The performance of the proposed method is illustrated using an example from metallurgical industry.**

I. Introduction

THE of estimation and selection of investment projects is often named in literature as ”capital budgeting”. An effective capital budget (a portfolio of investments) is a budget which provides the maximum NPV (Net Present Value) for an acceptable level of risk or the lowest level of risk for a given acceptable NPV of a portfolio. The choice of an appropriate method for risk assessment is associated, among others, with the problem of description of uncertainty in business activity.

For many years, probabilistic calculus was considered as the only appropriate way to mathematically describe and deal with uncertainty. However, in real problems of assessing risk in business activity, not only randomness, but also imprecise or incomplete data is an important source of information. For this reason, many researchers often use alternative ways of modelling of uncertainty, such as fuzzy sets or interval numbers. Also for the selection of investment portfolio, the most appropriate approach to risk assessment is to develop and use methods that allow different representations of uncertainty (e.g., by probability distributions, fuzzy numbers and intervals) to be processed according to their nature and only finally combine them into a synthetic easy-to-interpret risk measure.

Another important aspect of defining an effective portfolio of investment projects is the analysis of the dependency problem. There are two kinds of dependency. The dependency between parameters is usually described statistically and, therefore, is called “statistical dependency”; statistical dependency is typically modelled by fuzzy or probabilistic correlation or regression. The second type of dependency is projects’ interdependency, usually called “economic dependency”. It is used to describe an interaction between investment projects. Interdependency is especially challenging to model, due to the difficulties with its description.

This paper briefly presents a novel framework for the selection of an efficient portfolio of investment projects. The proposed framework integrates a non-linear programming with tools that enable to describe interdependency between projects in a situation when model parameters are described both using probability distributions and fuzzy numbers. The paper has the following structure. Section II outlines different ways of description of uncertainty. In Section III, the current state of art in the selection of a portfolio of investment project is presented. A novel framework for the selection of an effective portfolio of investment projects is suggested in Section IV. In Section V a numerical example is solved using the proposed framework to demonstrate the effectiveness of the latter.

II. Description of uncertainty in evaluation of investment projects

Most of models of the real-world investment projects contain a mixture of quantitative and qualitative data. Therefore, increasingly often alternative descriptions of uncertainty in the assessment of the efficiency of investment projects are applied. The most common situation is when some parameters are described by probability distributions (statistical data), while others are given in the form of possibility degrees (subjective assessments of phenomena made by experts [27], [4]), i.e., the available data is heterogeneous in nature. To sum up, one may say, after Baudrit *et al.* [4], that randomness and imprecise or missing information are two reasons of uncertainty, which have an impact on the analysis of economic efficiency. Therefore, in the process of the evaluation of investment projects (estimation of efficiency and risk of projects), it is inevitable to deal with uncertainty caused by vagueness intrinsic to human knowledge and imprecision or incompleteness resulting from the limit of human knowledge [13], [20]. Hence, it is necessary to use a scheme for representing and processing vague, imprecise, and incomplete information in conjunction with precise and statistical data [4], [8], [13].

978-83-60810-58-3/$25.00

There are hardly a few studies that describe the use of hybrid data [25], i.e., data partially described by probability distributions, and partially by possibility distributions. The use of such data allows to reflect more properly the knowledge on parameters of economic calculus. However, very often, in the assessment of efficiency of investment projects, no distinction is made between these two types of uncertainty, both being represented by means of probability distributions [27]-[13]. Whereas, as suggested by Ferson and Ginzburg [11], distinct methods are needed to adequately represent random variability ("objective uncertainty") and imprecision ("subjective uncertainty").

III. Methods for the selection of effective portfolios of investment projects

The problem of capital budgeting was for the first time formulated by Lorie and Savage [20]. Later on, it was solved using mathematical programming methods. First works on this subject date back to 1960s and 70s [6]-[2]. The problem of determining the capital budget was also solved using linear programming, linear programming with binary variables and multi-objective programming methods.

A lot of attention, especially in the recent years, is given to the risk of investment projects. A method for the construction of an effective portfolio of investment projects on the capital market was first presented by Markowitz [22]. Seitz has adopted the ideas of Markowitz for capital budgeting [26] by using the binary quadratic programming. Methods for the selection of an effective portfolio of investment projects are being constantly improved [7], [9], [1], [3], [24]. Probability distributions of selected parameters were used to describe the uncertainties in these models. In the literature, also presented are methods for the selection of a portfolio of investment projects in the case when uncertain parameters of efficiency calculus are described by means of fuzzy numbers. Such methods were proposed by Huang [15], [16] and Liu and Iwamura [19] and Kahraman [18].

Guyonnet *et al.* [13] has proposed a method which facilitates estimation of risk in the case when probability and possibility distributions are used simultaneously. This method was a modification of the method proposed previously by Cooper *et al.* [8]. Methods for processing hybrid data combine stochastic simulation with arithmetic of fuzzy numbers. As a result of processing of such data, Guyonnet *et al.* [13] define two cumulative distribution functions: optimistic and pessimistic. Similarly, Baudrit *et al.* [4] use probability and possibility distributions in risk analysis. As a result of processing of such data, authors obtain random fuzzy variable, which characterizes the examined phenomenon.

Dickinson *et al.* [10] presented a method for optimal scheduling of investment projects, which takes into account the fact that particular projects can be complementary or substitutive to each other. Santhanam and Kyparisis [29] presented a mathematical model for the selection of a portfolio from economically dependent investment projects associated with the development of information systems. Zuluaga *et al.* [31] presented a model that enables the selection and scheduling of economically dependent investment projects. However, the models of Dickson, Santhanam and Kyparisis and Zuluga do not take into account uncertainty of cash flows generated by investment projects and stochastic dependencies between projects. Medaglia *et al.* [23] proposed the usage of evolutionary algorithms for the selection of economically and stochastically dependent investment projects.

It must be, however, highlighted that there are no methods for the selection of effective investment portfolio, which could process hybrid data, e.g., data expressed in the form of fuzzy numbers and probability distributions. In most of the existing approaches, different ways of uncertainty representation are usually unified by transforming one form of uncertainty into another. Obviously, such transformation entails some problems. For example, transformation of a probability distribution into a possibility distribution causes the loss of information, whereas the opposite one requires additional information to be introduced. This leads to systematic errors in the estimation of efficiency. It is, therefore, necessary to elaborate a framework for representing and processing stochastic, vague, imprecise, and incomplete information in conjunction with precise data for selection of investment project portfolio. Such a framework should also be able to take into account stochastic and economic dependencies.

IV. A framework for the selection of investment project

The process of building an effective capital budget consists of three phases [14]: strategic consideration, individual project evaluation and portfolio selection. Because the approach proposed here focuses on interdependency between projects, the problem of building an effective capital budget is divided into two models – *portfolio selection model* (PSM) and *portfolio evaluation model* (PEM). The purpose of first model is to find selection of the investment projects to gain the best evaluation parameters. Second model is used to determine evaluation parameters for a given set of investment projects.

A. Projects interdependency in uncertain environment

PSM focuses on the selection of projects. Most of project portfolio optimization methods and tools treat each project in a portfolio as an isolated entity. This leads to systematic errors in the estimation of risk and efficiency, and usually produces large overestimation. In order to eliminate these deficiencies, the interdependency between project should be considered. Three types of projects interdependencies are recognized in the literature: benefit, resource and technical [12].

Resources interdependency occurs when the demand for resources to develop projects independently is greater than amount of resources required when all of projects are selected. *Benefit interdependency* occurs when the total advantage of at least two independent projects increases or decreases when these projects are treated as interrelated. *Technical interdependency* occurs when there is a set of exclusive projects such that only one of them may be selected.

There are five classes of project portfolio selection models [2]: ad hoc approaches (e.g., profiles), comparative approaches [28] (e.g., AHP), scoring models, portfolio matrices, and optimization models. PSM is multi criteria linear programming model, and PEM is non-linear programming model combined with stochastic simulation.

B. Portfolio selection model (PSM)

Let us consider a company which plans to launch m potential projects. Due to the changing environment, the company must select a proper subset of those projects. Let each project create new or modify existing primary process steps. A portfolio of investment projects is defined as $(x_1, \ldots, x_m)$ and $x_i = 1$ when project i is selected and 0 otherwise ($i \in I = \{1, \ldots, m\}$). Let b defines overall budget allocation for a selected portfolio, and c_i initial cost of implementation of i-th investment. Let $fin(x_1, \ldots, x_m)$ denote financial evaluation parameter for a given portfolio of investments. The performance of the selected portfolio is measured by two functions: $E(fin(x_1, \ldots, x_m)$ and $\sigma(fin(x_1, \ldots, x_m))$. Then, the selection of the portfolio of investments is defined as follows: find $(x_1, \ldots, x_m)$ that maximize of the expected value of $fin(x_1, \ldots, x_m)$ and minimize $\sigma(fin(x_1, \ldots, x_m))$ subject to:

- *portfolio selection constraints* – for each investment, the cost of implementation cannot exceed the overall budget $\sum_{i \in I} c_i * x_i \leqslant b$
- *integrability constraints*: $x_i \in \{0, 1\}, i \in I$

In order to solve PSO problem, for each portfolio of investments, PEM model must be invoked in order to compute evaluation parameters.

C. Portfolio evaluation model (PEM)

PEM computes evaluation parameters for a given set of investment projects $(x_1, \ldots, x_m)$. Since some of the model parameters are given as fuzzy numbers, thus, the value of the $fin()$ function is a fuzzy variable, which results from the simulation combined with non-linear programming.

In order to evaluate a portfolio of investments, a mathematical model of an enterprise is built. Mathematical model consists of two groups of equations. First group of equations includes balances of the enterprise manufacturing capacities and material balances. It allows to determine size of the total production and size of sale achieved by enterprise. It determines also conditions of the selection of projects to be implemented. These conditions result form manufacturing capacities balance, material balances and availability of capital allocated for investments. The second group of equations are financial equations.

- equations of manufacturing capacities balance for primary production departments

$$\sum_{i \in I} X_{ijw}^{t\tau} \leqslant v_{jw}^{\varsigma} \cdot \Delta_{jw}^{\tau}, \tag{1}$$

where $\tau = 0, \ldots, \bar{\tau}, \tau \leqslant t, j \in J, w \in W_j$ and $t = \tau, \ldots, \tau + \bar{t}_{jw}$.

$$X_{ijw}^{t\tau} \geqslant 0\,,\ \varsigma = t - \tau \tag{2}$$

$$\Delta_{jw}^{\tau} = \begin{cases} 1 & \text{for } w \in \bar{W} \\ 0 & \text{for } w \in W - \bar{W} \end{cases} \tag{3}$$

$$\kappa\left(\bar{W}\right) = 1, \tag{4}$$

$$\eta^{\tau}\left(\bar{W}\right) \leqslant \bar{\eta}^{\tau},\ \tau = 0, 1, \cdots, \bar{\tau} \tag{5}$$

- equations of the enterprise material balance

$$\sum_{j \in J} \sum_{w \in W_j} \sum_{\substack{\tau=1 \\ \tau \leqslant t}}^{\bar{\tau}} X_{ijw}^{t,\tau} - \sum_{j \in J} \sum_{w \in W_j} \sum_{z \in I} \sum_{\substack{\tau=1 \\ \tau \leqslant t}}^{\bar{\tau}} m_{izjw} X_{zjw}^{t,\tau} = G_i^t \tag{6}$$

$$G_i^t \leqslant \bar{g}_i^t\left(\bar{W}\right), \tag{7}$$

where:

- $X_{ijw}^{t\tau}$ - quantity of the gross output of product i produced in j department in t year, in case of qualifying to realization project w in τ year,
- G_i^t - size of sale of the product i in year t,
- KRK^t - value of short-term credit in year t,
- KRD^t - value of long-term credit in year t,
- ZB^t - gross profit in year t,
- I - set of product indexes,
- I_j - set of indexes of products produced in j department,
- W - set of project indexes,
- W_j - set of indexes of projects connected with j department,
- $\bar{W}$ - set of indexes of projects qualified to realization,
- J - set of primary production department indexes
- v_{jw}^{ς} - manufacturing capacity of the j department after realization of w project in ς year of the duration
- $\bar{\eta}^{\tau}$ - limit of investment outlays in the τ year,
- m_{izjw} - consumption per unit of the i product for producing the z product in the j department after realizing the w project,
- $\bar{t}_{jw}$ - duration of the w project being realized in j department
- c_i^t - selling price for product i in year t
- kz_{ijw}^{ς} - variable cost of processing the product i by department j after realization of w project in ς year of the duration
- r_d - long-term interest rate
- r_k - short-term interest rate
- $\bar{\kappa} : 2^W \rightarrow \{0, 1\}$ - function determining sets of projects being possible for realization, value 1 means a set possible to realization, value 0 means set impossible to realization,
- $\eta^{\tau} : 2^W \rightarrow R$ - function assigning to $\bar{W}$ set of the projects an investment outlay for realization of this set in τ year of capital budgeting period
- $\bar{g}_i^t : 2^W \rightarrow R$ - function assigning to $\bar{W}$ set of the projects possible sale of the product in the t year

The second set of equations of the model are financial equations. They are linear equations, which for all the

above-mentioned parameters determined by equations (1)-(7), determine specific items of the company's balance sheet, P&L account and cash flows (NCF) used to calculate the NPV. As an example, an equation for calculating a company's gross profit is presented below.

$$ZB^t = P - C_1 - C_2 - C_3, \tag{8}$$

where $i \in I;\ t = 0, 1, 2, \cdots, T$ and

$$P = \sum_{j \in J} \sum_{i \in I_j} c_i^t G_i^t \tag{9}$$

$$C_1 = \sum_{\tau=1, \tau \leqslant t}^{\bar{\tau}} \sum_{w \in W_j} \sum_{j \in J} \sum_{i \in I} k_{ijw}^{\varsigma} X_{ijw}^{t,\tau} \tag{10}$$

$$C_2 = r_k K R K^t + r_d K R D^t \tag{11}$$

$$C_3 = \chi^t \left(\bar{W}\right) - \xi^t \ \left(\bar{W}\right) \tag{12}$$

The remaining financial equations express commonly known dependencies. A detailed presentation on them would considerably increase the volume of the article. Therefore, it is omitted.

In the next step, an appropriate model of uncertainty is assigned for every parameters. In the proposed framework, material consumption and product cost are characterized by fuzzy numbers. Demand and selling prices are described by probability distributions. Then, fuzzy simulation is employed, which allows different representations of uncertainty to be processed according to their nature. Moreover, the proposed framework takes into account economic dependencies in the process of selection of an effective portfolio of investment projects. Statistical dependency is used for describing relation between model parameters. Dependency between parameters characterized by fuzzy numbers are described by interval regression. Interval regression is an extension of the classical (crisp) regression where regression parameters are bounded closed intervals. For probabilistic parameters their dependency is determined by the correlation matrix. To process them, a method presented by Yang [30] based on Cholesky decomposition of the correlation matrix is utilized.

D. Procedure of determining portfolio evaluation model

The proposed procedure of determining the effectiveness of investment portfolio consists of two stages. It combines the procedure of stochastic simulation with execution of arithmetic operations on interactive fuzzy numbers. To execute such arithmetic operations non-linear programming is used. Computation procedure in this case is the following. Random variable values are drawn from among mentioned above parameters expressed in the form of the probability distribution. The procedure of generation accounts statistical dependency between variables. These values and remaining parameters expressed in the form of fuzzy numbers allow to determine evaluation parameter as fuzzy number. The problem of determining the fuzzy number characterizing evaluation parameter may be written owing to use of the concept of α-levels of fuzzy sets. Thus, the variables y corresponding to the parameters that are expressed in the form of fuzzy numbers are introduced, and then the parameters are replaced for those variables. Additionally, the following constraints are imposed:

$$\inf\left(\tilde{Y}_i\right)_\alpha \leq y_i \leq \sup\left(\tilde{Y}_i\right)_\alpha \tag{13}$$

$$y_i \geq \inf\left(a_1^{iz}\right) \cdot y_z + \inf\left(a_2^{iz}\right) \tag{14}$$

$$y_i \leq \sup\left(a_1^{iz}\right) \cdot y_z + \sup\left(a_2^{iz}\right) \tag{15}$$

where

- $\inf\left(\tilde{Y}_i\right)_\alpha$, $\sup\left(\tilde{Y}_i\right)_\alpha$ - respectively lower and upper bounds of α-level of the fuzzy parameter $\tilde{Y}_i$
- $\sup\left(a_1^{iz}\right)$, $\inf\left(a_1^{iz}\right)$, $\sup\left(a_2^{iz}\right)$, $\inf\left(a_2^{iz}\right)$ - respectively lower and upper bounds of interval regression coefficients describing dependency between parameters $\tilde{Y}_z$ and $\tilde{Y}_i$

Next, in order to determine the lower and upper bounds of the respective α-level of the efficiency parameter, the following constrained optimization problems must be solved:

$$NPV_\alpha \longrightarrow \min \tag{16}$$

for the definition of the lower bound of the α-level of the NPV,

$$NPV_\alpha \longrightarrow \max \tag{17}$$

for the definition of the upper bound of the α-level of the NPV. Drawing probabilistic values and determining NPV is repeated $n_{\max}$ times. As result $n_{\max}$ fuzzy sets characterized by membership functions $(\mu_1^{NPV}, \ldots, \mu^{NPV})$ are obtained and thus NPV is represented by a random fuzzy variable. Based on the vector $(\mu_1^{NPV}, \ldots, \mu^{NPV})$, the mean value, standard deviation as well as lower and upper cumulative distributions for the NPV are calculated. The hybrid procedure which implements the described approach is presented in the following algorithm.

Algorithm 1 Procedure of determining evaluation model

$n \leftarrow 1$;
2: Randomly generate vector probabilistic variables taking into account the correlation between them
$\alpha = 0$;
4: Define α-levels for fuzzy variables defining efficiency parameter
Calculate (sup) and (inf) for defined α-levels by finding: $eff_\alpha \leftarrow \min$ and $eff_\alpha \leftarrow \max$ under the problem constrains specified by constraints
6: $\alpha = \alpha + \phi$
If $\alpha \leqslant 1$ goto **STEP 4** else $n = n + 1$
8: if $n \leqslant n_{\max}$ goto **STEP 2**
Calculate mean value, standard deviation, and lower and upper cumulative distributions of the NPV.

V. NUMERICAL EXAMPLE

The capital budget was determined for the production process presented on the fig. 1. This setup includes the production cycle in steel industry, from production of the pig iron, production of steel, hot rolling products to production products coated with metal and plastics.

We take into consideration five investment projects: steel making plant, hot rolled sheet mill, cold-rolled sheet mill, hot-dip galvanizing sheet plant, sheet organic

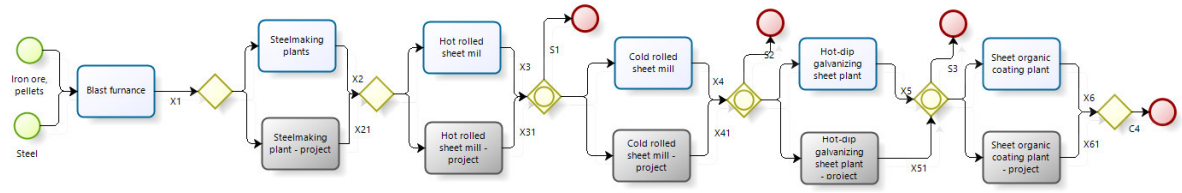


Fig. 1. Diagram of the analyzed technological setup

coating plant. Those investments are highlighted in gray. Decisive variables for the estimation of efficiency and the risk of investment projects in case of the investment in iron metallurgy are: quantity and selling prices, costs of materials and quantity of investment outlays. It was recognized also from here, that in the simulation experiment is necessary to take into consideration the uncertainty of the possible quantity of sale for each of products ranges being produced by the company, prices of these products, prices of metallurgic raw materials (prices of iron ores and the pellets), consumption per unit indexes, quantity of investment outlays. It was assumed, that remaining parameters of the efficiency calculus were determined. Prices of individual assortments of metallurgic products and metallurgic raw materials are correlated strongly. Similarly, sale quantities of each assortment of metallurgic products are correlated. This fact was taken into consideration when processing the values of efficiency calculus uncertain parameters.

In the computational experiment it was taken into consideration the uncertainty of the possible quantity of sales for each of products ranges being produced by the company, prices of these products, prices of semi-finished steel (prices of continuous casting stands), investment outlay for projects, construction period for investment projects and consumption per unit indexes for all products. Sales and possible quantity of sale for each product are described be probability distribution (in this case normal distribution). Rest of parameters were presented as a triangular fuzzy numbers.

In numerical example, we identify two types of dependencies. Prices of individual assortments of metallurgic products and metallurgic raw materials are correlated. Also sale quantities of each assortment of metallurgic products are correlated. Those parameters are using to describe benefit interdependency. Equations of manufacturing capacity balance are using for describing resource dependency and technical dependency.

The following new investment projects are considered: steel mill, rolling mill of cold milled steel sheets, hot dip galvanized coating, organic coating, rolling mill of hot milled steel sheets. Material consumption as well as product and half-product prices are given in the form of fuzzy numbers. They are presented, respectively, in Table I and Table II.

Sale parameters are given by normal probability distributions given in Table III.

TABLE I
TRAPEZOIDAL FUZZY NUMBERS (TFN) INDICATING MATERIAL CONSUMPTION

Material consumption	**TFN**
steel half-products – molten iron	$(0.855, 0.860, 0.870, 0.875)$
half-products – hot rolled steel sheets	$(1.058, 1.064, 1.075, 1.078)$
hot rolled steel sheets – cold rolled sheets	$(1.105, 1.111, 1.124, 1.130)$
cold rolled sheets – dip galvanized sheets	$(1.010, 1.020, 1.026, 1.031)$
dip galvanized sheets – organic coated sheets	$(0.998, 0.999, 1.000, 1.001)$

TABLE II
TRAPEZOIDAL FUZZY NUMBERS (TFN) FOR PRICES

Price	**TFN (USD/t)**
iron ore	$(335, 360, 400, 425)$
lumps	$(375, 400, 440, 470)$
steel scrap	$(940, 960, 1010, 1035)$
hot rolled sheets	$(2040, 2080.8, 2177.7, 2228.7)$
cold rolled sheets	$(2220.08, 2266.65, 2370.15, 2427.08)$
hot dip galvanized sheets and strips	$(2535.75, 2588.25, 2709, 2772)$
organic coated sheets and tapes	$(3450.6, 3519.82, 3684.9, 3754.13)$

TABLE III
PROBABILITY DISTRIBUTIONS INDICATING SALE PARAMETERS

Sale	**Mean value**	**Std. dev.**
hot rolled sheets	4704.0	117.5
cold rolled sheets	2750.0	51.4
hot dip galvanized – sheets and tapes	1147.9	52.4
organic coated – sheets and tapes	708.4	30.8

The Cholesky matrix which describes the dependencies between sale parameters is given by the equation (18).

$$\begin{pmatrix} 1.00000 & 0.87786 & 0.91142 & 0.86321 \\ 0.00000 & 0.47891 & 0.24007 & 0.27276 \\ 0.00000 & 0.00000 & 0.33418 & 0.34165 \\ 0.00000 & 0.00000 & 0.00000 & 0.25249 \end{pmatrix} \tag{18}$$

For the computational example, the α-level for fuzzy variables are set at 10 and the number of simulation are set to 100. The result for the computational example is shown in Fig. 2.

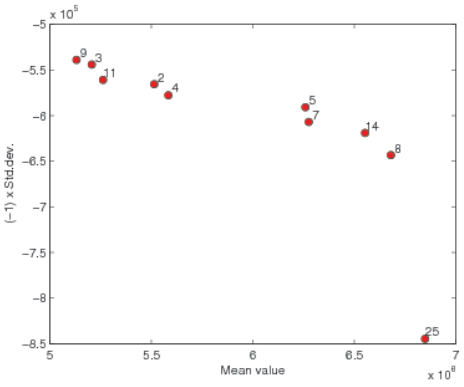


Fig. 2. Pareto optimal solutions for the problem of selection of an efficient portfolio of investment projects

VI. Conclusions

Above, the new method of choice of the effective portfolio of investment projects was presented. Presented concept of the mathematical model and the algorithm elaborated are making it possible to generate the set of Pareto-optimal solutions. The method is allowing for flexible formulating of dependence between projects. It concerns to dependencies with the technical character, how for example mutual excluding of projects. The structure of the model is causing, that dependencies with the economic character are also taken into account, it means, that projects are substitutionary or complementary in relation to themselves. Elaborated model has been utilized for selection of projects for the chosen production setup. The model is generating the set of Pareto-optimal solutions, which may to be the subject of the further analysis with taking extra criteria of quality into account.

References

[1] J. April, F. Glover, J. P. Kelly, "OptFolio - A Simulation Optimization System For Project Portfolio Planning", *in Proceedings of the 2003 Winter Simulation Conference*, vol.1, 2003, pp. 301–309. (DOI:10.1109/WSC.2003.1261437)

[2] N. P. Archer, F. Ghasemzadeh. "An integrated framework for project portfolio selection", *International Journal of Project Management*, vol. 17(4), 1999, pp. 207–216. (DOI:10.1016/S0263-7863(98)00032-5)

[3] M. A. Badri, D. Davis, D. Davis, "A comprehensive 0-1 goal programming model for project selection", *International Journal of Project Management*, vol. 19(4), 2001, pp. 243–252. (DOI:10.1016/S0263-7863(99)00078-2)

[4] C. Baudrit, D. Dubois, D. Guyonet, "Joint Propagation and Exploitation of Probabilistic and Possibilistic information in Risk Assessment", *IEEE Transaction on Fuzzy Systems*, vol. 14, 2006, pp. 593–607. (DOI:10.1109/TFUZZ.2006.876720)

[5] R. H. Bernhard, "Mathematical programming models for capital budgeting.-survey, generalization and critique", *J Financ Quant Anal*, vol. 4, 1969, pp. 111–158. (DOI:10.2307/2329837)

[6] S. P. Bradley, S. C. Frey, "Equivalent Mathematical Programming Models of Pure Capital Rationing", *J Financ Quant Anal*, vol. 6, 1978, pp. 345–361. (DOI:10.2307/2330391)

[7] W. T. Carleton, "Linear programming and Capital Budgeting Models: A New Interpretation", *Journal of Finance*, vol. 23, 1974, pp. 825–833. (DOI:10.1111/j.1540-6261.1969.tb01695.x)

[8] J. A. Cooper, S. Ferson, L. Ginzburg, "Hybrid processing of stochastic and subjective uncertainty data", *Risk Analysis*, vol. 16, 1996, pp. 785–791. (DOI:10.1111/j.1539-6924.1996.tb00829.x)

[9] P. K. De, D. Acharaya, K. C. Sahu, "A Chance-Constrained Goal Programming Model for Capital Budgeting", *Journal for the Operational Research Society*, vol. 33, 1982, pp. 635–638. (DOI:10.2307/2581726)

[10] M. W. Dickinson, A. C. Thomton, S. Graves, "Technology portfolio management. Optimizing interdependent projects over multiple time period", *IEEE Transaction on Engineering Management*, vol. 48(4), 2001, pp. 518–527. (DOI:10.1109/17.969428)

[11] S. Ferson, L. R. Ginzburg, "Difference method are needed to propagate ignorance and variability", *Reliab Eng Syst Safe*, vol. 54, 1996, pp. 133–144. (DOI:10.1016/S0951-8320(96)00071-3)

[12] H. Eilat, B. Golany, A. Shtub, "Constructing and evaluating balanced portfolios of R&D projects with interactions: A dea based methodology", *Eur J Oper Res*, vol. 172(3), 2006, pp. 1018–1039. (DOI:10.1016/j.ejor.2004.12.001)

[13] D. Guyonnet, B. Bourgine, D. Dubois, H. Fargier, B. Cme, P. J. Chils, "Hybrid Approach for addressing uncertainty in risk assessment", *Journal of Environmental Engineering*, vol. 126, 2003, pp. 68–76. (DOI:10.1061/(ASCE)0733-9372(2003)129:1(68))

[14] D. L. Hall, A. Nauda, "An interactive approach for selecting IR&D projects", *IEEE Trans. Eng. Management*, vol. 37(2), 1990, pp. 126–133. (DOI:10.1109/17.53715)

[15] X. Huang, "Credibility-based chance-constrained integer programming models with fuzzy parameters", *Information Sciences*, vol. 176(18), 2006, pp. 2698–2712. (DOI:10.1016/j.ins.2005.11.012)

[16] X. Huang, "Fuzzy chance-constrained portfolio selection", *Applied Mathematics and Computation*, vol. 177(2), 2006, pp. 500–507. (DOI:10.1016/j.amc.2005.11.027)

[17] J. P. Ignazio, "An approach to the Capital Budgeting Problem with Multiple Objectives", *The Engineering Economist*, vol. 21, 1976, pp. 259–272. (DOI:10.1080/00137917608902798)

[18] C. Kahraman, D. Ruan, C. E. Dozdag, "Optimization of Multilevel Investments Using Dynamic Programming Based on Fuzzy Cash Flows", *Fuzzy Optimization and Decision Making*, vol. 2(2), 2003, pp. 101–122. (DOI:10.1023/A:1023443116850)

[19] B. Liu, K. Iwamura, "Chance constrained programming with fuzzy parameters", *Fuzzy Sets and Systems*, vol. 94(2), 1998, pp. 227–237. (DOI:10.1016/S0165-0114(96)00236-9)

[20] J. H. Lorie, L. J. Savage, "Three problems in capital rationing", *Journal of Business* vol. 28, 1955, pp. 229–239. (DOI:10.1086/294081)

[21] P. Lusztig, B. Schwab, "A Note of the Application of Linear Programming to Capital Budgeting", *J Financ Quant Anal*, vol. 3, 1968, pp. 427–431. (DOI:10.2307/2329582)

[22] H. M. Markowitz, *Portfolio Selection Efficient Diversification of Investment*, Wiley, New York, 1959.

[23] A. L. Medaglia, S. B. Graves, J. L. Ringuest, "A multiobjective evolutionary approach for linearly constrained project selection under uncertainty", *Eur J Oper Res*, vol. 179(3), 2007, pp. 869–894. (DOI:10.1016/j.ejor.2005.03.068)

[24] D. L. Olson, *Decision Aids for Selection Problems*, New York, Springer Series in Operations Research, 1996.

[25] B. Rębiasz, "Selection of efficient portfoliosprobabilistic and fuzzy approach, comparative study", *Computers & Industrial Engineering*, vol. 64(4), 2013, pp. 1019–1032. (DOI:10.1016/j.cie.2013.01.011)

[26] N. E. Seitz, *Capital Budgeting and Long-Term Financing Decisions*. 3rd ed. USA, Dryden Press, 1999.

[27] B. Rębiasz, "Fuzziness and randomness in investment project risk appraisal", *Computers and Operations Research*, vol. 34, 2007, pp. 199–210 (DOI:10.1016/j.cor.2005.05.006).

[28] B. Rębiasz, B. Gaweł, I. Skalna, "Fuzzy multi-attribute evaluation of investments", *in Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2013.

[29] R. Santhanam , G. J. Kyparisis, "A decision model for interdependent information system project selection", *European Journal of Operational Research*, vol. 89(2), 1996, pp. 380–399. (DOI:10.1016/0377-2217(94)00257-6)

[30] I.-T. Yang, "Simulation-based estimation for correlated cost elements", *International Journal of Project Management*, vol. 23(4), 2005, pp. 275–282. (DOI:10.1016/j.ijproman.2004.12.002)

[31] A. Zuluaga, J. Sefair, A. Medaglia, "Model for the Selection and Scheduling of Interdependent Projects", *in Proceedings of the 2007 Systems and Information Engineering Design Symposium*, University of Virginia, 2007. (DOI:10.1109/SIEDS.2007.4374020)

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1123–1130
DOI: 10.15439/2014F346
ACSIS, Vol. 2

Analysis of Aggregated Bot and Human Traffic on E-Commerce Site

Grażyna Suchacka
Opole University
ul. Oleska 48,
45-052 Opole, Poland
Email: gsuchacka@uni.opole.pl

***Abstract*—A significant volume of Web traffic nowadays can be attributed to robots. Although some of them, e.g., search-engine crawlers, perform useful tasks on a website, others may be malicious and should be banned. Consequently, there is a growing need to identify bots and to characterize their behavior. This paper investigates the share of bot-generated traffic on an e-commerce site and studies differences in bots' and humans' session-based traffic by analyzing data recorded in Web server log files. Results show that both kinds of sessions reveal different characteristics, including the session duration, the number of pages visited in session, the number of requests, the volume of data transferred, the mean time per page, the number of images per page, and the percentage of pages with unassigned referrers.**

I. Introduction

ALONG with the growing popularity of search engines and other Web-based applications there has been the growing need to develop advanced tools for retrieving information on the Web content, structure, and usage. Such tools are Web bots (also called Web robots, spiders, or crawlers). They can traverse the Web autonomously by following the structure of hyperlinks, collect different kinds of information, and perform specific tasks on websites.

The most common bots are search engine crawlers, which visit Web pages on a regular basis to build and maintain huge search indexes [1], [2]. Popular bots visiting e-commerce sites are shopping bots, which collect information on products in various Web stores on behalf of product search engines or price comparison services. SEO spybots and content scrapers can expeditiously scrape from websites large amounts of information which may be valuable for SEO professionals or competitive e-business companies [3]. Other examples of robots include resource archivers, link checkers, e-mail harvesters, chat bots [4], spambots [5], hacking bots, artificial actors in e-dating [6], or automatic online game players [7].

Robot traffic on a website should be identified and sometimes also banned for several reasons. The most obvious ones are connected with potential threats of malicious bot activities [8], [9], [10]. Bot-generated click frauds in pay-per-click advertising result in higher fees paid by the advertisers [11]. Content-stealing bots may gather valuable business intelligence knowledge from websites and thus indirectly harm e-business competitiveness. High bot activity may also negatively affect the position of a website in search-engine rankings. Besides, bots consume network bandwidth and server resources; thus, they may cause degradation of the server performance and the quality of service offered to human users. Especially dangerous are automated DoS (Denial of Service) attacks, which may even make the server stall or crash [9]. Lastly, identification of robot traffic is essential when analyzing behavior of human users, who are characterized by different navigational patterns than bots [1], [8], [9], [12], [13], [14], [15].

Although some studies have addressed the problem of robot traffic characterization based on Web server logs, very little research has been done for e-commerce sites ([16], [17], [18]). Our study aims to partially fill this gap by comparing key characteristics of bot- and human-generated traffic on a Web bookstore site. This issue is crucial for e-commerce sites where human users are potential buyers and their activity on the site is directly related to the profitability of the online store.

The paper is organized as follows. Section II presents Web server log data underlying our research and the research methodology. Section III discusses the share of bots in the analyzed Web traffic, whereas Section IV compares key characteristics of bot and human sessions. Section V concludes the paper and suggests prospective future work.

II. Research Methodology

A. Web Server Log Data Description

When an Internet user visits a website, their Web browser (which is a Web client, in fact) communicates via the HTTP protocol with the server hosting the site. For each Web page requested by the user, their client typically issues a series of HTTP requests to the server: one request for a page description file and the following requests for objects embedded in the page, such as images or video files. After

978-83-60810-58-3/$25.00

receiving HTTP responses the client assembles the page and displays it in a browser window. A Web client may represent not only a human user but it may also be a computer program, i.e. a Web bot.

Data concerning each incoming HTTP request is recorded in the access log file stored at the Web server. That data includes some client data (a client IP address, a client identifier, a user agent field, a user identifier, a referrer field), the requested resource data (an URI identifying the requested server resource, a transfer size), the HTTP-related data (a method, a protocol version, a status code), and a timestamp. As an example, let us consider the following log entry, representing one HTTP request:

```
66.249.66.52 - - [03/Dec/2013:08:55:59 +0100] "GET shopping/images/pict21.jpg HTTP/1.1" 200 242 "-" "Mozilla/5.0 (compatible;Googlebot/2.1;+http://www.google.com/bot.html)".
```

This line describes a request sent by a Web client with the IP address 66.249.66.52, whose user identifier is not available. The request was served 3 December 2013 at 8:55:59 (according to Central European Time) and it concerned downloading (by using the GET method) an image file identified by URI "shopping/images/pict21.jpg". The request was successfully served (a status code is 200) and the server sent to the client 242 bytes in response. A referrer field is unassigned. The client was Mozilla 5.0 which used the protocol HTTP/1.1. One can notice that the user was not a human but Google's web crawling bot (the user agent field contains the bot's name, "Googlebot").

Our analysis was based on access logs for an online store (the store name is not given in the paper due to a non-disclosure agreement). The data covered the period of one month, December 2013.

A dedicated computer program was used to read, preprocess, clean, and analyze the data. The program was implemented in C++ using MS Visual Studio. Its most important modules include:

- Input/Output Module containing functions for reading raw data from the input log files and saving the results to the output files;
- Basic Functions Module with functions for parsing each HTTP request's line in order to distinguish individual data describing the request and transform it to the format suitable for the analysis;
- Request Module for managing and processing HTTP requests, e.g., checking whether a request was generated by bot;
- Session Module for reconstructing and processing user sessions;
- Robot Module for identifying and processing sessions generated by bots;
- Statistics Module containing functions for computing all the necessary statistics;
- other modules implementing the operation of visual forms.

B. Reconstruction and Characterization of User Sessions

Based on HTTP requests user sessions were reconstructed. A `user session` means a sequence of requests issued by a Web client during the single visit to the Web store. Each individual user was identified based on two data fields describing HTTP requests: the client IP address and the user agent field. Consecutive user sessions were reconstructed based on the requests' timestamps, assuming a minimum 30-minute interval between two subsequent sessions of a given user (the value of 30 minutes has been commonly applied in previous Web traffic analyses, e.g. in [9], [19]).

Afterwards, each user session was described with a number of attributes:

- session length – the number of pages visited in session;
- session duration – time interval (in seconds) between the times of the last and the first requests in session (session duration is shorter than the actual time of the user-site interaction because the time of browsing the last page in session by the user in unknown at the server side; for the same reason this attribute cannot be determined for sessions containing only one page);
- mean time per page – the average time (in seconds) the user browsed a single page in session (this attribute may be derived only for sessions containing more than one page);
- volume of data transferred to the Web client (in MB);
- number of HTTP requests;
- image-to-page ratio – the average number of image file requests over the number of page requests in session;
- percentage of pages with unassigned referrers – the percentage of page requests with unassigned or blank referrer fields;
- percentage of requests with unassigned referrers – the percentage of HTTP requests with unassigned or blank referrer fields;
- percentage of requests of type HEAD – the percentage of HTTP requests with HEAD method;
- percentage of 4xx responses – the percentage of erroneous HTTP requests in session (i.e. requests with status codes starting with "4").

We decided to compute the aforementioned attributes because some previous user session analyses for non-e-commerce environments reported that these session features may be useful in distinguishing Web robots from human users [1], [8], [9].

Some sessions contained no Web page request and only one request for an image file (such a situation is often connected with displaying a banner advertisement of the store on another Web page). As these sessions cannot be regarded as intended visits to the store, we did not take them into consideration in our analysis.

C. Identification of Bot Sessions

There are a few ways to identify at least some part of user sessions issued by Web robots.

First, one should check if the file "robots.txt" was accessed in a session. Cooperative robots should request this file at the beginning of each visit to a site in order to read which parts of the site they can access.

Second, "ethical" bots should inform a Web server about their identities via their user agent fields, containing the name of the robot. We implemented a function verifying HTTP requests' user agent fields for compliance with user agents of known robots, available on online databases [20] and [21]. Moreover, some robots not included in these databases were identified based on keywords contained in user agent fields ("bot", "spider", "crawler", "worm", "search", "track", "harvest", "dig", "hack", "trap", "archive", or "scrap"), as well as through a semi-automatic inspection of user agent fields.

In practice, not all robots access the file "robots.txt" or declare their identities in user agent fields. However, some of such bots may be still identified based on the character of their interaction with the site, which proceeds differently from the interaction of human users. Humans usually communicate with the site via the Web interface and follow navigation paths according to the site topology. Each Web page request is typically followed by a group of requests for embedded objects (usually images). Moreover, the successive page requests are separated with some time intervals called "user think times". In contrast, robots tend to reveal navigational patterns incompatible with the site topology and have unintuitive session characteristics, e.g., the extremely low mean time per page. We assumed the following three groups of session characteristics that indicate Web robots:

- the mean time per page shorter than 0.5 second;
- an unassigned referrer field in the first request in session, the percentage of pages or requests with unassigned referrers equal to 100, and the percentage of requests of type HEAD equal to 100;
- an unassigned referrer field in the first request in session, the percentage of pages or requests with unassigned referrers equal to 100, the percentage of 4xx responses equal to 100, and the image-to-page ratio equal to 0.

All sessions which were not classified as performed by robots, after excluding sessions connected with executing administrative tasks on the site, were assumed to be performed by human users.

III. Bot Share in Overall Web Server Traffic

According to the results of earlier analyses for e-business workloads the share of robot requests has differed from several (3.2% in [17]) to a dozen or so (15% in [22], 16% in [23]) percent. In our data set 22.3% of all HTTP requests were identified as generated by bots (Fig. 1). However, as regards the number of user sessions, as many as 79.3% were performed by bots. Bot sessions seem to contain on average less requests and consume less server resources than human sessions (the volume of data transferred to bot clients comprised 38% of the overall data transfer).

In regards to known bots, possible to recognize by checking requests' user agent fields, 76 different robots were identified. The most active of them were popular search engine crawlers (Bingbot, MJ12bot, Googlebot, Google AdsBot, Yandex bot, MSNBot, Baidu spider). Large part of bot traffic was also generated by SEO and e-commerce crawlers (AhrefsBot, ShopWiki, WillyBot), link checkers (SEOkicks robot, SpBot), and social media agent FacebookExternalHit.

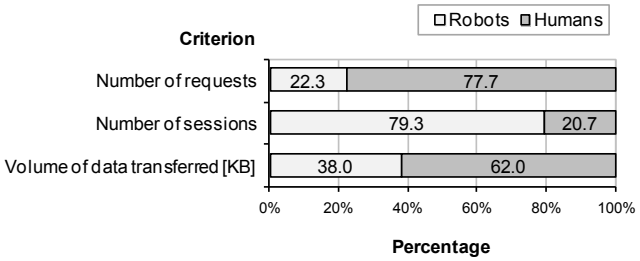


Fig. 1 Percentage breakdown of robots' and humans' data

It is worth noting that only 11% of all bot sessions (including 22.7% of known bots' sessions) accessed the file "robots.txt".

IV. Comparison of Bot and Human Sessions

An insight into the session characteristics revealed that a large number of all user sessions contained only one page and/or lasted only one second. One of the justifications of such user behavior may be that some users were referred to the store site by following a search engine link or by clicking a store advertisement placed on another site but the content of the store website was not what they had searched for. Such users left the site immediately after entering it. We decided to exclude from the statistical analysis sessions containing only one page and sessions lasting only one second.

Furthermore, using the graphical method, we excluded from the analysis a few outlying sessions which were extremely long and lasted for an extremely long time compared to other sessions. The outliers were two human sessions (containing 202 and 312 pages) and eight bot sessions (containing from 1 453 to 6 029 pages and/or lasting longer than 48 hours) – the reason for the exclusion of these several sessions was that they strongly distorted the statistical results for all human and bot sessions, respectively. Finally, only 27% of sessions were left in our dataset.

A. Session Length

An important aspect of user session characterization in the context of distinguishing bots from humans is the session length in the number of pages visited in session. Intuitively, there is some upper limit on the maximum number of human user's clicks, i.e. the number of pages a human user can open and browse during a single visit to a website. This limitation does not apply to automatic computer programs, such as bots, which are able to automatically traverse all pages belonging to a site in a relatively short time. For the same reasons the maximum time of a human-website interaction is limited as well, so bot-generated sessions tend to last much longer than human ones.

Our results achieved for the e-commerce site confirm these observations. Session length statistics presented in Table I show that robots requested on average above four times more pages in session than humans and the maximum session length was an order of magnitude higher for bots than for humans. (Taking into consideration the outlying sessions excluded from the analysis one can notice that the longest human session contained 312 pages whereas the longest bot session contained as many as 6 029 pages.) However, session length distributions presented in Fig. 2 are similar for robots and humans: both histograms illustrate a strong right-skew of session length distribution. Over 95% of human users opened less than 26 pages during their visits in the store and 97% of bots requested less than 176 pages.

B. Session Duration

Session duration statistics presented in Table II show that Web bots tend to spend much more time on the website than human users. The mean session duration is fifteen times longer for bots than for humans. The maximum session duration for bots is 39 hours (and up to 181 hours taking into consideration the excluded outlying bot sessions!) whereas for humans it is less than two hours. Distributions of session durations, shown in Fig. 3, are very similar to the distributions of session lengths in Fig. 2. For both kinds of sessions they are strongly right-skewed and heavy-tailed.

Intuitively, bigger numbers of pages in session should correspond to longer session durations so it is worth graphically examining this relationship. Fig. 4 presents a two-dimensional scatter plot of the session duration against the session length (to improve the graph readability robot sessions with lengths exceeding 800 pages were not shown in the figure). One can see the correlation between the number of pages visited in session and the duration of a visit for both kinds of session: as the session length increases, the session duration tends to increase as well. Human sessions form a quite well-knit group in the two-dimensional area whereas robot sessions are rather dispersed and seem to form a few (at least five) separate clusters. Fig. 4 suggests that different kinds of bots may reveal different behavior so it would make sense to separately characterize behavior of various bots (search engine crawlers, image indexers, link checkers, e-

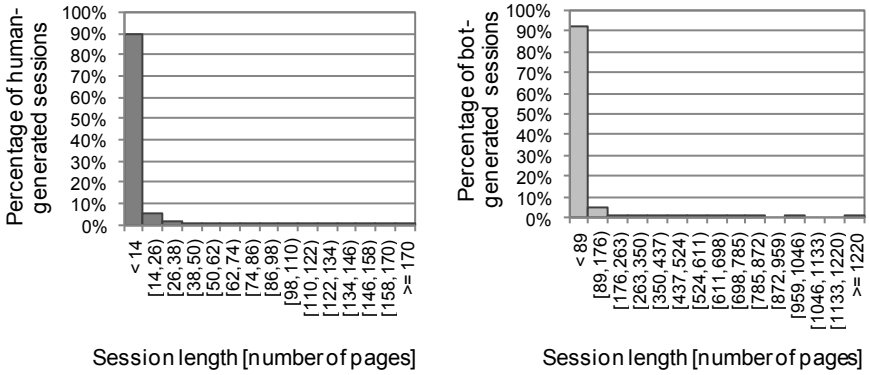


Fig. 2 Histogram of session lengths: (left) for humans, (right) for bots

TABLE I.
SESSION LENGTH STATISTICS
(IN NUMBER OF PAGES VISITED)

Statistics	Humans	Bots
Mean	7.2	30.1
Median	3	8
Mode	2	2
Std. dev.	13.1	73.7
Minimum	2	2
Maximum	173	1 253

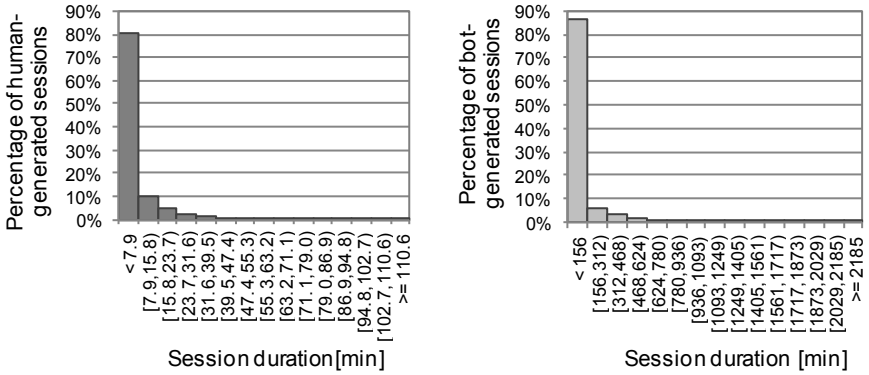


Fig. 3 Histogram of session durations: (left) for humans, (right) for bots

TABLE II.
SESSION DURATION STATISTICS
(IN MINUTES)

Statistics	Humans	Bots
Mean	5.4	82.6
Median	1.4	11.4
Mode	0.05	0.03
Std. dev.	10.2	213
Minimum	0.03	0.03
Maximum	1 18.4	2 341

mail collectors, etc.). One could also apply classification or clustering methods to determine classes or clusters of robot sessions. We leave these issues for our future work.

C. Mean Time per Page

Based on the session length and the session duration one can determine the mean time per page for each user session containing more than one page. The mean time per page in session was computed according to the formula:

$$\bar{d}_s = \frac{d_s}{l_s - 1} \quad (1)$$

where d_s is the session duration (in seconds) and l_s is the session length (in number of pages), $l_s > 1$. Unlike the session duration, which does not include the time for the last page visited in session, the mean time per page is not underrepresented as it is computed for all visited pages except one.

Mean time per page statistics, presented in Table III, show significant differences between bot and human sessions. It may be surprising that bots spend more time analyzing Web pages than human users and the average is equal to as much as 5.3 minutes. However, the median equal to 1.9 minute and the mode equal to 2 seconds are much lower. Besides, a relatively high value of the standard deviation, 8 minutes, indicates that the mean time per page is rather differentiated for bots. Histograms in Fig. 5 also show a bigger differentiation of mean times per page for bots than for humans. For robots the distribution of mean times per page is not so strongly right-skewed and does not include such a long tail as for human users. These results also indicate that it may be worth performing the statistical characterization of various kinds of robots visiting the Web store site.

D. Image-to-Page Ratio

Some previous studies reported that robots (especially crawlers) request mostly Web page files and ignore image files [8], [9], [24]. In contrast, human users navigate through the website following the structure of hyperlinks and when they open a new page, they usually download the page description file along with image files embedded in the page. Hence, such metrics as image-to-page ratio or percentage of image requests in session belong to strongly distinguishable characteristics between bots and humans. Image-to-page ratio statistics in Table IV, as well as histograms in Fig. 6 confirm these results. Humans request more than 22 images per page; the median is equal to 19 and the mode is equal to 36. In contrast, robot requests for image files are negligible: the mean number of images per page is only 0.3 and what is more, both the median and the mode are equal to 0. Among all robot sessions almost 50% did not request any image at all. Surprisingly, 0.6% of human sessions also contained no image request (it may indicate that some sessions considered as generated by humans were performed by bots, in fact).

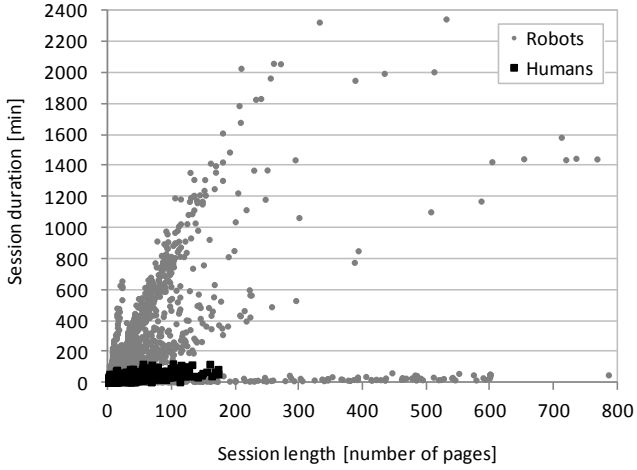


Fig. 4 Scatter plot of session duration vs. session length for robots and humans

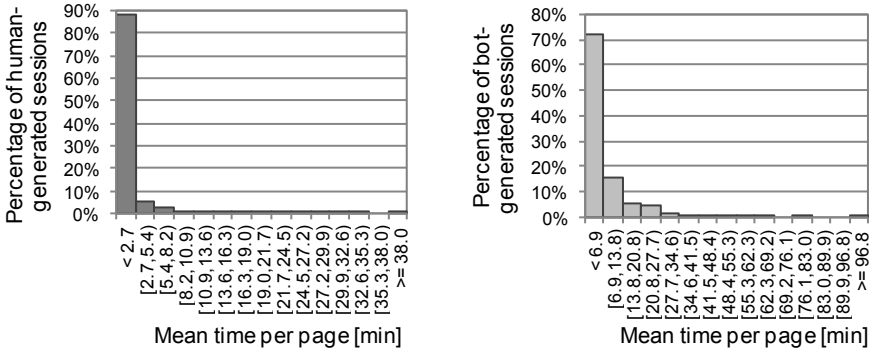


Fig. 5 Histogram of mean times per page: (left) for humans, (right) for bots

TABLE III.
MEAN TIME PER PAGE STATISTICS
(IN MINUTES)

Statistics	Humans	Bots
Mean	1.4	5.3
Median	0.4	1.9
Mode	0.1	0.03
Std. dev.	3.2	8.0
Minimum	0.01	0.001
Maximum	40.6	103.6

E. Volume of Data Transferred to Web Clients

We decided to examine if large numbers of image requests correspond to large volumes of data transferred to Web clients, i.e. whether data transfers are bigger for humans than for bots. Typically, a significant part of Web traffic concerns transmitting small files and messages (e.g. messages that the requested resource has not been modified or that resource could not be found on the server). In contrast, graphical and multimedia Web resources are relatively big files.

As can be seen in Table V, volumes of data transferred to human users tend to be much bigger than for robots. Although this metric in both cases ranges from 0 to over 14, distributions of data transfer volumes are quite different (Fig. 7). For robot sessions the histogram is extremely heavy-tailed. Average bot data transfer is 227 KB, however the median transfer is only 65 KB and the mode is merely 1 KB. For 98% of bot sessions the transmitted data did not exceed 1.7 MB. On the contrary, average data transfer for humans is 1.2 MB, the median is 765 KB and the mode is 263 KB. Data sent in 98% of human sessions were up to 5.1 MB.

It is interesting to observe that for both kinds of sessions the distributions of data transfer volumes (Fig. 7) correspond very accurately to the distributions of the numbers of HTTP requests in session (Fig. 8). Also the range of the numbers of HTTP requests in session is almost the same for bots and humans (Table VI).

F. Percentage of Pages with Unassigned Referrers

Users may reach the store website in many different ways, e.g. by following a search engine link (one of organic search engine results or sponsored links), or by clicking a banner ad on another website. In such cases an address of the referring

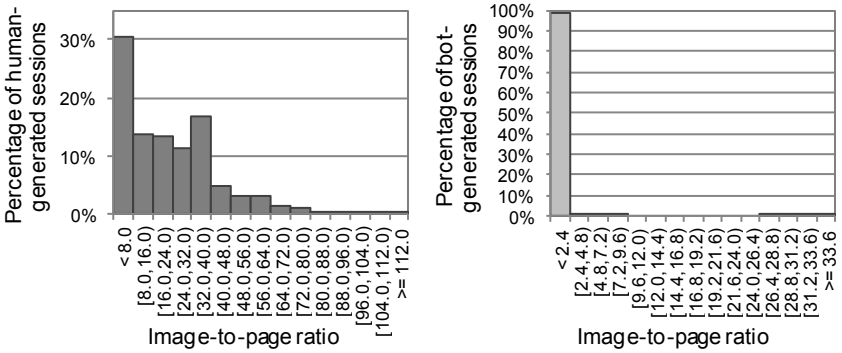


Fig. 6 Histogram of image-to-page ratios: (left) for humans, (right) for bots

TABLE IV.
IMAGE-TO-PAGE RATIO STATISTICS

Statistics	Humans	Bots
Mean	22.4	0.3
Median	19.1	0
Mode	36	0
Std. dev.	19	1.8
Minimum	0	0
Maximum	119.5	35.5

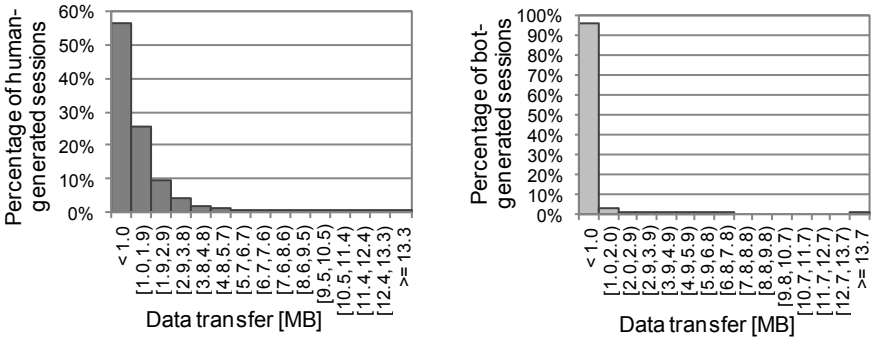


Fig. 7 Histogram of data transfer volumes: (left) for humans, (right) for bots

TABLE V.
DATA TRANSFER VOLUME STATISTICS (IN MEGABYTES)

Statistics	Humans	Bots
Mean	1.2	0.2
Median	0.7	0.1
Mode	0.3	0.001
Std. dev.	1.3	0.6
Minimum	0	0
Maximum	14.3	14.6

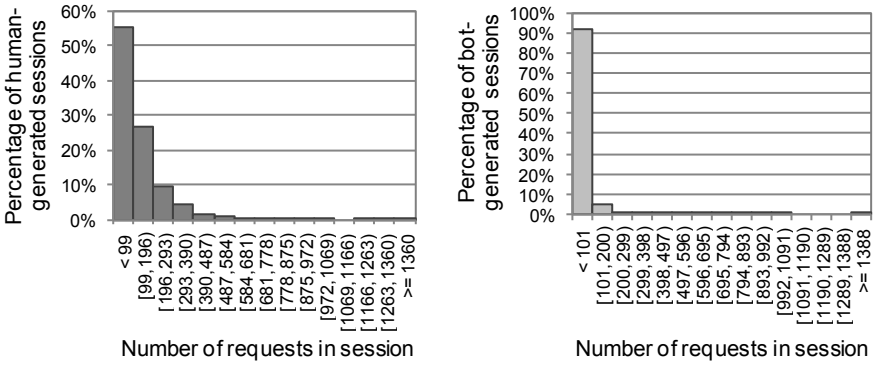


Fig. 8 Histogram of the numbers of HTTP requests in session: (left) for humans, (right) for bots

TABLE VI.
NUMBER OF REQUESTS IN SESSION STATISTICS

Statistics	Humans	Bots
Mean	120.9	35.8
Median	80	10
Mode	76	2
Std. dev.	130.4	80.8
Minimum	2	2
Maximum	1 450	1 484

page is contained in the referrer field of the first HTTP request in session. Sometimes this field may be empty, e.g. when a user enters the site directly by typing the site address in a browser's address bar or clicking on a bookmarked page. However, as a human user navigates through the website, each newly opened Web page request will contain in its referrer field the address of the previously browsed page. On the contrary, the vast majority of Web robots initiate their sessions (or even all HTTP requests in session) with unassigned referrer fields, so it may be a good indicator of a bot-generated session. This was confirmed in previous Web characterization studies, e.g. in [10] and [25].

We computed the percentage of page requests with unassigned referrer fields in each session. Our results, presented in Fig. 9 and Table VIII, are similar to observations reported in earlier studies. In fact, 98.4% of all robot sessions had all pages with unassigned referrers. For comparison, among human sessions there were only 2.1% of such sessions (it is very likely that they were actually unidentified bot-generated sessions).

G. Percentage of Requests of Type HEAD

The most common HTTP method is GET, which is used do download contents from Web servers. By default, when a human user browses a website via a browser, requests sent to the server by the browser will be of type GET. Other possible HTTP method is HEAD, used to retrieve only Web metadata. In contrast to humans, robots are expected to use HEAD method instead of GET when possible (e.g. to download only recently updated contents) in order to reduce the amount of data downloaded from servers and to minimize the consumption of server resources.

Some previous workload characterization studies showed that the percentage of requests of type HEAD is higher for bots than for humans [25]. Other studies reported that nearly all crawler requests were of type GET [24]. Our results signalize an advantage of bots over humans in this respect, however the percentage of requests of type HEAD for bot sessions was not very high (Table VII). Only 0.4% of bot sessions had some requests of type HEAD, compared to 0.1% of human sessions. After taking into consideration all robot sessions (even those containing only one request, excluded from our statistical analysis), the mean percentage of HEAD requests for bots increases to 0.8 and 0.9% of bots sessions contain only HEAD requests.

TABLE VII.
PERCENTAGE OF REQUESTS OF TYPE HEAD STATISTICS

Statistics	Humans	Bots
Mean	0.004	0.2
Median	0	0
Mode	0	0
Std. dev.	0.15	4.1
Minimum	0	0
Maximum	8.3	100

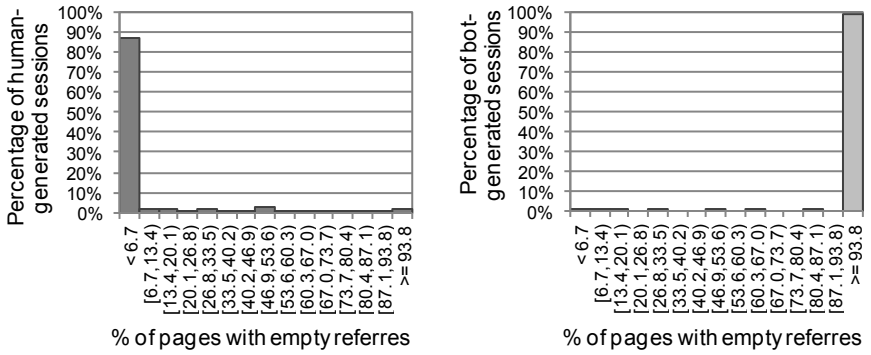


TABLE VIII.
PERCENTAGE OF PAGES WITH UNASSIGNED REFERRERS STATISTICS

Statistics	Humans	Bots
Mean	5.9	99
Median	0	100
Mode	0	100
Std. dev.	18.4	9.8
Minimum	0	0
Maximum	100	100

Fig. 9 Histogram of percentages of pages with unassigned referrer fields: (left) for humans, (right) for bots

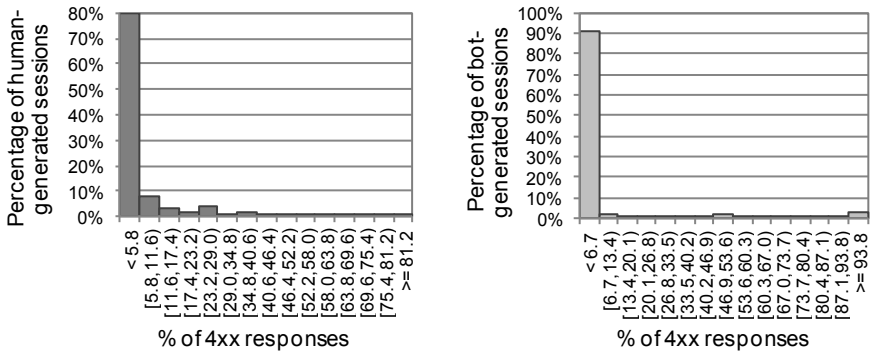


TABLE IX.
PERCENTAGE OF 4XX RESPONSES STATISTICS

Statistics	Humans	Bots
Mean	5	4.9
Median	1.3	0
Mode	0	0
Std. dev.	9	18.6
Minimum	0	0
Maximum	87	100

Fig. 10 Histogram of percentages of 4xx responses: (left) for humans, (right) for bots

H. Percentage of 4xx Responses

Web robots tend to have a higher rate of erroneous requests (i.e. requests with status codes of type 4xx), because it is more likely that they request outdated or deleted files [1], [24], [25]. However, we observed that for our data set the rate of erroneous requests was a little bit higher for human sessions (the mean equal to 5) than for bot ones (the mean equal to 4.9) (Table IX). For robots the mean is a bit lower but more variable. Moreover, 2.8% of bot sessions had 100% of erroneous responses (compared to 0% of such human sessions). After taking into consideration also sessions containing only one request, the statistics for humans increase insignificantly whereas for bots they increase notably: the mean is equal to 10.8, the standard deviation is equal to 22.5, and 3.8% of bot sessions have 100% of erroneous responses.

V. Conclusion

The paper discusses key characteristics of sessions realized by Web robots and human users on the e-commerce site. Our results confirm some earlier findings of Web workload analyses, concerning differences between bots and humans in the following session characteristics: the session duration, the number of pages visited in session, the number of HTTP requests in session, the volume of data transferred, the percentage of pages with unassigned referrers, and the number of images per page. However, such characteristics as the percentage of requests of type HEAD and the percentage of erroneous responses turned out not to be as good indicators of bot sessions as reported in previous studies.

The analysis was done for the aggregated Web robot traffic. However, our observations suggest that the behavior of bots is not homogenous and various kinds of bots may reveal different navigational patterns. In our future work we plan to address this issue. We also plan to extend our research to another Web stores of different sizes and branches to verify the reliability of our conclusions for other e-commerce scenarios. Our findings may be applied in classification and segmentation methods aiming at identifying sessions of unknown bots on the e-commerce website.

References

[1] A. Balla, A. Stassopoulou, M. D. Dikaiakos, "Real-time Web crawler detection," in Proc. 18th ICT, Ayia Napa, Cyprus, 2011, pp. 428-432, http://dx.doi.org/10.1109/CTS.2011.5898963.

[2] H. Kang, K. Wang, D. Soukal, F. Behr, Z. Zheng, "Large-scale bot detection for search engines," in Proc. 19th WWW, Raleigh, NC, USA, 2010, pp. 501-510, http://dx.doi.org/10.1145/1772690.1772742.

[3] N. Poggi, J. L. Berral, T. Moreno, R. Gavaldà, J. Torres, "Automatic detection and banning of content stealing bots for e-commerce", in Proc. NIPS Workshop on Machine Learning in Adversarial Environments for Computer Security, British Columbia, Canada, 2007, pp. 7-8.

[4] S. Gianvecchio, M. Xie, Z. Wu, H. Wang, "Humans and bots in Internet chat: measurement, analysis, and automated classification," IEEE/ACM Trans. Netw. 19(5), 2011, pp. 1557-1571, http://dx.doi.org/10.1109/TNET.2011.2126591.

[5] P. Hayati, V. Potdar, K. Chai, A. Talevski, "Web spambot detection based on Web navigation behaviour," in Proc. AINA, Perth, 2010, pp. 797-803, http://dx.doi.org/10.1109/AINA.2010.92.

[6] A. Schmitz, O. Yanenko, M. Hebing, "Identifying artificial actors in e-dating: a probabilistic segmentation based on interactional pattern analysis," Challenges at the Interface of Data Analysis, Computer Science, and Optimization - Studies in Classification, Data Analysis, and Knowledge Organization, 2012, pp 319-327, http://dx.doi.org/10.1007/978-3-642-24466-7_33.

[7] A. R. Kang, H. K. Kim, J. Woo, "Chatting pattern based game BOT detection: do they talk like us?," KSII TIIS 6(11), 2012, pp. 2866-2879.

[8] A. Stassopoulou, M. D. Dikaiakos, "Web robot detection: a probabilistic reasoning approach," Comput. Netw. 53(3), 2009, pp. 265-278, http://dx.doi.org/10.1016/j.comnet.2008.09.021.

[9] D. Stevanovic, N. Vlajic, A. An, "Unsupervised clustering of Web sessions to detect malicious and non-malicious website users," Procedia Computer Science 5, Elsevier, 2011, pp. 123–131, http://dx.doi.org/10.1016/j.procs.2011.07.018.

[10] P.-N. Tan, V. Kumar, "Discovery of web robot sessions based on their navigational patterns," Data Min. Knowl. Discov. 6 (1), 2002, pp. 9–35, http://dx.doi.org/10.1023/A:1013228602957.

[11] K. Springborn, P. Barford, "Impression fraud in online advertising via pay-per-view networks," in Proc. 22nd USENIX Conference on Security, Washington, D.C., 2013, pp. 211-226.

[12] S. Kwon, M. Oh, D. Kim, J. Lee, Y.-G. Kim, S. Cha, "Web robot detection based on monotonous behavior," ASTL 4, Springer-Verlag, 2012, pp. 43-48.

[13] C. H. Saputra, E. Adi, S. Revina, "Comparison of classification algorithms to tell bots and humans apart," JNIT 4(7), 2013, pp. 23-32.

[14] G. Suchacka, G. Chodak, "Practical aspects of log file analysis for e-commerce," CCIS 370, Springer, 2013, pp. 562-572, http://dx.doi.org/10.1007/978-3-642-38865-1_56.

[15] G. Suchacka, "Statistical analysis of buying and non-buying user sessions in a Web store," Information Systems Architecture and Technology – Network Architecture and Applications, Wroclaw, Poland, 2013, pp. 163-172.

[16] V. Almeida, D. Menascé, R. Riedi, F. Peligrinelli, R. Fonseca, W. Meira Jr., "Analyzing robot behavior in e-business sites," in Proc. ACM SIGMETRICS, Cambridge, MA, USA, 2001, pp. 338–339, http://dx.doi.org/10.1145/378420.378838.

[17] D. Doran, S. S. Gokhale, "Long range dependence (LRD) in the arrival process of Web robots," in Proc. ICCTS, New Delhi, India, 2012, pp. 176-180, http://dx.doi.org/10.7763/IPCSIT.2012.V47.33.

[18] N. Poggi, J. L. Berral, T. Moreno, R. Gavaldà, J. Torres, "Automatic detection and banning of content stealing bots for e-commerce," in Workshop on Machine Learning in Adversarial Environments for Computer Security, British Columbia, Canada, 2007.

[19] D. Doran, S. S. Gokhale, "Searching for heavy tails in Web robot traffic," in Proc. 7th Int. Conf. QEST, Williamsburg, Virginia, USA, 2010, pp. 282-291, http://dx.doi.org/10.1109/QEST.2010.42.

[20] List of user-agents (spiders, robots, crawler, browser), http://www.user-agents.org.

[21] List of user agent strings – Robots (crawlers), http://user-agent-string.info/list-of-ua/bots.

[22] N. Poggi, D. Carrera, R. Gavaldà, E. Ayguadé, J. Torres, "A methodology for the evaluation of high response time on e-commerce users and sales," Inform. Syst. Front., 2012, http://dx.doi.org/10.1007/s10796-012-9387-4.

[23] D. A. Menascé, V. Almeida, R. H. Riedi, F. Ribeiro, R. C. Fonseca, W. Meira Jr., "In search of invariants for e-business workloads," in Proc. 2nd ACM-EC Conf., Minneapolis, MN, USA, 2000, pp. 56-65, http://dx.doi.org/10.1145/352871.352878.

[24] M. D. Dikaiakos, A. Stassopoulou, L. Papageorgiou, "An investigation of Web crawler behavior: characterization and metrics," Comput. Commun. 28(8), 2005, pp. 880-897, http://dx.doi.org/10.1016/j.comcom.2005.01.003.

[25] C. Bomhardt, W. Gaul, L. Schmidt-Thieme, "Web robot detection - preprocessing Web logfiles for robot detection," New Developments in Classification and Data Analysis - Studies in Classification, Data Analysis, and Knowledge Organization, 2005, pp. 113-124, http://dx.doi.org/10.1007/3-540-27373-5_14.

12th Conference on Advanced Information Technologies for Management

We are pleased to invite you to participate in the 10th edition of Conference on "Advanced Information Technologies for Management AITM'2012". The main purpose of the conference is to provide a forum for researchers and practitioners to present and discuss the current issues of IT in business applications. There will be also the opportunity to demonstrate by the software houses and firms their solutions as well as achievements in management information systems.

Topics

The topics of interest include but are not limited to:

- Concepts and methods of business informatics
- Business Process Management and Management Systems (BPM and BPMS)
- Management Information Systems (MIS)
- Enterprise information systems (ERP, CRM, SCM, etc.)
- Business Intelligence methods and tools
- Strategies and methodologies of IT implementation
- IT projects & IT projects management
- IT governance, efficiency and effectiveness
- Decision Support Systems and data mining
- Intelligence and mobile IT
- Cloud computing, SOA, Web services
- Agent-based systems
- Business-oriented ontologies, topic maps
- Knowledge-based and intelligent systems in management

Event Chairs

Dudycz, Helena, Wrocław University of Economics, Poland

Dyczkowski, Mirosław, Wrocław University of Economics, Poland

Korczak, Jerzy, Wroclaw University of Economics, Poland

Program Committee

Abramowicz, Witold, Poznan University of Economics, Poland

Ahlemann, Frederik, University of Duisburg-Essen, Germany

Andres, Frederic, National Institute of Informatics, Tokyo, Japan

Brown, Kenneth, Communigram SA, France

Chmielarz, Witold, University of Warsaw, Poland

Cortesi, Agostino, Università Ca' Foscari, Venezia, Italy

Czarnacka-Chrobot, Beata, Warsaw School of Economics, Poland

De, Suparna, University of Surrey, Guildford, United Kingdom

Dufourd, Jean-François, University of Strasbourg, France

Franczyk, Bogdan, Universitat Leipzig, Germany

Kannan, Rajkumar, Bishop Heber College (Autonomous), Tiruchirappalli, India

Kersten, Grzegorz, Concordia University, Montreal, Poland

Kowalczyk, Ryszard, Swinburne University of Technology, Melbourne, Victoria, Australia

Ligęza, Antoni, AGH University of Science and Technology, Poland

Ludwig, André, University of Leipzig, Germany

Maciaszek, Leszek, Wroclaw University of Economics, Poland and Macquarie University ~ Sydney, Australia

Magoni, Damien, University of Bordeaux – LaBRI, France

Michalak, Krzysztof, Wroclaw University of Economics

Pankowska, Malgorzata, University of Economics in Katowice, Poland

Stanek, Stanisław, General Tadeusz Kosciuszko Military Academy of Land Forces in Wroclaw, Poland

Teufel, Stephanie, University of Fribourg, Switzerland

Tsang, Edward, University of Essex, United Kingdom

Zanni-Merk, Cecilia, Universite de Strasbourg, France

Ziemba, Ewa, University of Economics in Katowice, Poland

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1133–1137
DOI: 10.15439/2014F317
ACSIS, Vol. 2

Towards Semantic-based Process-oriented Control in Digital Home

Tatiana Atanasova
Institute of Information and
Communication Technologies-BAS
Acad. G. Bonchev 2, 1113
Sofia, Bulgaria
Email: atanasova@iit.bas.bg

***Abstract*—The purpose of this work is to investigate the specifics in the development of technologies for heterogeneous data and process integration in digital home and to show possible solutions during design of integrated applications. The analysis of the integrated data can be useful for the development of improved algorithms for monitoring and control of digital networked home.**

I. Introduction

THE tendency for bringing more intelligence into building automation can be seen. It is observed that smart environments have growing demand. The technology development provides a new kind of lifestyle and designing of smart environment attracts attention of researches, home techniques manufacturers, mobile operators, civil engineers and other organizations. The scope of arising problems in digital networked homes is very wide and covers different scientific, technological and psychological aspects [8], [9], [13], [18], [20]. Difficulties results from rapid growing of heterogeneity of electronic devices and communications networks in modern buildings. Home appliances are evolving from purely components devices to complex systems that content processors, sensors and use interfaces. The complexity of the underlying infrastructure is increasing too. The broadband is widely available now in living environment, personal digital devices became very popular, local networks and wireless technologies get emergent interest.

Smart home digital systems have network functions and can be supplemented with connection to the Internet. Network access needs to be available on a range of devices over Wi-Fi and cellular links as well as wired connections. This gives a possibility to monitor and control various home appliances by network and to extend their capabilities through connections in the cloud. Thus home networked system transfers significant functionality from it to the cloud and allows simplifying its design and integration with other systems and services.

The research work reported in the paper is partly supported by the project AComIn "Advanced Computing for Innovation", grant 316087, funded by the FP7 Capacity Programme (Research Potential of Convergence Regions).

Technologies themselves are rapidly changing. The next generation networks are moving to Software-Defined Networking (SDN) where the network's data layer is decoupling from the control layer. SDN is relatively new field in research and consider involving intelligent control methods in network management. This will contribute to a better communication between the various actors involved with various objectives. These new tendencies can be observed in digital networked home environment too – in intention to construct intelligent control and monitoring methods.

However, a lot of problems still have to be resolved, for example: how all these house's devices will communicate, how they will be managed, aggregated, and how the data will be distributed. Besides that, methods for automation in living environment are focused at present on the construction of relatively static structures, designed in advance.

Uniform technology and methods for integrated interoperation of heterogeneous digital systems in living environment that are orientated to optimal using of resources and ensuring of comfort conditions are not developed yet.

The investigation in this paper is focused on problem formulation and directions, in which methods and tools have to be developed for inter-operation of heterogeneous digital systems in smart living environment that have extendible functionality.

First some problems connected with heterogeneous data and information sources are outlined. After then we consider involving semantics into infrastructures. Finally some proposals for solutions are suggested. The extended functionality can be oriented to improving subjective perception of quality of life as well as optimal using of resources in digital homes.

II. Data and Processes Integration

A. Levels of integration

The integration of heterogeneous data and processes can be accomplished at several levels (fig. 1).

978-83-60810-58-3/$25.00

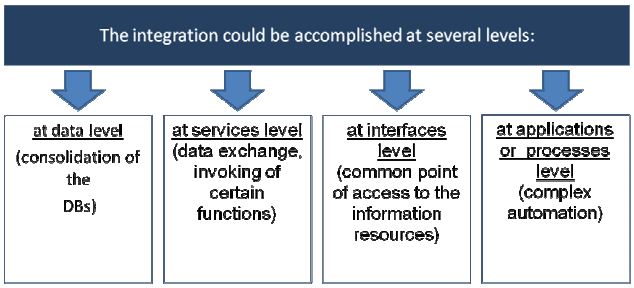


Fig. 1 Integration levels

The existing devices, systems and local networks in a digital house are usually realized with different technologies, regarding the volume of information that is transferred from the integrated devices. We have to take into consideration the heterogeneous data processes and the signals to be conveyed: Ethernet, RF TV and radio signals for wireless-end-connectivity [18] in the distributed building network and necessity of integration of various interfaces for different devices. When new objectives and tasks arise it may be hard to provide interaction between the diversity of the existing appliances.

Each device can be connected with a particular service or a set of services. Thus a multiservice network [12], [15] is established. The multiservice network gives a base for existence of multiple traffic types within the house.

The service may be shared between different devices and dynamically assigned to some of them depending on inhabitants' desires. For very simple example, music or video call may follow the indwellers everywhere in the house. The service can be transferred from one device to another with inference about possibilities for the transfer and service delivery.

Besides certain functionality, the data and process integration requires the construction of an infrastructure, providing safety and security.

B. Ontologies

The semantic description and realization of methods for semantic processing may be the key to achievement of common integration objectives. Several researches suggest an idea to enhance sensors and actuators with a semantic description of their capabilities [8], [9], [13], [20].

Ontologies as a core of semantics can be used for the purposes of information integration, sharing and reuse. The main components of ontology are concepts, relationship, rules and instances. Concept is a class of objects (entities) in the area. Relationships describe the interactions between concepts or properties; they can be in form of taxonomy or associative relationship. Taxonomies systematize concepts as a hierarchical tree, and the associative relationship disposals the concept on the tree.

Instances are specifications of concepts, together with the taxonomy and the relationships they form the knowledge domain. Axioms are used to restrict the values of classes or objects (examples).

Ontology may have logic inference, and then it is so-called formal ontology. Formal ontology must have axioms that restrict the possible interpretations of logical expressions. Web Ontology Language (OWL) can been used to describe each element in the ontology.

Ontologies are created in various forms - from lexicon to dictionary terms, or as first-order logic.

In a broad sense, they can be distributed over three categories: general, domain or applied ontology.

The domain ontology focuses on the refinement of a more narrow meaning of the terms used in a certain area, and may represent a basic reality, in this specific area, but independent of a specific task.

Applied ontology is a specific sub-ontology that contains concepts and relationships which are relevant only to the definite task, such as thesauri, which are semantic relations between lexical units. Usually they contain a small number of concepts with relationships and inference rules, which are defined in detail for solution of particular independent task.

The choice of an appropriate semantic model to represent ontology depends on the purpose for which the ontology is build and the underlying assumptions for achieving these goals.

As an ontology a symbolic system $\{C, T, P, F, A\}$ is considered, where

C is the set of concepts,

T - a thesaurus, or partial order on the set C, the hierarchy of relationships, "subclass" and "super-class";

P - the set of predicates (properties);

F - a function that assigns to each element of P an element from the set of C (considering them in T);

A -is a set of axioms of the ontology.

A hierarchy of concepts is represented as a graph $G = (N, E)$, N - the set of nodes, E - the set of branches, $N=\{n_1, n_2,\ldots n_n\}$, $E=\{e_1,e_2,\ldots e_m\}$.

The graph can be described using XML Schema Datatype (XSD).

At development time ontologies are used to provide ontology–driven development (for example, to describe a domain) or ontology–enabled development (to support developers with their activities).

At run time ontologies form ontology–based architectures (as part of the system architecture) or ontology–enabled architectures (to provide support to the users).

C. Process Ontology

Fundamental process ontologies are becoming more important in recent times [4], [6], [7]. For example, in [19] the idea is discussed that everything is a process and consists of the processes.

The basic postulates are:

- the world is represented as an interconnected system of large and small events;

- some of them are relatively stable;
- the events are always changing;
- the changes represent the actualization of certain features and disappearance of others.

The processes can be divided into:

- constant processes that are interpreted as *concepts*,
- processes which are interpreted as *events*, represent a finite set of four-dimensional space-time.

Thus, the world is built from events, i.e., ontologically, all consists of processes.

The consideration of processes includes:

- when a process should be initiated and finished;
- who participates in this process;
- how this process should be performed;
- which results must be examined, analyzed and taken into account.

The surrounding of a process in such a way consists of data, event, resources, goal and output as a result of a process (fig. 2).

The processes are divided in three main classes:

- *basic* processes;
- *composition* of basic processes and
- *external* processes.

Additionally, processes can be identified that determine the *trends* and directions of changing of basic processes, depending on the analysis and estimated data. The processes are available in the streams of data as implicit patterns. The data is contained in a multitude of sources R (data sensors, files, databases, external resources), $R = \{R_1, ..., R_q\}$.

Extracting knowledge from a specific domain can be considered as the construction of structural design pattern – that process ontology [19]. The objective is to identify processes that have brought to the particular event, and to predict future events based on the past experience.

The process ontology is the key for combining the device and system knowledge.

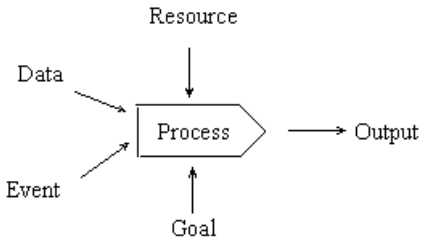


Fig. 2 Process surroundings

III. Proposed Trends and Solutions

The intelligent control methods in digital networked home aim to achieve three goals: semantic integration, providing interface to various devices and ensuring adaptation.

To supply resource control, interoperation and possibility for reconfiguration of digital systems we need to integrate the infrastructure with services.

The semantic and formal description of services and resources is relevant to digital home, where a large diversity of resources have to be described and managed in a highly dynamic way [14]. Services with different purposes can collaborate to offer new and more complex functionalities to the user transparently.

Buildings can be considered as a software problem. This problem addresses integration of information and resources, which are invisible in everyday life. One possible way is to develop operating system for buildings and drivers for every device enriched with semantics. Thus systems can be constructed that are connected to the Internet and are controlled automatically.

It is recognized that there are obstacles in extensive use of embedded devices with limited characteristics of mobility, computing resources and memory. Semantic description may be a way to overcome this large handicap.

Semantic description and modelling of services, together with constructing and using process ontologies provided to users is a key component to autonomic service management, service negotiation and configuration.

The full exploitation of semantics in user and device description has several benefits [21], [22]. The integration of knowledge representation features and reasoning techniques into standard home automation protocols can offer high-level services to users [7].

Current experiences suggest that trends from device-oriented to process-oriented control of home appliances can be seen.

Discovering process models from system event logs is definitely non-trivial. Within the analysis of event logs, process can be defined as the automated construction of structured process models. Each event is a part of the chain process.

The main goal is to suggest a model in which the decision support system provides solution on choosing the optimal set of services using the given network resources on the base of reasoning on process ontologies.

Decision support system (fig. 3) is a coordination unit that integrates heterogeneous data. It is on the top layer of the data processing and provides semantic reasoning. Middle layer realize control, monitoring and visualization functions by event processing, as each device is enhanced with metadata and it is associated with service. This allows discovering functionalities and request services from other devices. Services are discovered by semantic matching. It has to be developed a logic-based ranking of approximated matches allowing to choose resources/services best satisfying a request, also taking user preferences and context into account.

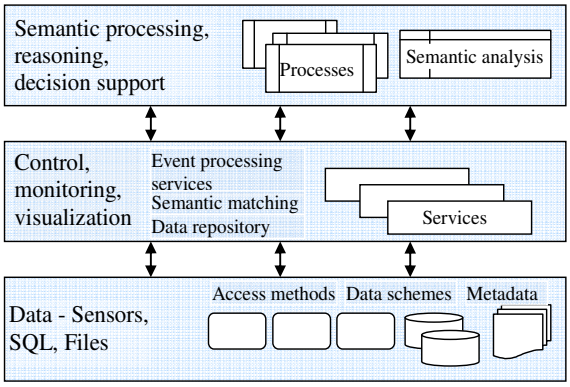


Fig. 3 Layers of data processing

In this framework ontologies are used twofold: for data integration from several sources and for intelligent systems operations.

The system evaluates and classifies records by the use of the process ontology. The ontology of processes will determine what information to extract and how to accelerate the semantic search.

The process-ontology will provide an appropriate philosophical foundation to integration problems. They will give a common conceptual framework for the researches as well as for the practice; it gives a possibility to compare different process-models and concepts and to interpret the dependencies between different models.

IV. Related Work

The development of ontologies in centralized settings is well studied and there are established methodologies. However, current experiences suggest, that ontology engineering should be subject to continuous improvement rather than a onetime effort and that ontologies promise the most benefits in decentralized rather than centralized systems [7].

Using ontologies for telecommunication is proposed by recently [1]-[3], [5]. The approach proposed in [1] discusses using ontologies to capture networking information as well as the domain and expert knowledge needed for network configuration tasks. By semantic description and the use of formal ontologies it is shown how task complexity can be reduced.

Collaborative mechanisms between services are a crucial aspect in the development of pervasive computing systems based on the paradigm of service-oriented architecture [15]. The current network can only provide syntax-layer services and not provide semantic-layer services [5].

The investigation in [20] describes a model of services composition based on a directed acyclic graph used in a service middleware for home-automation, in which loosely coupled services-oriented systems is suggested over the peer-to-peer technology.

The use of semantics (including the integration of ontologies with rule-based systems) has been proposed in early works on context-awareness in the home environment [23]-[26].

The presented in [8] approach proposes use of semantic description that can potentially make the digital networked home more adaptable, agile, sustainable, and dependable given the requirements of changing environment.

The development of multi-service model is discussed in [15] and it is concluded that is still at an early phase.

V. Conclusion

The upward trend in ubiquity and heterogeneity of networked home services and resources demands for a formal and systematic approach to home management tasks.

Current solutions of automated control systems in digital networked homes poorly support dynamic scenarios and context-awareness. Ongoing research covers sensors and how to include the description of these sensors in the control system for smart living environment [9].

In this work it is shown that:

- The primary aim to use of ontologies is to integrate different applications;
- Ontology is proposed to model the relations between events and to manage process configurations.

The rapid development and emerging demands for process automation and interoperability requires systematic modelling methodology and increased semantic information.

The work outlined some solutions for the hard problems during integration of heterogeneous data and processes and shows trends to overcome the lack of standard methods for integration of the information resources and processes that hampers the supply and the efficient use of the information by the users. Process ontologies are a base for providing the whole functionality of the digital networked home in a coordinated and controllable manner.

The future living environment will need to be more intelligent and adaptive, optimizing continuously the use of its resources without any impact on the demanding services.

An effective way to acquire knowledge and share it internally and with outside strategic sources is needed. It is still a challenge to provide a robust and acceptable solution for knowledge capture from different sources.

Methods that ensure integration between the various subsystems automatically and in real time have to be designed. It is hoped that the proposed model will invite further work on integrated framework and shall become reality in the near future.

References

[1] D. Cleary, B. Danev, D. O' Donoghue, “Using ontologies to simplify wireless network configuration”, in *Formal Ontologies Meet Industry*, Verona, Italy, June 9-10, 2005.

[2] X. Qiao, X. Li and J. Chen, “Telecommunications service domain ontology: semantic interoperation foundation of intelligent integrated

services", in *"Telecommunications Networks – Current Status and Future Trends",* J. Ortiz (Ed.), InTech, 2012, pp. 183-210.

[3] D. Cao, X. Li; X. Qiao; L. Meng, "Ontology-based modelling method for semantic telecommunication services", in *Proc. 5th Int. Conf. on Fuzzy Systems and Knowledge Discovery FSKD '08,* 2008, vol. 4, pp. 449 – 453.

[4] T. Liu, H. Wang, L. Liu, "Extended ontology-based process management architecture" in *Cooperative Design, Visualization, and Engineering,* LNCS, vol. 6874, 2011, pp 114-120.

[5] Z. Shangguan, Z. Gao, K. Zhu, "Ontology-based process modeling using eTOM and ITIL" in *Proc. IFIP Int. Conf. on Research and Practical Issues of Enterprise Information Systems,* 2008, vol. 2, pp. 1001-1010.

[6] St. Aitken, J. Curtis, "A process ontology", in *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web,* LNCS, vol. 2473, 2002, pp 108-113.

[7] Ch. Tempich, H. S. Pinto, Y. Sure, and St. Staab, "An argumentation ontology for DIstributed, Loosely-controlled and evolvInG Engineering processes of oNTologies (DILIGENT)", In A. Gómez-Pérez, J. Euzenat (eds.) *"The Semantic Web: Research and Applications", Proc. Second European Semantic Web conference,* ESWC'2005, LNCS 3532, 2005, pp. 241-256.

[8] M. J. Kofler, C. Reinisch, W. Kastner, "A semantic representation of energy-related information in future smart homes", *Energy and Buildings,* April 2012, pp. 169-179.

[9] M. Chan, D. Est`eve, C. Escriba, E. Campo, "A Review of Smart Homes – Present State and Future Challenges", *Computer Methods and Programs in Biomedicine,* Elsevier Ireland, 2008, pp. 55-81.

[10] V. Monov, "Energy consumption and efficiency in industrial processes", *Proc. Int. Conference "Robotics, Automation and Mechatronics'12",* 2012, pp. a9-a12.

[11] T. D. Tashev, H.R. Hristov, "Modelling and synthesis of information interactions", *Problems of Technical Cybernetics and Robotics,* No 52, 2001, pp. 75-80.

[12] Mc. Dysan D., N. Björkman, "Multiservice networking using a component-based switch and router architecture", *in IEEE 2000 Multi-Service Forum, Technical Library,* http://www.msforum.org/techinfo/library.shtml

[13] F. Colace, M. De Santo, "A network management system based on ontology and slow intelligence system", *Int. Journal of Smart Home,* vol. 5, No. 3, July, 2011, pp. 25-38.

[14] C. Rodrigues, S. R. Lima, L. M. Sabucedo, P. Carvalho, "An ontology for managing network services quality" *Expert Systems with Applications,* vol. 39, Issue 9, 2012, pp. 7938-7946.

[15] A. Vilkman, R. Hautala, E. Pilli-Sihvola, "Multi-service model for mobility and logistics", *VTT Symposium on Service Innovation;* Issue 271, 2011, pp. 105-11.

[16] K. S. Munasinghe, F. Javadi, A. Jamalipour, "Resource competition at the NGN core network: An ecologically inspired analysis", *Proc. 18th Int. Conf. on Telecommunications, ICT 2011,* pp. 72-77.

[17] M. Charalambides, G. Pavlou, P. Flegkas, N. Wang, D. Tuncer, "Managing the future internet through intelligent in-network substrates", *IEEE Network,* vol. 25, 2011, pp. 34-40.

[18] J. Guillory, F. Richard, Ph. Guignard, A. Pizzinat, S. Meyer, B. Charbonnier, L. Guillo, C. Algani, H.W. Li, E. Tanguy, "Towards a multiservice & multiformat optical home area network" in *Proc. 14th ITG Conf. on Electronic Media Technology,* CEMT 2011.

[19] J. Palomäki and H. Keto, "A formal presentation of the process-ontological model", in *Information Modelling and Knowledge Bases XXII,* A. Heimburger et al. (Eds.) IOS Press, 2011, pp. 194-205.

[20] J. Holgado-Terriza, S. Rodriguez-Valenzuela, "Services composition model for home-automation peer-to-peer pervasive computing" in *Proc. of the Federated Conference on Computer Science and Information Systems, International Symposium on Services Science,* 2011, pp. 529–536.

[21] T. Atanasova, "Digital networked houses: problems, challenges and trends" in *Proc. 20th Telecommunication forum TELFOR 2012,* Belgrad, Serbia, November 20-22, 2012, pp. 1417-1420.

[22] M. Fathalipour, A. Selamat, and J. J. Jung, "Ontology-based, process-oriented, and society-independent agent system for cloud computing", *International Journal of Distributed Sensor Networks,* 2014, article ID 689650, 17 pages.

[23] H. Hu, D. Yang, L. Fu, H. Xiang, C. Fu, J. Sang, C. Ye, R. Li, "Semantic Web-based policy interaction detection method with rules in smart home for detecting interactions among user policies", *IET Communications,* 2011, vol. 5, No. 17.

[24] Y. Zhiwen, Z. Xingshe, Z. Daqing, C. Chung-Yau, W. Xiaohang, and M. Ji, "Supporting Context-Aware Media Recommendations for Smart Phones," *IEEE Pervasive Computing,* 2006, vol. 5, pp. 68-75.

[25] S. De, K. Moessner, "A framework for mobile, context-aware applications", Marrakech, Morocco, *Proc. IEEE 16th International Conference on Telecommunications (ICT),* 2009, pp. 232-237.

[26] V. Ricquebourg, D. Durand, D. Menga, B. Marhic, L. Delahoche, C. Loge, and A.-M. Jolly-Desodt, "Context inferring in the smart home: an SWRL approach," *Proc. 21st International Conference on Advanced Information Networking and Applications Workshops,* (AINAW '07), 2007.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1139–1146
DOI: 10.15439/2014F213
ACSIS, Vol. 2

Implementation of Virtual Desktop Infrastructure in academic laboratories

Pawel Chrobak
Wroclaw, University of Economy
Komandorska 118-120,
53-345 Wroclaw, Poland
Email: pawel.chrobak@ue.wroc.pl

***Abstrac*—The article describes the causes of the economic and organizational case for implementing VDI solutions in the learning centers of Academic Centers.**

The analysis laboratory infrastructure allows to better understand the broad ability to adapt to VDI and range of benefits they receive, administrators and staff research and teaching. Described implementation is based on VMware Horizon View 5.

The author was the originator of the concept of VDI implementation at the University of Economics in Wroclaw and participant of the project team.

I. Introduction

THE increasing popularization of centralized computing centers popularly referred to as the clouds [5] caused the Universities of Poland to begin building their own private clouds, not only to support their internal processes, but also to provide the students with the virtualized workstation model DaS (Desktop as a Service) [7].

The concept of DaS is to use a VDI environment to offer customers a persistent, highly available desktop that can be accessed from all of their mobile devices. This is significant because right now there are no services available that offer those services. The idea is to take the burden off of the customer by removing the tedious upkeep that the facilities must regularly go through. Desktop as a Service, or DaaS, is used in enterprises for a similar reason. The employees no longer have to worry about maintaining a PC, since the operating system would be centrally managed. The idea is to reduce complexity by centralizing management, which leads to a more productive and efficient IT organization.

This is significant because as the times change consumers are looking for different, innovative solutions to their problems. One problem that consumers have is that they have to regularly maintain their data and hardware. Consumer DaaS would provide a turnkey solution for consumers looking for a simpler computer experience [10].

The Article omits the description of Virtualization as it is generally known. Instead, after a brief description of the VDI we will focus on the analysis of the suitability and potential benefits for virtualization of student laboratory.

The author is an employee of one of the first universities in Poland (University of Economics in Wroclaw) which implemented this solution on a massive scale in their teaching process. In this article, we describe some of the benefits and other experiences gained from the implementation of this project.

The subject of this study are student academic laboratories, although the content contained herein also refers to any type of training centers and educational solutions in schools or centers of learning.

II. Desktop Virtualization (VDI)

For high end graphics workstations, consolidation seeks to migrate the processing load from multiple high end workstations to enterprise class servers in the data center. The primary goal is to offload the work from the local station and provide a desktop environment to users remotely. Rather than stack up a collection of workstations in the data center and provide remote access on a one-to-one basis, desktop virtualization utilizes virtualization techniques to consolidate multiple desktop workstations onto a single server. VMware, an industry leader in virtualization technologies, has created an alliance of vendors and service providers that support what they have coined the Virtual Desktop Infrastructure (VDI). Basically VDI is a server-based computing offering that provides desktop environments as an enterprise hosted service [19]. The core of the VDI initiative can be VMware's ESX server (or other virtualization solution) technology which provides hardware virtualization. Multiple separate operating system images and associated software packages share a single hardware server. Each instance is called a virtual machine (VM). For example, a VMware ESX server might host Microsoft Windows XP, Microsoft Windows Server family, Windows 7/8, and Linux virtual machines at the same time. In its simplest form, each user can connect to a specific virtual machine using some kind of remote desktop protocol. However, having a

978-83-60810-58-3/$25.00

dedicated virtual machine for each user is often impractical, unnecessary and cumbersome. Therefore, VDI solutions normally include some kind of connection broker to connect users to available VM's. Connection brokers are a part of a rapidly developing suite of management tools that can help minimize the support overhead of a VDI solution. Management tools may include services to connect users to the correct pool of VM's, determine which VM's are in use, locate active users, automatically reconnect disconnected remote sessions, provision additional VM's on demand, take VM's offline for testing, updating or troubleshooting, remotely relocate, reboot and reset running and offline VM's, anticipate performance issues or equipment failures, monitor performance, and perform load balancing. It is helpful to compare and contrast VDI with previous generations of remote desktop solutions. Server Based Computing (SBC) is one solution that has developed over the last 10 years, providing applications and desktops to users. Users connect to a remote server, sharing a single instance of an operating system and applications. The application access is increasingly seamless, providing users an illusion that they are working on the application locally even though it is executing on the remote server. Rather than each user getting a VM to themselves, users share connections to the server operating system and installed applications. Citrix Systems has coined the Dynamic Desktop Initiative (DDI) to contrast VMware's VDI. DDI is a Windows based desktop that's delivered over any network and optimized for office … tasks – from simple to complex. DDI is a developing initiative that builds on current solutions. Desktop virtualization is a term that can apply to other types of virtualization strategies. For example, a desktop workstation can be utilizing a desktop virtualization product to allow several operating systems to run simultaneously on one local desktop machine. One common example of this type of desktop virtualization is in the software development life cycles, where it is helpful to have a virtualized production environment available to the developer immediately. This is particularly useful in software development test-bed scenarios. Other desktop virtualization strategies focus on getting a standardized application or operating system image out to local workstations, streaming applications or operating systems out to office computers or unsecured terminals. In these scenarios, the local workstation hardware runs the operating system and/or software that are being provided from a remote source. Note that this does not match our scenario. In both notebook computer desktop virtualization and high end graphic workstation consolidation, the applications and operating system will be running on the remote servers. VDI and DDI approaches each have their strengths and weaknesses. For instance, due to lower overhead, Citrix Presentation Server can support more users per server. However, the applications run in the server operating system environment, rather than the Windows XP professional as they could in a VDI solution. This poses some challenges for us because the applications our users employ are created for use in desktop level operating systems such as Windows XP. They are often not well tested and qualified in the Windows Server environment. How would the developing VDI (VMware) and DDI (Citrix) solutions meet the desktop virtualization challenges presented by Ringling College? Only hands on testing would tell[8].

III. General benefits of implementation of VDI solutions

Computer labs students are usually characterized by low utilization of resources at a fairly fast computer purchased to laboratories. The low utilization statistics indicate that workstation consolidation could achieve great savings in infrastructure, networking, power consumption, and maintenance costs. In addition, we would spend less time in deployment, security, and fault isolation without compromising performance.

With the ever increasing prices of upgrading desktop computers, virtualization of the desktop is becoming very appealing. Here are some of the benefits of virtual desktop infrastructure (VDI).

- **Management** - In a typical corporate infrastructure, you manage desktops using remote software technology such as Altiris or some other push technology. It is really hard to man-age hundreds of desktops as you are well aware if you administer desktops in your corporate infrastructure. Using technology such as virtual desktop infrastructure (VDI) allows you to have central management of all your desktops and really control what is being installed and used on the desktops. Deployment of virtual desktops is lightning fast as opposed to using imaging technology such as Norton or other antiquated technologies. Would you like to manage 500 desktops all over the United States or Europe or manage them from one data center?
- **Security** - Security is a key factor in rolling out VDI. With VDI, you have greater control of how you secure your desktop. You can lock down the image from external devices or pre-vent copying data from the image to your local machine; I'm a big fan of this feature of VDI. Remote users or road warriors also benefit as sensitive data is stored on the server in the data center and not the device. If the device is stolen, the information is protected.
- **OS migrations** - Let's say you want to roll out Windows Vista to a select few managers. Prior to VDI, you would have to look at their equipment and most likely upgrade hardware, memory, disk space, etc. With VDI, you can just push out a Windows Vista image from a central location to the group of managers.
- **VDI image** - We can create a library of VDI images to meet all of your company needs. If your company is seasonal, you can have extra images to handle the increased employee traffic. If you use third-party vendors/contractors/consultants, you can use secure/encrypted locked down images to allow them to work in your environment.

• **Snapshot technology** - With VDI, you have the ability to roll back desktops to different states. This is a great feature, and it allows you to give a lot of flexibility to your end users.

• **Go green** - A thin client VDI session will use less electricity than a desktop computer. Using VDI is a way to reduce your carbon footprint on our planet and save your company money in power costs.

• **Independence** - With VDI, who cares what device you use? A thin client, a PC, Apple, Linux, etc. As long as you can connect to your VDI with ICA or RDP, you are golden[11].

IV. Economic and organizational causes of implementation of VDI in academic environment

Universities have a few reasons for deploying VDI in teaching laboratories: economic, organizational, technological or marketing. In fact almost all aspects have also the implied economic savings (redundancies, etc.), so in the characteristics below will focus on the economic and organizational aspects.

The reality was that desktop support had focused on bringing broken machines back from the dead and not on discovering new tools or ways in which they could be used. Desktop and laptop computer users were supporting themselves in regard to using software and the number of trouble tickets handled clearly indicated that it wasn't about the user. IT wasn't supporting the user of the desktop but instead the computer that was on the desktop. To improve customer service and deliver better support, this paradigm had to change and the only way to change it was to substitute the need for hardware fixes with the ability to apply user fixes. Solving the problems that the user had required extending the knowledge base to them and eliminating the break fix endless loop that existed [9].

V. Characteristics of student's laboratories

Computer workstations in the student laboratories have their own characteristics, the analysis and understanding of them allow us to understand the ability to adapt to VDI in such an environment. Let's try to describe these specifics:

1 For most classes are using the same software configuration of work stations (package Office + software specific to the various classes as Mathematica, Visual Studio package, graphics packages, etc.)

2 For some of the more advanced subjects a specific configuration is needed (usually requiring a more powerful computers) mainly in the case of configurations where the database is locally installed. Mostly these are the classes related to databases, ERP software (eg SAP with local database or computer networks (another network configuration or other elements of the local virtualization). This situation often is solved by separating the specialized laboratories or the ability to run different systems during take-off of systems (separate systems with different configuration on different disk partitions)

3 It is desirable that every student joining the course has a "clean machine " with no files or configuration changes that could leave the previous student

4 The key principle is that all the computers in the lab have exactly the same software configuration

5 One of the main problem of the administrators are configuration changes caused by students. The complexity of Windows causes that despite imposing further restrictions on student accounts, receiving permission to install there are still a lot of gaps, that clever students use to show their abilities (eg changing of desktop background, install add-ons to the browser, vulnerability installation malware). On the other hand, revoking causes problems with the software update, drivers (eg drivers for USB sticks), which handicap the life of students and teachers.

6 Most students perform simple tasks (eg working in Excel) by which the CPU utilization remains at 5-10% for 90 % of the time.

So we summarize some of the implications arising from these observations:

Very time consuming and tedious for administrators is to maintain a number of such labs (in the form of a PC) and the continuous outgoing students of ingenuity who more or less deliberately modify the standard configurations of the operating system or application. Of course, this time consumption expresses the amount of administrators posts, who spend half of their time to perform the same, often unnecessary tasks. Of course there are administrative tools which allow you to automate part of the action (Active Directory, automation software installation and other), but here are new difficulties arising:

• not always administrators working on Colleges are proficient in implementing of new solutions (which implicitly from the earnings on such positions in the Colleges)

• tools to automate software deployment often require administrators to take the time to be able to perform this process. (not all programs can be automatically installed or require a rewrite to version installation "msi"

In addition, there are a number of activities which cannot be automated and require intervention as system recovery, repair damaged units etc.

We also need to take a minute to analyze the second point of described characteristic. Both of described variants are inefficient and cumbersome in practice. Creating specialized laboratories causes the difficulties in allocating and scheduling classes and must lead to not optimal time

management of use of these laboratories or problems with their availability, especially if there are more. Often used on colleges the second option (multi - system configurations) is more efficient but requires a restart of computer before classes which practically can take up to 5 minutes and then require another restart (at the end or at the beginning of successive classes).

VI. Economy benefits of a VDI implementation in student's labs

From the analysis in the previous section is emerging organizational model of the optimal solution: ideally it would be if the student at the beginning of classes receives a computer with a freshly installed operating system and needed him during the classes applications and could do on it what he wants (and even in some cases have administrative privileges) and after completing the course, such a system were completely erased and in its place would be substituted entirely new.

Exactly this possibility gives the replacement of traditional PC by VDI architecture and terminal devices class "zero client".

However, all these considerations are only an introduction to change of the organizational model of maintaining laboratories, further we focus on the economic benefits and we will show savings that will allow institutions to manage more effectively its budget in the IT area :

• **Expenses of maintaining administration** - the implementation of the VDI architecture can significantly simplify the process of administration and maintaining laboratories and0 computer workstations. In the corporate environment in the 90s were taking that one administrative post covered support of about 50 workstations. With the passage of time and the expanding range of tools to automate administrative processes today it is assumed that one administrator (1 full time post) sufficient to support about 500 or more workstations. However in academicals environment (especially in the state Colleges) because of different not essential reasons, that go beyond this elaboration, development of increasing employment is observed in this segment. VDI specification itself causes that whether you have 100 or 500 workstations - for maintenance of laboratories and implementation of demand from the leading ½ of the post is enough. The method of its distribution and scheduling is already in the hands of the management of the institution.

• **Operating costs** - a typical terminal integrated with LED monitor made with "zero client" technology consumes in average 40-50W of electricity which is 4 times less consumption of a typical workstation (which consumes about 200W with the monitor). Of course the VDI infrastructure includes also a set of servers and disk array, so averaging the results for a typical example of ten 30 - bench laboratories we can assume electricity savings of 50%

• **Costs of equipment replacement** - assumed average amortization time of workstation to be three years, and in the university practice this time is estimated at 5 years. VDI equipment manufacturers as one of the advantages of VDI indicate twice as long amortization period of VDI client compare to typical workstation. The key is the fact that the VDI terminal has not in itself any components that determine aging hardware. So there is no processor or graphics card that will be too slow, disc you can drive crash, ending RAM, etc.. All of these components are found in servers, so long as the terminal does not break down naturally in the aging process of electronics or display - no need to replace it. VDI terminals have no mechanical parts including the lack of fans, so MTBF indicators (Mean Time between Failures) is for them around 70 000 hours which is a value more than twice that of a typical workstation (MTBF - 30 000 hours).

It should be noted, that alone cost of VDI terminal is about ¼ cheaper than the average computer set to the lab.

VII. Technical benefits of a VDI implementations

Further description of the functionality is based on the author's experience with VMware Horizon View [16], however, Microsoft's [17] solution offers similar functionality. Of course solutions based on VMware is currently considered the most technologically advanced, however, keep in mind that this solution is most expensive and, if in addition the University has the ability of implementation licensed under the MSDN Academic Alliance [18] (or another program) Microsoft solution can significantly reduce licensing costs and thus the entire implementation.

Philosophy and VDI infrastructure give administrators of the University efficient and stable management environment for teaching laboratories, automating many processes and increasing the reliability of the entire solution. The main aspects that empower administrators :

Central deployment and maintenance of virtual systems - the administrator prepares a single system image, so called " Gold Image " which will be available in read-only option then cloning of each image and creating a virtual system does not copy the entire image. The system reads the data from the golden image and all the changes that are implemented in the virtual system are stored in so called paintings " Linked Clone " This process is visualized in the figure below. We compared here the volume of disk space for five virtual desktops, providing that each prospective of the images saved 1 GB of data for its own needs, while the volume of the golden image is 20 GB (see fig. 1). Of course, space images "Linked Clone" will be automatically increased while writing new data until achieving maximum value (defined by the administrator) [3].

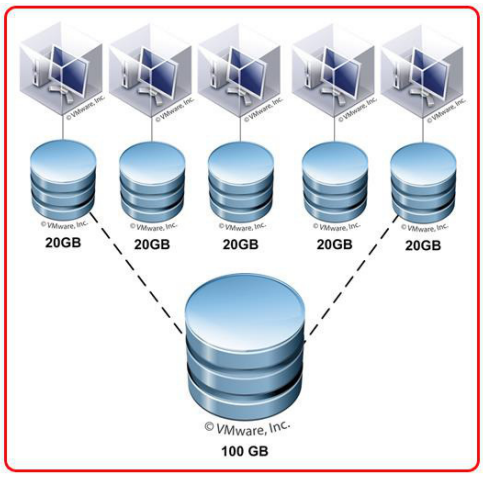


Fig 1. Compare disk space allocation.

For given example a space – saving (average) is 75 % and in the process of increasing the number of working stations factor will still grow. Of course, this technique requires a sufficiently fast disk to store the golden image. It should be stored on disks SSD or disk data should be cached, preferably also using the SSD. Using such technology also gives you another key benefit for our case and laboratories namely fast refresh of virtual systems.

Technology "link cloned 'is particularly useful in these environments. We have here a lot in the exact same virtual machines allowing large savings relate the space of disk. For example, with 100 of the same virtual machine where each would take a 20GB and "link clone" would be at the level of savings we achieve 1GB we can save almost 18 times more space on hard drives. (20GB x100 machines = 2TB and 100 x 1GB + 20GB = 120GB). Estimate the space required (for similar parameters 20GB - used space and 1GB of "link cloned" is shown in Figure 2. The figure shows estimated disk storage capacity requirements (in GB) depending on the amount of the same virtual systems.

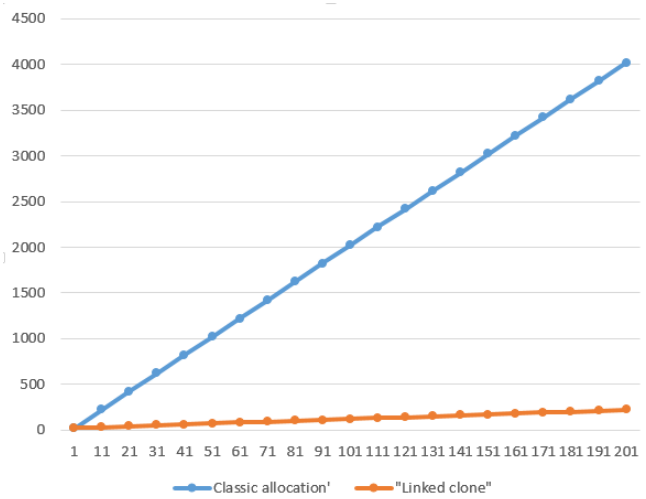


Fig 2. Estimated space required for virtual systems (in GB)

Instant Refreshing virtual systems - one of the implications of the golden images is the fact that if a virtual operating system only reads data from the golden image (without the possibility of writing anything on it) and all the changes differential writes on "Linked Clone", is to delete the data in this place immediately restore the clean image of the system (see fig. 3). The system automatically disconnects the user after 15 minutes of inactivity and instantly refreshes the image of the virtual system, which restores it to its original state (this process takes about five seconds for each system) [6][4].

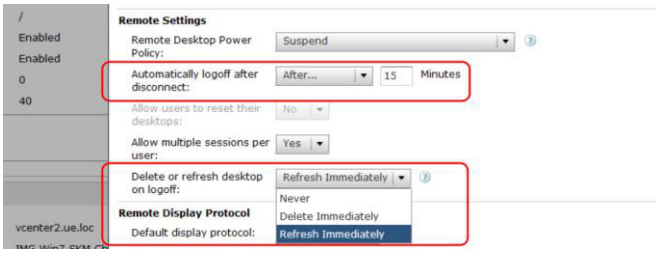


Fig 3. Refreshing virtual OS configuration

Business continuity of laboratory - User cannot spoil anything - unless he physically damages the terminal. Since the virtual system which works will be cleared immediately after work no matter how much he can reconfigure it, disorder or virused by student.

Security policy - depending on the philosophy adopted by the security administrators can be installed one central antivirus program with agents for individual virtual systems, or even resign from anti-virus software on workstations based on the assumption that even if the student hacked the computer, or even computers in the vicinity of this, anyway they will not exist after classes.

In addition, the VDI environment offers many useful additional features that allow you to: load balancing (moving virtual systems to the other servers in the cluster), shutdown of servers with lighter loads snapshots and much more.

Prior to the implementation, it is necessary, of course to

ensure the layout of server resources necessary to ensure adequate performance - the next section describes an example configuration and load up to 300 images, on pages could be found relevant calculators to calculate the parameters of the environment.

VIII. The limitation of VDI technology

Although the systems used for laboratories are not especially graphically demanding (not taking in consideration specific graphics applications) you should be aware that the real-time transmission of multiple video streams can cause performance problems of the entire ecosystem of VDI. These problems are growing not only with an increase of the amount of terminals but also with an increase in resolution of displayed images (which begins to be particularly important in Full HD resolutions and larger).

The specificity of VDI[12] and streaming overshadowing many virtual desktops through a common internet network, the servers even when you start Youtube are very heavily loaded for two reasons:

FIRST REASON - PCoIP stream is encoded in the default workstation and sent to the LAN. While decoding the entire image (especially for the Full HD resolution and larger), it turns out that the process of decoding the stream can take as many as 50-60 % of the two- core virtual processor. To remedy this server can be equipped with a special card decoding hardware PCoIP streams. (Card Teradici Apex 2800 [15]. Application card offloads streaming process several times. One card is able to decode at the same time 40 HD streams or 25 streams of 2560x1600.

SECOND REASON - problem with such high resolutions: default graphics 2/3D is emulated by VMware as for the desktop becomes a bottleneck (especially using newer Windows systems, even with off AERO). The solution to this problem is the use of specialized graphics cards designed for VDI environments - (eg Nvidia Nvidia Grid K1 or K2 Grid)[13] which can be used SVGA (shared VGA) graphics card that is shared for multiple virtual systems. Although the power of the card (and the price as well) seems to be huge in the case of dividing its power for 20-30 virtual machines, can barely support the basic graphical operations.

• Both of the above measures cannot be applied in Blade servers, though the HP Blade Gen8 solutions it is possible to use a miniaturized version of the card. Projecting such solutions for large format displays, should mount them on traditional servers (like dell R720) [14].

IX. Examples of implementations of VDI in University Of Economy in Wroclaw

The concept of rebuilding IT Infrastructure on University of Economics and building its own private cloud began to emerge in mid-2010 and after passaging of all the procedures, fixing concept and finding financing of the project, in September 2011 begin the process of public procurement and project was run in early 2012. The University of Economics was the first university in Poland, which has implemented such solutions on such a large scale.

On the university are working more than 240 terminals (mainly Samsung NC240), more than 400 virtual systems and students can connect to one of 5 available images, depending on the classes and necessary configuration. For the purposes of VDI six 2-processor servers are dedicated, giving a total of 1.2 TB of RAM. Disk array and some servers are equipped with cache memory for SSD -based. One of the servers is equipped with an nVidia card GRID K1 and streaming card supporting hardware PCoIP (Apex 2800).

As storage is used EMC disk array with a total of 20 Tb working gross capacity.

X. Analysis of system performance VDI in practical example

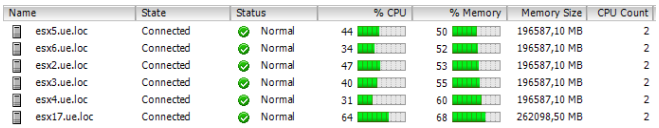
Instead of carrying out theoretical calculations functioning VDI performance will show a practical example of the above-described example, the our University (University of Economy in Wroclaw).

As we mentioned we use a few images of virtual system and the main virtual computers are:

1. **Windows XP** - with 1,5Gb RAM and one core processor, 60Gb provisioned disk and emulated graphics card by vmWare system. Most uses for office and economics applications.
2. **Windows XP with Oracle local database** - with 2GB RAM, two cores, 40Gb provisioned disk, shared graphics card like Nvidia Grid K1 and PcoIP accretion card (Apex 2800). This system is used for classes of databases
3. **Windows 7** – with 3 Gb RAM, two cores, 50Gb provisioned disk and shared graphics shared graphics card like Nvidia Grid K1 and PcoIP accretion card (Apex 2800). This system has software for computer networks and some graphics and business applications.

Below we present some examples of server load these virtual systems in a typical working day. The fig. 4 shows typical load the entire cluster. The cluster consists of:

- Five two-processors Xeon E5645 – 2,4GHz (16 logical cores) servers (Dell M710HD) with 196Gb RAM each –described as ESX2 – ESX6 on the figs.
- One two-processors Xeon E5-2630 – 2,3GHz (24 logical cores) server (Dell R720) with 262Gb RAM each, with NVIdia Grid K1 Video Card and Hardware Acceleration PcoIP Card (Apex 2800) –described as ESX17 on the figs.



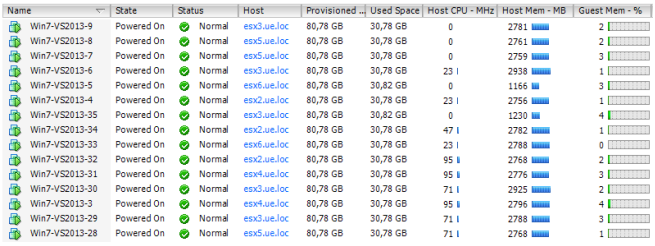
Name	State	Status	% CPU	% Memory	Memory Size	CPU Count
esx5.ue.loc	Connected	Normal	44	50	196587,10 MB	2
esx6.ue.loc	Connected	Normal	34	52	196587,10 MB	2
esx2.ue.loc	Connected	Normal	47	53	196587,10 MB	2
esx3.ue.loc	Connected	Normal	40	55	196587,10 MB	2
esx4.ue.loc	Connected	Normal	31	60	196587,10 MB	2
esx17.ue.loc	Connected	Normal	64	68	262098,50 MB	2

Fig 4. Typically loaded VDI cluster

On the ESX17 there is 84 virtual systems:

- 32 virtual systems of **Windows XP with Oracle local database**
- 52 virtual systems of **Windows 7**

On the fig. 5 and 6 we can see typically load virtual systems when the student class working.



Name	State	Status	Host	Provisioned ...	Used Space	Host CPU - MHz	Host Mem - MB	Guest Mem - %
Win7-VS2013-9	Powered On	Normal	esx3.ue.loc	80,78 GB	30,78 GB	0	2781	2
Win7-VS2013-8	Powered On	Normal	esx5.ue.loc	80,78 GB	30,78 GB	0	2761	2
Win7-VS2013-7	Powered On	Normal	esx5.ue.loc	80,78 GB	30,78 GB	0	2759	3
Win7-VS2013-6	Powered On	Normal	esx3.ue.loc	80,78 GB	30,78 GB	23	2938	1
Win7-VS2013-5	Powered On	Normal	esx6.ue.loc	80,78 GB	30,82 GB	0	1166	3
Win7-VS2013-4	Powered On	Normal	esx2.ue.loc	80,78 GB	30,78 GB	23	2756	1
Win7-VS2013-35	Powered On	Normal	esx3.ue.loc	80,78 GB	30,82 GB	0	1230	4
Win7-VS2013-34	Powered On	Normal	esx2.ue.loc	80,78 GB	30,78 GB	47	2782	1
Win7-VS2013-33	Powered On	Normal	esx6.ue.loc	80,78 GB	30,78 GB	23	2788	0
Win7-VS2013-32	Powered On	Normal	esx2.ue.loc	80,78 GB	30,78 GB	95	2768	2
Win7-VS2013-31	Powered On	Normal	esx4.ue.loc	80,78 GB	30,78 GB	95	2776	3
Win7-VS2013-30	Powered On	Normal	esx3.ue.loc	80,78 GB	30,78 GB	71	2925	2
Win7-VS2013-3	Powered On	Normal	esx4.ue.loc	80,78 GB	30,78 GB	95	2796	4
Win7-VS2013-29	Powered On	Normal	esx3.ue.loc	80,78 GB	30,78 GB	71	2788	3
Win7-VS2013-28	Powered On	Normal	esx5.ue.loc	80,78 GB	30,78 GB	71	2768	1

Fig 5. Typically loaded Windows 7

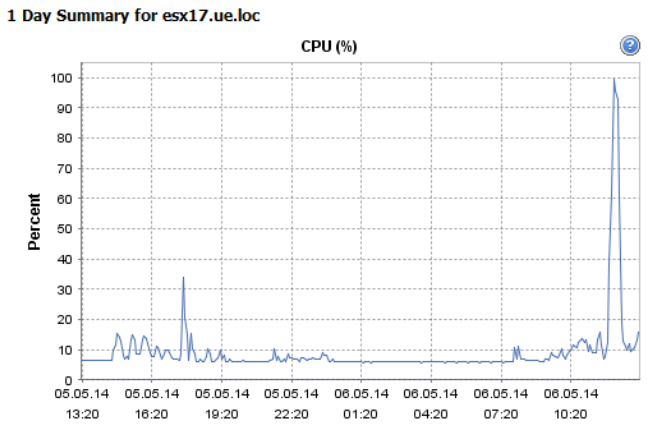
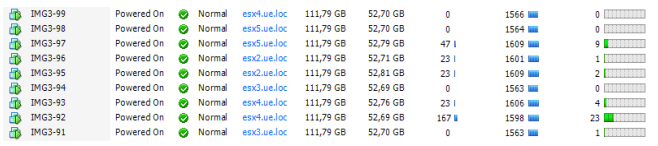


Fig 6. Typically loaded CPUs at ESX7 at worked day

On the ESX2- ESX6 we have 84 virtual systems:

- 32 virtual systems of **Windows XP with Oracle local database**
- 52 virtual systems of **Windows 7**

On the fig. 7 we can see typically load virtual systems with windows XP.



IMG3-99	Powered On	Normal	esx4.ue.loc	111,79 GB	52,70 GB	0	1566	0
IMG3-98	Powered On	Normal	esx5.ue.loc	111,79 GB	52,70 GB	0	1564	0
IMG3-97	Powered On	Normal	esx5.ue.loc	111,79 GB	52,79 GB	47	1609	9
IMG3-96	Powered On	Normal	esx2.ue.loc	111,79 GB	52,71 GB	23	1601	1
IMG3-95	Powered On	Normal	esx2.ue.loc	111,79 GB	52,81 GB	23	1609	2
IMG3-94	Powered On	Normal	esx3.ue.loc	111,79 GB	52,69 GB	0	1563	0
IMG3-93	Powered On	Normal	esx4.ue.loc	111,79 GB	52,76 GB	23	1606	4
IMG3-92	Powered On	Normal	esx4.ue.loc	111,79 GB	52,69 GB	167	1598	23
IMG3-91	Powered On	Normal	esx3.ue.loc	111,79 GB	52,70 GB	0	1563	1

Fig 6. Typically loaded Windows XP with emulated graphics cards

Of course the Windows XP is the best system for virtualization because of low demand for computing power and RAM size as we can see on fig. 7.

XI. Conclusion

After 2 years from implementing VDI on owner University (UE in Wroclaw) we know that VDI promises more efficient use of the university's resources, and could offer students the convenience of accessing specialty software from any device, at any time, from anywhere. Now we do not provide access from anywhere for students (just for teachers and some students involved in student's organization) but I's the next planned step.

ACKNOWLEDGMENT

Special thanks to other employees University of Economy in Wroclaw (administrators and the IT department of managers, employees of the Institute of Business Informatics, IT staff procurement and other departments involved in the project implementation and technology private cloud vdi on our University.

Keywords

VDI, Virtual Desktop Infrastructure, Virtualization, Academic Computing, Private Cloud, Management academic laboratories

References

[1] Serafin M., 2011, Wirtualizacja w Praktyce, Helion, Warszawa.

[2] Finn A., Luescher M., Lownds P., 2012, Windows Server 2012 Hyper-V. *Podręcznik instalacji i konfiguracji*, Helion, Warszawa.

[3] Lowe S.,Marshall N., 2013, Mastering VMware vSphere 5.5, John Wiley & Sons, Indianapilis, Indiana

[4] Guthrie F., Lowe S., Coleman K., 2013, VMware vSphere design, John Wiley & Sons, Indianapilis, Indiana

[5] Rosenberg J., Mateos A.,2012, Chmura obliczeniowa. *Rozwiązania dla biznesu*, Helion, Warszawa

[6] Asselin S., O'DohertyP., 2014, VMware Horizon Suite: Building End User Services, VMware Press

[7] Madden B., Knuth G., 2014, Desktops as a Service: Everything You Need to Know About DaaS & Hosted VDI, Burning Troll Production, San Francisco, California

[8] Miller K., Pegah M.,Virtualization, Virtually at the Desktop

[9] Vieira S., Why Virtual Desktop at CCRI ? Finding Sustainability for Desktop Support

[10] Eaves A., Stockman M., 2012, Desktop as a Service Proof of Concept, 13th annual conference on Information technology education, pp. 85-86, ACM New York, USA

[11] What are the benefits of VDI? *http://www.techrepublic.com/blog/the-enterprise-cloud/what-are-the-benefits-of-vdi/579/*

[12] Wady i zalety wirtualizacji stacji roboczych *http://www.itwadministracji.pl/numery/marzec-2014/wady-i-zalety-wirtualizacji-stacji-roboczych.html*

[13] GRID GPUS *http://www.nvidia.com/object/grid-boards.html*

[14] GPU accelerators and coprocessors for PowerEdge servers *http://www.dell.com/learn/us/en/04/campaigns/poweredge-gpu*

[15] Accelerate the Virtual Workspace: Take a Load off Servers *http://www.teradici.com/products-and-solutions/pcoip-products/hardware-accelerator*

[16] VMware Horizon View *http://www.vmware.com/pl/products/horizon-view*

[17] Microsoft Virtual Desktop Infrastructure (VDI) *http://www.microsoft.com/pl-pl/windows/enterprise/products-and-technologies/virtualization/vdi.aspx*

[18] Microsoft DreamSpark *https://www.dreamspark.com/*

[19] VMware Infrastructure VDI Server sizing and scaling *http://www.vmware.com/pdf/vdi_sizing_vi3.pdf*

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1147–1155
DOI: 10.15439/2014F388
ACSIS, Vol. 2

Multi-criteria Evaluation of the Intelligent Dashboard for SME Managers based on Scorecard Framework

Mirosław Dyczkowski, Jerzy Korczak, Helena Dudycz
Wroclaw University of Economics
ul. Komandorska 118/120 53-345 Wroclaw, Poland
Email: {miroslaw.dyczkowski, jerzy.korczak, helena.dudycz, }@ue.wroc.pl

Abstract — The article presents an approach to evaluate the Decision Support System applied in the InKoM project. The evaluation method is based on a scorecard framework, oriented towards Business Intelligence (BI) systems and projects dedicated to the management supporting of small and medium enterprises (SME). To design the method, known existing commercial and no-commercial BI maturity models, usability standards, and scorecard frameworks have been analyzed and adapted to SMEs area. Notably, the scorecard framework was extended to the new evaluation criteria associated with innovative knowledge-based functions created in the InKoM project, especially such as ontologies of economic and financial knowledge, and visual navigation and exploratory interface based on topic maps. The main elements of the scorecard framework and usage in InKoM of multi-criteria evaluation are illustrated and discussed in this paper.

I. Introduction

The current economic situation forces the decision-makers of small and medium enterprises (SMEs) to have at their disposal current and appropriate knowledge about the economic and financial situation of the enterprise and its environment. Because of that, decision-makers must have the efficient methods and tools to identify and analyze key performance indicators that have an impact on the operations of the enterprise. Analysis and interpretation of information in the traditional way becomes very difficult, sometimes even impossible. Discovering all dependences between various financial ratios is necessary, because they alert managers about anomalies and dangers (see [23]). Decision-makers in these enterprises, in comparison to managers of big companies, may not have access to all essential strategic information. Usually financial expertise is either not available or too expensive. Big companies have at their disposal strategic consultation and possess standard procedures to solve problems in the case of essential changes in the business environment. For financial and personnel reasons most SMEs cannot afford these types of facilities. It should be noted that SMEs operate in a definitely more uncertain and risky environment than big enterprises, because of a complex and dynamic market that has a much more important impact on SMEs' financial situation than on big companies'. Tolerance of mistakes is narrower (see among others [11, pp. 74–91]). In these circumstances, SMEs' decision-makers often act intuitively and as a result, the rationality of their decisions is significantly weaker. Moreover, SMEs' decision-makers often do not have a solid knowledge of economics and finance.

In general, most existing Business Intelligence (BI) and Executive Information Systems (EIS) provide the functionality of data aggregation and visualization (see among others [26], [31]). Many reports and papers in this domain underline the fact that decision makers expect new ICT solutions to interactively provide not only relevant and up-to-date information on the economic and financial situation of their companies, but also explanations taking into account the contextual relationships.

The aim of this article is to present the approach to multi-criteria evaluation of BI innovative functions created and used in the Intelligent Dashboard for Managers (further referred to as InKoM). The InKoM system has been developed by the consortium consisting of the Wroclaw University of Economics (WUE), which is the leader, and a company UNIT4 TETA BI Center Ltd. (TETA BIC). Credit Agricole Bank Polska S.A. also participates in the project.

Figure 1 presents the main components of the InKoM: a comprehensive description of the TETA BI system with examples of its application is available on the website: [27] (see also UNIT4 TETA presentations [1], [29] and other papers published by the authors; see among others [17], [18], [19], [20]). It can be seen that the InKoM uses TETA BI mechanisms for extracting source data from ERP and non-ERP transactional systems internally (ETL), its data warehouse, and analytical database. However, the available solutions – in particular the standard analyses, reports and analytical statements generated by the system – are complemented by economic and financial knowledge – most

This work has been supported by the National Research and Development Centre within the Innotech program (track In-Tech), grant agreement INNOTECH-K1/IN1/34/153437/NCBR/12.

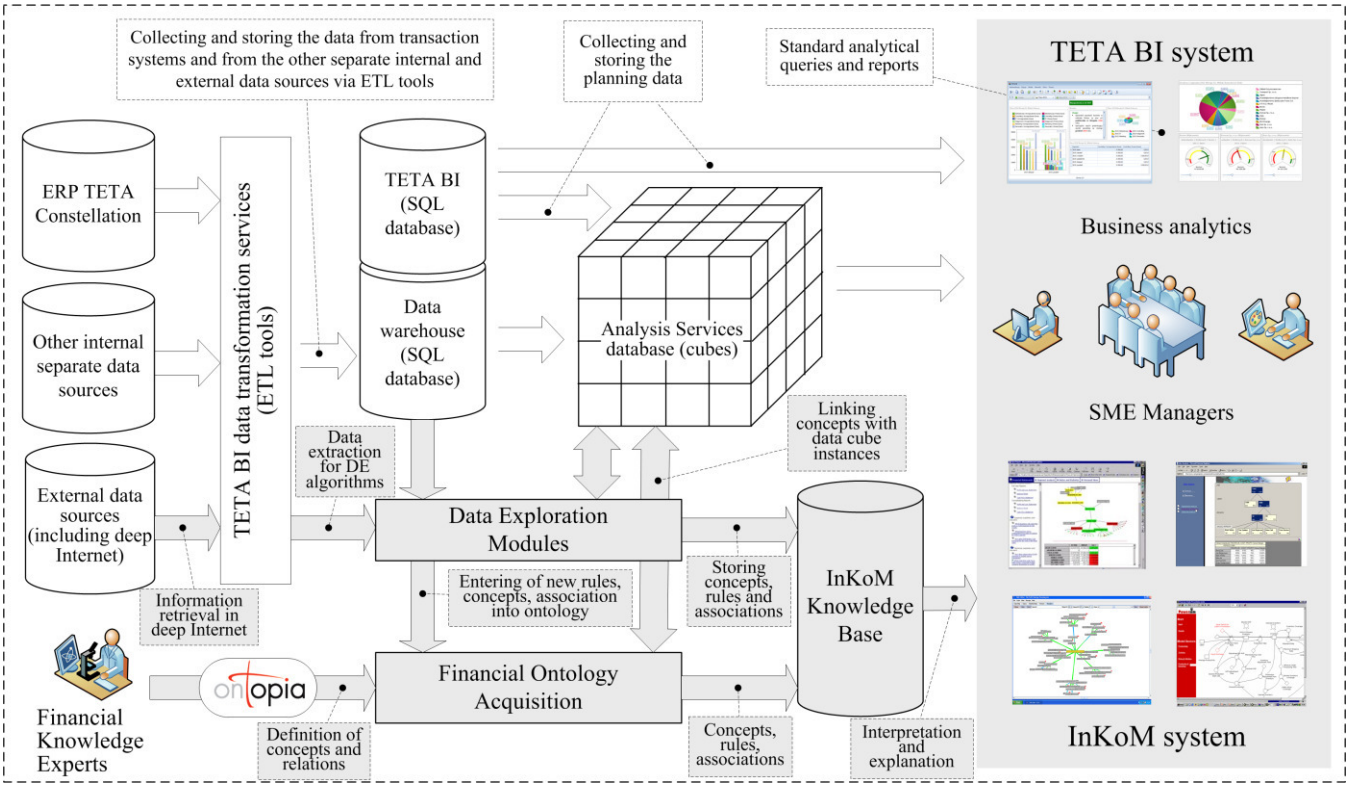


Fig. 1. Components of the Intelligent Dashboard for Managers and their location in the TETA BI system (The InKoM components are shaded in grey)

importantly ontologies and topic maps – and financial data mining algorithms, including mechanisms for extracting business knowledge from the deep Web (indicated as gray color boxes). This enables a dynamic, on-line, interactive analysis of key business indicators.

The transactional data obtained from external sources, supplemented with planning data, e.g., budgets in the form of multidimensional data structures, or cubes, which are stored in a TETA BI Analysis Services database and provide a basis for the on-line, interactive creation of standard analytical queries and/or reports. The InKoM system complements and extends these processes[1]. By providing economic and financial knowledge stored in ontologies and presented in the form of topic maps to facilitate the perception of concepts, InKoM can make the analysis more comprehensive and simpler. This is particularly important for users who are not specialists in the analysis and interpretation of economics and finance.

The structure of the paper is the following: In the next section the overview of approaches to evaluation BI systems and projects evaluation is briefly described. The third section presents the discussion about extension of criteria of evaluation for BI functions oriented on SMEs. To illustrate the use of the extended scorecard framework, a case study of the InKoM Dashboard evaluation is characterized in the fourth section. The last section summarizes the work already carried out and points out the most important conclusions.

II. AN OVERVIEW OF APPROACHES TO BI SYSTEMS AND PROJECTS EVALUATION.

The main goal of any BI system is to access the right data at the right time to allow proactive decision-making (see among others [6], [34]). The users of BI systems expect access to useful information and knowledge through an interface easy to understand and use[2]. However, most of existing BI solutions are designed primarily for users who are able to understand the business data models and semantic and/or algorithmic relationships between financial and economic objects/concepts (data, information, measure, key performance indicator, gauge etc.) used in analytical processes. Today the development of new BI systems is oriented towards BI 2.0 (using semantic search) and 3.0, Service Oriented Architecture (SOA), Software as a Service (SaaS), mobile BI, Big Data technologies using BI etc. (see among others [14], [23], [26], [32]). The typical features of the systems include: proactive alerts and notifications, event driven (real time) access to information, advanced and predictive analytics, mobile and ubiquitous access, improved visualization, and semantic search information.

[1] The InKoM architecture and functionalities have been presented in [17], [18], [19], [20].

[2] Generally the BI system interface should allow users to do: *"find what* they need, understand what they find, and act appropriately, within the *limits of time and effort that they consider adequate for the task"* (http://www.dataprix.com/en/bi-usability-evolution-and-tendencies).

But no matter what BI applications we implement, they should always meet the expectations and needs of their business users. Helpful in achieving these goals can be the systematic, continuous and multi-criteria evaluation of BI systems and projects based on formalized and verified in practice approaches to BI evaluation process [8]. The most important of these are:

1. BI Maturity Models,
2. BI commercial and non-commercial frameworks used to compare BI systems, projects and/or vendors,
3. BI scorecards,
4. Standards of BI systems and/or projects usability and quality,
5. Methods and tools dedicated for evaluation of economic efficiency of BI systems and/or projects.

There are many Business Intelligence Maturity Models (BI MM) developed by different authors such as Business intelligence Development Model (BIDM), TDWI's maturity model, Business Intelligence Maturity Hierarchy, Hewlett Package Business Intelligence Maturity Model, Gartner's Maturity Models, Business Information Maturity Model, AMR Research's Business Intelligence and Performance Management Maturity Model, Infrastructure Optimization Maturity Model, Ladder of Business Intelligence (LOBI) etc. All of these models and case studies of their use are widely described and compared in the available literature and on the websites of their owners, vendors and/or consulting firms applying them (see among others [3], [7], [8], [9], [15], [21], [22], [24], [25]). Because BI Maturity Models primarily assess the maturity of BI solutions used in decision-making processes, BI frameworks are more useful for the evaluation of development projects of the BI application.

BI frameworks are used to compare BI applications, projects and/or vendors. Examples of selected frameworks to the evaluation of BI systems are presented in table 1.

Generally, BI frameworks define the people, processes, platforms and technologies that need to be integrated and aligned to take a more strategic approach to business intelligence, analytics and performance management initiatives [2, p. 1]. There is no single or right instantiation of the BI frameworks. Different configurations can be supported by the framework based on business objectives and constraints.

Often BI frameworks owners as consulting companies create and provide BI evaluation scorecards. A good example is the BI Scorecard® (table 1). The BI evaluation scorecard is a tool to support the evaluation process based on the multi-level pre-defined breakdown structure of the evaluation criteria and scoring technique. The changes of the scoring evaluation of BI system/project from the "as-was" to "as-is" and/or "to-be" status can be monitored and visualized and then can be used for the continuous improvement of BI initiatives. The structure and the few elements of BI Scorecard® used to evaluate the InKoM system are described in Section IV.

The two last, but not the least, "sources of knowledge" need to create a system of evaluation that – as we noted above – use standards of BI usability/quality and methods/tools dedicated for measurement of economic efficiency of BI systems and/or projects.

In the first area the most important are ISO (IEEE, BSI) standards for software/systems usability/quality evaluation such as "the old" ISO/IEC 9126 (usability), "the new"

TABLE 1.
THE COMPARISON OF SELECTED FRAMEWORKS TO THE EVALUATION OF BI SYSTEMS

Owner and framework solution	**Gartner** (Business Intelligence and Analytics Platforms Magic Quadrant and Gartner's Business Analytics Framework) (see [2], [26])	**Dresner Advisory Services** (Small and Mid-Sized Enterprise Business Intelligence Market Study) (see [31], [32], [33])	**BI Scorecard®** (BI Scorecard Evaluation Frameworks) (see [13], [14])
Main (1st level) evaluation categories and selected (2nd level) subcategories	1. Integration: 1.1. BI infrastructure 1.2. Metadata management 1.3. Development tools 1.4. Collaboration 2. Information Delivery 2.1. Reporting 2.2. Dashboards 2.3. Ad hoc query 2.4. Microsoft Office integration 2.5. Search-based BI 2.6. Mobile BI 3. Analysis 3.1. Online analytical processing (OLAP) 3.2. Interactive visualization 3.3. Predictive modeling and data mining 3.4. Scorecards 3.5. Prescriptive modeling, simulation and optimization	1. Ability to write to transactional applications 2. Ad-hoc query 3. Advanced visualization 4. Big data support 5. Collaborative support for group-based analysis 6. Complex event processing 7. Data mining and advanced algorithms 8. Data visualization 9. End user "self-service" 10. In-memory support 11. Interactive analysis 12. Personalized dashboards 13. Pre-packaged 14. Vertical/functional analytical applications 15. Production reporting 16. Social media analysis (Social BI) 17. Text analytics/Data integration/Data quality tools/ETL 18. "Embedded" BI	1. Information delivery and business intelligence reach 2. Business query and reporting 3. Production reporting 4. OLAP support 5. Dashboard capabilities 5.1. Dashboard layout 5.2. Dashboard design 5.3. Presentation 5.4. Alerting 5.5. Analysis 5.6. KPIs/metrics 5.7. Dashboard interactivity 5.8. Delivery 5.9. Architecture 5.10. Other 6. Delivery and Exploration 7. Spreadsheet Integration

SQuaRE (Systems and software Quality Requirements and Evaluation) ISO/IEC 25000:2014, 25010:2011, 25051:2014 and ISO 9241-171:2008 (ergonomics of human-system interaction). The ISO standards defined usability as the software's capacity to be understood, learned, used, and to be attractive to the user in specific use conditions. ISO also establishes four basic principles on which usability is based: ease of learning, ease of use, flexibility, and robustness. These principles were used in the heuristic evaluation of user interface (e.g. dashboards) based on topic maps and visual navigation as a part of the InKoM system usability evaluation [5, pp. 50-58].

The important part of the BI evaluation framework concerns the economic efficiency/effectiveness of BI systems and/or projects (see among others [10], [30]). From this point of view, the evaluation of a given solution represents a process of analysis of costs, benefits and risks, of BI solution, which must be done by a team of both business and IT personnel. The initial evaluation is followed by a series of analyses made before the start of the project (a priori) and after each year of use in order to verify the initial estimation and to adjust the BI solutions.

The main problem that confronts the current frameworks for the measurement of BI solutions is the fact that much of the benefits are strategic benefits, hard to quantify and only appearing several years after the implementation of the solution. Thus, many of the effects of the BI solution are nonfinancial, sometimes intangible effects that lead to financial results after a certain period of time. These benefits come from improved decision-making, and increased quality of information, and often are not financial incomes directly quantifiable (see among others [10], [12], [30]).

There are different methods to evaluate an investment into IT (including BI) solutions. The most important of these is the Cost-Benefits Analysis (CBA) method based on discounted cash flows. CBA used well-known and widely recommended detailed measures and indicators such as IRR (Internal Rate of Return), MIRR (Modified Internal Rate of Return), NPV (Net Present Value) and ROI (Return On Investment). CBA can be extended by the TCO (total cost of ownership) analysis, where TCO/ROI calculators can be used. A good example of such a tool is TDWI Business Intelligence ROI Calculator (www.tdwi.org).

All of these presented "sources of knowledge" are very useful to design multi-criteria evaluation of BI systems and projects. But as a lot of works have noted, most of them are available for large or mid large companies (see among others [9], [23]). However, none of these tools address the project of designing and implementing BI systems in SMEs specifically. Also, there is a lack of guidelines informing how to create BI systems that might be used as reference examples for SMEs.

There is a very important need, because of the role of SMEs as catalysts for the EU (and also Polish) economy, to accelerate SMEs' growth and to improve their competitiveness. This is recognized by the European Commission, which has developed the set of 10 principles to guide the design and implementation of policies both at EU and Member State level, called "Small Business Act" (SBA). The VIII principle of SBA specifies that "The EU and Member States should promote the upgrading of skills in SMEs and all forms of innovation. They should encourage investment in research by SMEs and their participation in R&D support programmes, transnational research, clustering and active intellectual property management by SMEs" [28].

Therefore in the next section we discuss the extension criteria of BI evaluation frameworks for BI functions oriented on SMEs.

III. THE INKOM PROJECT AND THE EXTENSION CRITERIA OF BI EVALUATION FRAMEWORKS FOR FUNCTIONS ORIENTED ON SMES

SMEs may differ from larger companies by a number of key characteristics, e.g. resource and knowledge limitations, lack of money, reliance on a small number of customers, and need for multi-skilled employees. Some of the above-mentioned characteristics are putting a greater strain on the SMEs, causing the successful implementation of BI to be possibly more challenging in this context.

SMEs are socially and economically important and need tools and solutions to preserve their competitiveness in challenging environments, particularly because they operate in highly competitive, turbulent and uncertain markets**.** Usually they do not have control or influence over the market and thus they need to adopt a reactive approach and adapt to market changes.

Scarcity of resources is one of the main problems and a typical characteristic of SMEs. In addition also skills are limited, not only among staff**,** but also owner-managers often do not have enough managerial expertise or organizational capabilities, and this implies poor strategic business planning and human resource management.

Some of the research has mentioned that for a successful BI project implementation and to bring tangible business benefits to SMEs in the future, it is necessary to meet the following critical success factors: well defined business problem and processes, well defined users' expectations, adjusting the BI solution to users' business expectations, integration between the BI system and other systems, data quality and the flexibility and responsiveness of BI on users' requirements, appropriate technology and tools, and "user friendly"/usability of BI system (see among others [9], [23]).

The analyses presented in the report "Small and Mid-Sized Enterprise Business Intelligence Market Study" specified that "making better decisions" was the most-sought outcome of BI, but SMEs show an even higher regard for revenue growth and competitive advantage stemming from Business Intelligence than their larger peers

[33, p. 15]. The technology priority changes among SMEs 2012-2013. Only three technologies related to BI increased in importance over 2012: Software-as-a-Service (Cloud BI), Dashboards, and Mobile Device Support [33, p. 25]. For 2013, top technologies related to BI in SMEs included: Dashboards, End User "Self-Service", Advanced Visualization and Data Warehousing. [33, p. 35-36]. The same survey noted that at SMEs, Executive Management, Sales, Finance, and Strategic Planning are most likely to drive BI initiatives and projects. Small Enterprises of one to 100 employees are the most likely of all to see Business Intelligence driven by Executive Management (which might describe CEO, CFO, COO or other titles) and are more likely to be driven from the Sales function [33, p. 26].

Features of SMEs and analysis of the BI market for SMEs indicate the directions of the development of modern BI systems. These directions are included in the InKoM project. In the development of InKoM, many new features are integrated, such as domain ontology covering key concepts of corporate finance and economics, knowledge discovery algorithms, semantic search mechanisms, explanation facilities, and tools for visual navigation in domain knowledge.

One of the main parts of modern BI systems is the ontology. In general, the ontology is used to define the necessary knowledge (see [19], [20]).

In the InKoM project, six ontologies were built, covering economic and financial areas: Cash Flow at Risk, Comprehensive Risk Measurement, Early Warning Models, Credit Scoring, the Financial Market, and General Financial Knowledge. Integration of these ontologies into the BI systems assures:

– support for the definition of business rules in order to get proactive information and advice in decision-making;
– a semantic layer describing relationships between the concepts and indicators;
– relevant information according to the different kinds of users that can be found in an organization;
– effective usage of existing data sources and data warehouse structure [20].

All of these benefits require the extension of the evaluation criteria of BI systems for domain-ontologies category.

The knowledge representation layer is the most critical aspect of a BI system, since it broadly shapes the core understanding of the information displayed on their screen [34]. In InKoM design, the basic assumption of navigation was that managers should be able to view focus and context areas at the same time to present the relevant knowledge structure.

Visual exploration in InKoM (see figure 2) is based on a standard Topic Map (TM – ISO/IEC 13250:2003). TM enables the representation of complex structures of knowledge bases and the delivery of a useful model of knowledge representation, where multiple contextual indexing can be used. Developed topic maps for analysis of economic indicators (see among others [5], [6], [16], [19], [20]) have demonstrated that the system [4]:

– can be easily used for the representation of economic knowledge about economic and financial measures,
– can express the organizational structure,
– can be adapted to new applications and managers' needs,
– can be supportive of the managerial staff by facilitating access to a wide range of relevant data resources,
– can assure a semantic information search and interpretation for non-technically-minded users,
– can visualize different connections between indicators that make possible the discovery of new relations between economic ratios constituting knowledge still unknown in this area,
– can improve the process of data analysis and reporting by facilitating the obtaining of data from different databases in an enterprise, and finally
– can be easily extended by users who are not IT specialists, e.g. by experts in economic analysis (using tools for creating a topic map application).

In turn, this group of features and benefits requires the extension of the evaluation criteria of BI on visual navigation and a data exploration interface based on standard topic maps categories.

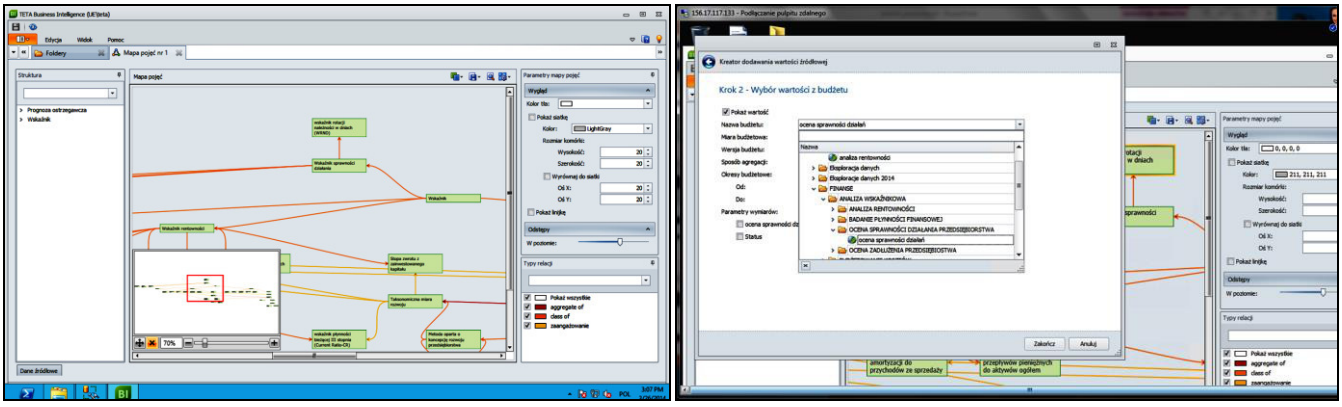


Fig. 2. The visual navigation and data exploration interface of the InKoM system based on topic maps with additional tools and wizards

This is very important in the case of SMEs, where a company does not employ experts in economic-financial analysis and using external consulting is too costly. Reproducing knowledge with the use of a topic map contributes inter alia to a better understanding of economic concepts and the interpretation of specific economic and financial indicators.

Data exploration algorithms (such as classification trees, association rules methods, clustering) have been integrated with topic maps (i.e. semantic search and visual data exploration). In general, data mining tools currently available on the market contain many knowledge extraction algorithms, but a lot of them are not applicable for SMEs. Moreover some of them are too complex and their usage requires costly expert support.

The data exploration module in InKoM not only is integrated with topic maps/ontologies and contains data exploration methods and algorithms dedicated for SMEs, but also is simple to use for non-analysts. Managers in the data exploration process use the built-in wizards to build step by step data mining models (see figure 3).

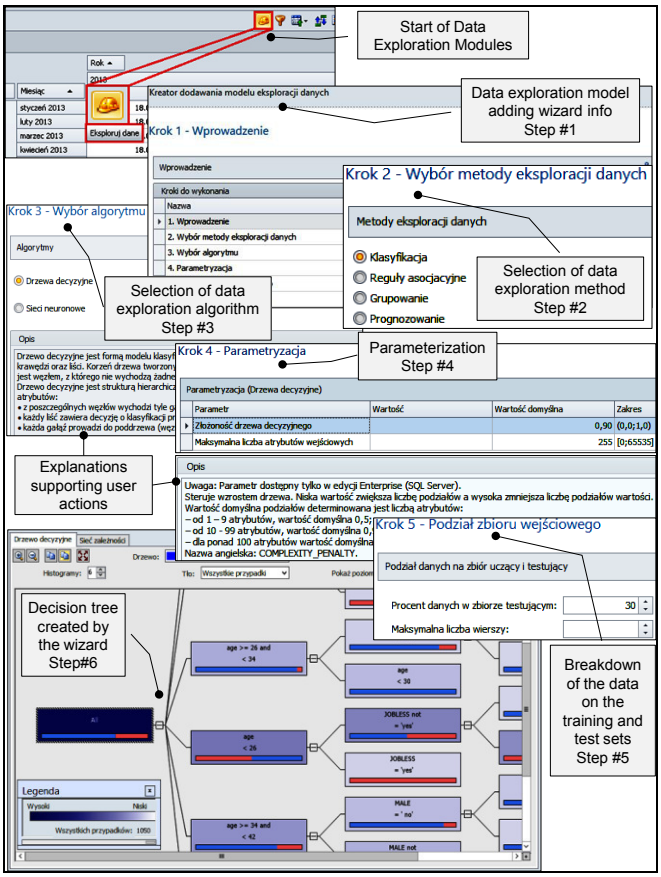


Fig. 3. Data exploration module wizard

These features require the extension of the evaluation criteria of BI related to topic maps/ontologies, dedicated to SMEs' exploration methods and built-in wizards.

IV. Case Study – Evaluation of Innovative Functions of the InKoM Dashboard

Evaluation of the InKoM Dashboard based on categories and subcategories used in BI Scorecard® with extensions was defined in the section III. All ratings were exposed using an approach based on the Delphi method. "As-was" assessment was issued on the basis of self-assessment by TETA BIC specialists. In turn, "as-is" assessment was prepared on the basis of internal expertise (developed by InKoM project teams from TETA BIC and WUE) and external expertise (developed by experts from universities and/or research centers and SME's managers). The results of the evaluation are presented in figures: 4 (business query and reporting category), 5 (delivery and exploration category), 6 (information delivery & BI reach category) and 7 (dashboard category). The detailed requirements for dashboard evaluation are reported in the tables 2÷10, namely:

- the dashboard layout category evaluation (table 2),
- the dashboard design category evaluation (table 3),
- the presentation category evaluation (table 4),
- the alerting category evaluation (table 5),
- the analysis category evaluation (table 6),
- the KPIs / metrics category evaluation (table 7),
- the dashboard interactivity category evaluation (table 8),
- the architecture category evaluation (table 9),
- the delivery and other category evaluation (table 10).

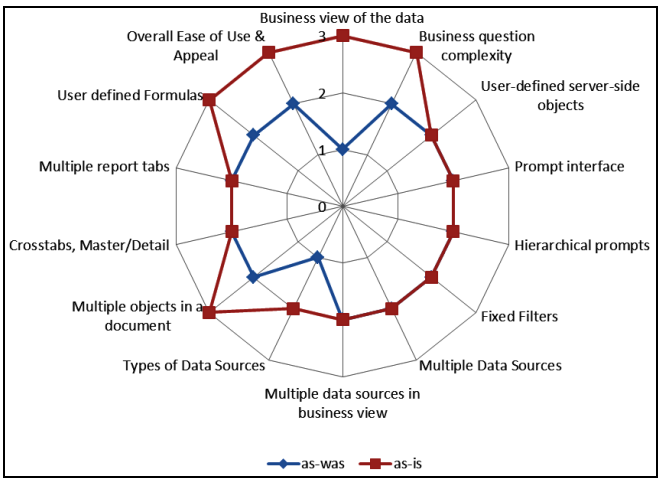


Fig. 4. The results of business query and reporting category evaluation of the InKoM project. Scoring changes from the TETA BI system (as-was) to TETA BI with InKoM functionalities (as-is)

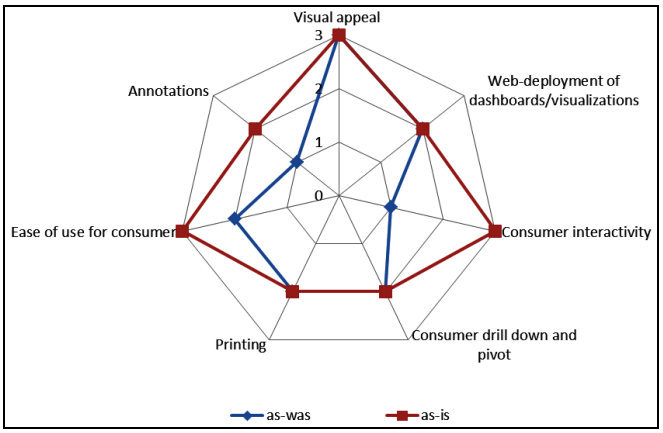


Fig. 5. The results of delivery and exploration category evaluation of the InKoM project. Scoring changes from the TETA BI system (as-was) to TETA BI with InKoM functionalities (as-is)

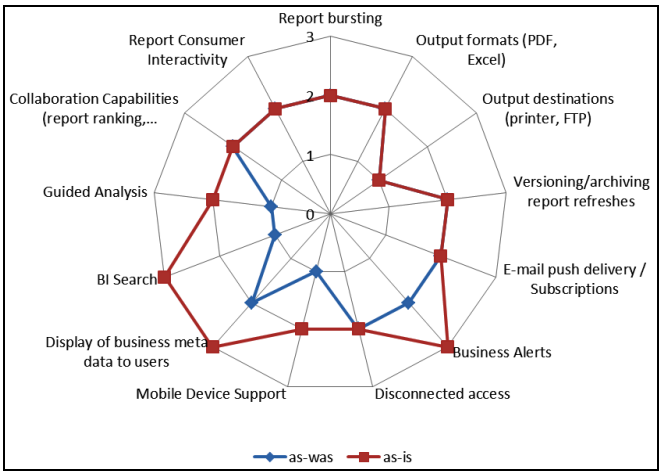


Fig. 6. The results of information delivery & BI reach category evaluation of the InKoM project. Scoring changes from the TETA BI system (as-was) to TETA BI with InKoM functionalities (as-is)

Table 2.
THE DASHBOARD LAYOUT CATEGORY EVALUATION
DETAILED REQUIREMENTS

Dashboard Layout - detailed requirements	as-was	as-is
Multiple objects on a page/display	2	3
Ability to resize portal objects independently	2	2
User defined dashboard layout, in addition to centrally built by IT	2	3
Multiple data sources within dashboard presentation	2	3
Average value	**2,00**	**2,75**

Table 3.
THE DASHBOARD DESIGN CATEGORY EVALUATION
DETAILED REQUIREMENTS

Dashboard Design - detailed requirements	as-was	as-is
Formatting templates for consistent look	0	1
WYSIWYG design mode	2	2
Structure mode for faster design without all data	1	2
Record limit in design mode	2	2
Undo	0	1
Java development environment or Visual Studio or SDK for embedding	2	2
Developer-defined calcs for data not in data warehouse	0	1
Elements re-usable in multiple dashboards	2	3
Web-based design environment	0	1
Ease of design and maintenance aspects	2	3
Average value	**1,10**	**1,80**

Table 4.
THE PRESENTATION CATEGORY EVALUATION
DETAILED REQUIREMENTS

Presentation - detailed requirements	as-was	as-is
Conditional formatting - traffic lights, trend arrows, highlighting of exceptions and variances within tabular display	2	3
Charts Overall	2	2
Hi/Lo Chart	2	2
Gauge Chart	2	2
Bullet Graphs	0	0
Spark Lines	0	1
Maps	1	2
Ability to create own visualizations	1	2
Average value	**1,25**	**1,75**

Table 5.
THE ALERTING CATEGORY EVALUATION
DETAILED REQUIREMENTS

Alerting - detailed requirements	as-was	as-is
Alerts - Visual display of exception values or text	2	3
Alerts - Email notification	2	2
Alerts - user defined in addition to centrally defined	2	3
Alert as RSS feed or textual display within dashboard	0	0
Average value	**1,50**	**2,00**

Table 6.
THE ANALYSIS CATEGORY EVALUATION
DETAILED REQUIREMENTS

Analysis - detailed requirements	as-was	as-is
This Year/Last Year analysis	2	3
Top 10 ranking	2	2
Asymmetrical reporting (expand Q4, collapse Q1-Q3)	1	2
Predictive analysis / what if	0	2
Advanced analysis (based on data exploration)	0	2
Average value	**1,00**	**2,20**

Table 7.
THE KPIS / METRICS CATEGORY EVALUATION
DETAILED REQUIREMENTS

KPIs / metrics - detailed requirements	as-was	as-is
Web-based screen for users to enter target for KPI	0	1
Multiple targets per metrics (stretch goals)	2	3
User-defined KPIs	2	3
IT-developed KPIs as part of dashboard	1	1
Predefined KPIs / metrics dedicated for managers	1	3
Average value	**1,20**	**2,20**

Table 8.
THE DASHBOARD INTERACTIVITY CATEGORY EVALUATION
DETAILED REQUIREMENTS

Dashboard Interactivity - detailed requirements	as-was	as-is
Global filter for all gadgets in dashboard	0	0
Re-sort data in a table within an existing dashboard	2	2
Drill-down	2	3
Pivot / drill by other dimensions	2	3
Drill from one dashboard to another with context passed	0	1
Sliders / Lassos to select content	0	0
Flash animation	0	0
Overall usability and navigation	2	3
Interactivity based on new visual tools (topics maps)	**0**	**3**
Average value	**0,89**	**1,67**

Table 9.
THE ARCHITECTURE CATEGORY EVALUATION
DETAILED REQUIREMENTS

Architecture - detailed requirements	as-was	as-is
Caching - consistently fast response time	2	3
Auto refresh/requery of dashboard objects	2	2
In-memory	0	0
Web-based dashboard delivery	1	2
Broad and Flexible data access (OLAP, relational, Web-Services, deep Internet)	2	3
Dashboard integration with other tools in the BI Suite	2	3
Average value	**1,50**	**2,17**

Table 10.
THE DELIVERY AND OTHER CATEGORY EVALUATION
DETAILED REQUIREMENTS

Delivery and other - detailed requirements	as-was	as-is
Print whole dashboard	2	2
Export to PDF	2	2
Export to Excel	2	2
Disconnected access	0	1
Live Excel connectivity	0	0
Guided analysis / workflow / link reports	1	2
Annotations /Collaboration	1	2
Integration with ontologies and topic maps	**0**	**3**
Average value	**1,00**	**1,75**

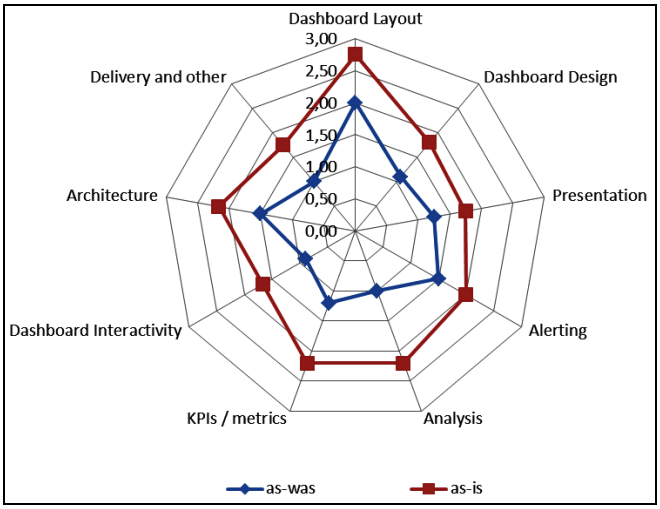


Fig. 7. The results of Dashboard category evaluation of the InKoM project. Average value of detailed requirements (see tables 2÷10) of scoring changes from the TETA BI system (as-was) to TETA BI with InKoM functionalities (as-is)

The evaluation of the InKoM system, especially the dashboard categories and subcategories (see the "greyed" cells), shows necessities for improvement of the BI evaluation frameworks and their customization to SMEs solutions and new innovative technologies and concepts.

V. SUMMARY AND FUTURE WORKS

In this paper, the multi-criteria evaluation of the Intelligent Dashboard for SME Managers used in the InKoM project environment was presented. Further studies will be conducted on empirical verification of the created framework in "real" SMEs, extension of the evaluation categories to support CBA analysis and measurement of ROI/TCO, and creation of a community of experts to continuously extend and update the evaluation tools.

REFERENCES

[1] Architektura systemu. *TETA Business Intelligence. Materiały* informacyjne, UNIT4 TETA Business Intelligence Center, Wrocław 2011.

[2] Chandler N., Hostmann B., Rayner N., Herschel G., Gartner's Business Analytics Framework, Gartner, Inc. no. G00219420, 20 September 2011, http://www.gartner.com/imagesrv/summits/docs/na/business-intelligence/gartners_business_analytics__219420.pdf.

[3] Describing the BI journey. The HP Business Intelligence (BI) Maturity Model. White paper no. 4AA3-9723EEW, created June 2012, http://h20195.www2.hp.com/v2/GetPDF.aspx%2F4AA3-9723EEW.pdf

[4] Dudycz H., "The concept of using standard topic map in Business Intelligence system", in: Proceedings of the 5th International Conference for Entrepreneurs, Innovation and Regional Development – ICEIRD 2012, D. Birov, Y. Todorova, eds., St. Kliment Ohridski University Press, Sofia, Bulgaria 2012, pp. 228-235.

[5] Dudycz H., "Research on usability of visualization in searching economic information in topic maps-based application for return on investment indicator", in: Advanced Information Technologies for

Management - *AITM'2011. Intelligent Technologies and Applications*, J. Korczak, H. Dudycz, M. Dyczkowski, Eds., Wroclaw University of Economics Research Papers no 206, Wrocław 2011, pp. 45-58.

[6] Dudycz H., "Visualization methods in Business Intelligence systems – an overview", in: Business Informatics (16). Data Mining and Business Intelligence, J. Korczak Ed., Research Papers of Wroclaw University of Economics, 2010, no. 104, pp. 9-24.

[7] Eckerson W., Business Intelligence Maturity Model, The Data Warehousing Institute TDWI, 1 March 2006, http://www.eurim.org.uk/activities/ig/voi/03-01-06_Executive_Series_Assessing_Your_BI_Maturity.pdf

[8] Farrokhi V., Pokorádi L., "The necessities for building a model to evaluate Business Intelligence projects – Literature Review", International Journal of Computer Science & Engineering Survey (IJCSES), vol. 3, no. 2, April 2012, pp. 1-10.

[9] Fedouaki F., Okar Ch.,, El Alami S., "A maturity model for Business Intelligence System project in Small and Medium-sized Enterprises: an empirical investigation", IJCSI International Journal of Computer Science Issues, vol. 10, issue 6, no 1, November 2013, pp. 61-69.

[10] Ghilic-Micu B., Stoica M., Mircea M., "A framework for measuring the impact of BI solution", in: Proceedings of the 9th WSEAS International Conference on Mathematics & Computers in Business and Economics (MCBE '08), 2008, http://www.wseas.us/e-library/conferences/2008/bucharest/mcbe/10mcbe.pdf

[11] Gibcus P., Vermeulen P.A.M., Jong J.P.J., "Strategic decision making in small firms: a taxonomy of small business owners", International Journal of Entrepreneurship and Small Business, vol. 7, no. 1, 2009, pp. 74-91.

[12] Gibson M., Arnott D., Jagielska I., Evaluating the Intangible Benefits of Business Intelligence: Review & Research Agenda, Decision Support Systems Laboratory, 2004, pp. 295-305.

[13] Howson C., Successful Business Intelligence: Secrets to Making BI a Killer Application, McGraw-Hill, New York, 2008.

[14] Howson C., Successful Business Intelligence, Second Edition: Unlock the Value of BI & Big Data, McGraw-Hill Education, New York, 2013.

[15] Hribar Rajterič I., "Overview of Business Intelligence Maturity Models", Management, vol. 15, no 1, 2010, pp. 47-67.

[16] Korczak J., Dudycz H., "Approach to visualization of financial information using topic maps", in: Information Management, B. F. Kubiak, A. Korowicki, Eds., Gdansk University Press, Gdansk 2009, pp. 86-97.

[17] Korczak J., Dudycz H, Dyczkowski M., "Intelligent Dashboard for SME Managers. Architecture and Functions", in: Proceedings of the Federated Conference on Computer Science and Information Systems FedCSIS 2012. M. Ganzha, L. Maciaszek, M. Paprzycki, Eds., Polskie Towarzystwo Informatyczne, IEEE Computer Society Press, Warsaw, Los Alamitos, CA 2012, pp. 1003–1007.

[18] Korczak J., Dudycz H., Dyczkowski M., "Intelligent decision support for SME managers – project InKoM", Business Informatics (Informatyka Ekonomiczna), no 3 (25), 2012, pp. 84-96.

[19] Korczak J., Dudycz H., Dyczkowski M., "Design of Financial Knowledge in Dashboard for SME Managers", in: Proceedings of the 2013 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, Eds., Polskie Towarzystwo Informatyczne, IEEE Computer Society Press, Warsaw, Los Alamitos, CA 2013, pp. 1111–1118.

[20] Korczak J., Dudycz H., Dyczkowski M., "Specification of financial knowledge – the case of Intelligent Dashboard for Managers", Business Informatics (Informatyka Ekonomiczna), no 2 (28), 2013, pp. 56-76.

[21] Lahrmann G., Marx F., Winter R., Wortmann F., "Business Intelligence Maturity: Development and Evaluation of a Theoretical Model", in: Proceedings of the 44 Hawaii International Conference on System Science, 2011.

[22] Olszak C., "Assessment of Business Intelligence Maturity in the Selected Organizations", in: Proceedings of the 2013 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, Eds., Polskie Towarzystwo Informatyczne, IEEE Computer Society Press, Warsaw, Los Alamitos, CA 2013, pp. 951–958.

[23] Olszak C., Ziemba E., "Critical Success Factors for Implementing Business Intelligence Systems in Small and Medium Enterprises on the Example of Upper Silesia, Poland", Interdisciplinary Journal of Information, Knowledge & Management, vol. 7, 2012, pp. 129-150.

[24] Popovič A., Turk T., Jaklič J., "Conceptual model of business value of business intelligence systems", Management, vol. 15, no 1, 2010, pp. 5-29.

[25] Raber D., Wortmann F., Winter R., "Towards the Measurement of Business Intelligence Maturity", in: Proceedings of the 21st European Conference on Information Systems, 2013, http://www.staff.science.uu.nl/~Vlaan107/ecis/

[26] Schlegel K., Sallam R.L., Yuen D., Tapadinhas J., Magic Quadrant for Business Intelligence and Analytics Platforms, Gartner, Inc. no. G00239854, 5 February 2013, http://www.walmeric.com/body/pm/2013_gartner_magic_quadrant_for_bi_and_analytics.pdf.

[27] TETA Business Intelligence, UNIT4 TETA Business Intelligence Center, http://tetabic.eu/pl/aplikacja.html.

[28] *"Think Small First". A "Small Business Act" for Europe*. Communication from the Commission of the European Communities COM (2008) 394 final, 25 June 2008, Brussels, http://www.socialeconomy.eu.org/spip.php?article531/COM-2008-0394-FIN-EN-TXT

[29] We change data into knowledge. TETA Business Intelligence. *Materiały informacyjne* UNIT4 TETA Business Intelligence Center, Wrocław 2011.

[30] Whittemore B., The Business Intelligence ROI Challenge: Putting It All Together, Business Intelligence Best Practices, 2008, http://www.bi-bestpractices.com/view/4782

[31] Wisdom of Crowds® Business Intelligence Market Study. 2013 Edition, Licensed to Information Builders, Dresner Advisory Services, LLC, 20 May 2013, http://www.informationbuilders.com/tracker/email/new/pdf/2013_wisdom_of_crowds_bi_market_study.pdf

[32] Wisdom of Crowds® Mobile Computing / Mobile Business Intelligence Market Study. 2013 Edition, Licensed to MicroStrategy, Dresner Advisory Services, LLC, 11 December 2013,https://www.microstrategy.com/Strategy/media/downloads/white-papers/mobile_dresner-mobile-bi-study-2013.pdf

[33] Wisdom of Crowds® Small and Mid-Sized Enterprise Business Intelligence Market Study. 2013 Edition, Licensed to TIBCO Software, Dresner Advisory Services, LLC, 6 November 2013, http://explore.tibco.com/rs/tibcospotfire/images/Wisdom_of_Crowds_SME_BI_Report-Licensed_to_TIBCO_Software-Copyright_2013.pdf

[34] Wise L., The emerging importance of data visualization, part 1, October 29, 2008, http://www.dashboardinsight.com/articles/business-performance-management/the-emerging-importance-of-data-visualization-part-1.aspx.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1157–1162
DOI: 10.15439/2014F102
ACSIS, Vol. 2

Identification of the knowledge conflicts' sources in the architecture of cognitive agents supporting decision-making process

Jadwiga Sobieska-Karpińska
Wrocław University of Economics
ul. Komandorska 118/120,
53-345 Wrocław, Poland
Email: jadwiga.sobieska-karpinska@ue.wroc.pl

Marcin Hernes
Wrocław University of Economics
ul. Komandorska 118/120,
53-345 Wrocław, Poland
Email: marcin.hernes@ue.wroc.pl

***Abstract*—This article presents the problem of knowledge conflicts identification in the architecture of cognitive agents. The agents operate at the decision support systems. The types and the sample of cognitive agents architecture was characterized in the first part of article. Next, the causes of knowledge conflicts was indicated. The final part of article contains the analysis of sources of knowledge conflicts and their examples related to decision-making process.**

I. Introduction

A DECISION-MAKING process may be supported by the use of tools of various kind, in particular IT systems. Currently used IT tools (systems) support decision-making mainly at an operational and tactical level but they become insufficient at a strategic level. They only enable the analysis of the form of information, the links between economic values and they are unable to analyse their meaning. Thus these tools mainly serve the conversion of the gathered data (mostly disordered and unstructured) into information - useful, legible and easily interpretable and thus more suitable to a decision-maker. However, for the definition of meaning of information, a human mind is necessary, and the change of knowledge into wisdom (necessary to take a good decision) - requires not only human intellect but even human genius [19]. Therefore, it seems justified to use the tools which perform cognitive and decision-making functions, the ones that take place in the human brain and owing to this are capable of understanding the real meaning of the observed phenomena and economic processes taking place in the organization environment. These tools include cognitive agents which often cooperate within the framework of a multi-agent system [e.g. 21] in order to effectively reach a set goal.

The architectures of cognitive agents are complex and their functioning is of asynchronous nature, which may be the reason for the occurrence of knowledge conflicts and have a negative impact on the results of cognitive and decision-making functions, which in turn may hinder supporting a decision-making process.

Previous research related to the issues of knowledge conflicts, and in particular with defining their sources [e.g. 10, 20] relate mostly to multi-agent systems composed of reactive agents, so the ones which are capable of drawing conclusions and adequately react to stimuli from the environment however do not have the cognitive function and have limited learning skills. With respect to the agents of this kind, knowledge conflicts occur in situations of opposition or in-coherence of the knowledge held by the agent [7, 14, 15, 18]. However, works concerning the sources of knowledge conflicts with respect to cognitive agent [e.g. 12, 16] are limited to very general approaches, and they do not take into consideration modules of agent's architecture. This may result from the fact that the implementation of various architectures of cognitive agents is currently mainly at a prototype stage and few of them function in commercial solutions and thus the problem of occurrence of knowledge conflicts is not raised. The work [13], for example, presents using intelligent technologies, such as Bayesian Network Case-Based Reasoning, Expert System, Fuzzy System, Genetic Algorithms and Ontology Based techniques for resolving different types of conflicts both reactive and cognitive agents, however they are not related to agents architecture. The work [3] presents cognitive agents resolving methods only on the development stage (design time, programming), the knowledge conflicts at the runtime are not taking into consideration.

However, more intensive development of cognitive agent is noticeable, which may lead to a situation in which the knowledge of these agents will be so extensive that the issue of defining the sources of knowledge conflicts as well as the methods of their solving will become very significant both from a theoretical point of view and from the point of view of persons dealing with designing cognitive agents and multi-agent systems made of them. An automatic solution to the knowledge conflict, as stated in the study [8], is a key element of the functioning of multi-agent systems.

Thus the purpose of this article is to analyse the sources of knowledge conflicts occurring in the architecture of cognitive agent supporting a decision-making process.

II. The Modules Of Cognitive Agent's Architecture As A Potential Places Of Knowledge Conflicts Sources

The most important features of all cognitive agents' architectures include the way of their memory organization and learning mechanisms. The memory is the repository of the knowledge about the world and oneself, the objectives and

978-83-60810-58-3/$25.00

current actions. The role of memory is understood differently by the authors [5, 6, 8, 9, 12]. The organization of the memory depends on the manner of knowledge representation. Learning is a process which transforms the remembered knowledge and the manner of its use. In the study [4] considering the taxonomy of cognitive agent architectures with respect to two above mentioned features, three main groups of the architectures were distinguished:

1. Symbolic architectures which use declarative knowledge included in relations recorded at the symbolic level, focusing on the use of this knowledge to solve problems. This group of architectures includes, among others: State, Operator And Result (SOAR), Executive Process Interactive Control (EPIC), Semantic Network Processing System (SNePS), CopyCat, Non-Axiomatic Reasoning System (NARS), Integrated Cognitive-Neuroscience Architectures for Understanding Sensemaking (ICARUS).
2. Emergent architectures using signal flows through the network of numerous, mutually interacting elements, in which emergent conditions occur, possible to be interpreted in a symbolic way. This group of architectures includes, among others: Neurally Organized Mobile Adaptive Device (NOMAD), Numenta Platform for Intelligent Computing (NuPIC) Cortronics, Brain-Emulating Cognition and Control Architecture (BECCA).
3. Hybrid architectures which are the combinations of the symbolic and emergent approach, combined in various ways. This group of architectures includes, among others: Adaptive Components of Thought-Rational (ACT-R),The Connectionist Learning Adaptive Rule Induction ON-line (CLARION), CogPrime, DUAL, Cortical Capacity-Constrained Concurrent Activation-based Production System (4CAPS), The Novamente AI Engine, Cognitive Agents Architecture (Cougaar), The Learning Intelligent Distribution Agent (LIDA).

It was decided to analyse in this article (due to its volume), with respect to the sources of knowledge conflicts, only the architectures of the LIDA cognitive agent, proposed by S. Franklin [11], presented in the Fig 1. This architecture is of emergent-symbolic nature, owing to which the processing of both structured (numerical and symbolic) knowledge and unstructured (recorded in the natural language) is possible. In addition, the Cognitive Computing Research Group established by S. Franklin, elaborated in 2011 the framework (in Java language) significantly facilitating the implementation of the cognitive agent. It should also be emphasized that the whole framework code is open, i.e. the developer has access to the definitions of all methods, as opposed to, for instance, Cougaar architecture framework software, in which the agent's software code constitutes the so-called "blackbox".

In the LIDA architecture, presented on Fig 1, it was adopted that the majority of basic operations are performed by the so-called codelets, namely specialized, mobile programmes processing in-formation in the model of global workspace. The functioning of the cognitive agent is performed within the framework of the cognitive cycle and it is divided into three phases: the understanding phase, the consciousness phase and the selection of actions and learning phase. At the beginning of the understanding phase the stimuli received from the environment activate the codelets of the low level features in the *sensory memory* [11]. The outlets of these codelets activate the *perceptual memory*, where high level feature codelets supply more abstract things such as objects, categories, actions or events. The perception results are transferred to *workspace* and on the basis of *episodic and declarative memory* local links are created and then, with the use of the occurrences of *perceptual memory*, a current situational model is generated; it other words the agent understands what phenomena are occurring in the environment of the organization. The consciousness phase starts with forming of the coalition of the most significant elements of the situational model, which then compete for attention so the place in the *workspace*, by using *attentional codelets*. The contents of the *workspace* module is then transferred to the *global workspace* (the so-called "broadcasting" is taking place), simultaneously initializing the phase of action selection. At this phase possible action schemes are taken from *procedural memory* and sent to the *action selection* module, where there compete for the selection in a given cycle. The selected actions activate *sensory-motor memory* for the purpose of creating an appropriate algorithm of their performance, which is the final stage of the cognitive cycle [1]. The cognitive cycle is repeated with the frequency of 5 - 10 times per second.

Parallelly with the previous actions the agent's learning is performed (Fig 1), which is divided into *perceptual learning* concerning the recognition of new objects, categories, relations; *episodic learning* which means remembering specific events: what, where, when, occurring in the working memory and thus available in the awareness; *procedural learning*, namely learning new actions and action sequences needed for solving the problems set; *conscious learning* relates to learning new, conscious behaviours or strengthening the existing conscious behaviours, which occurs when a given element of the situational model is often in the *workspace*. The agent's learning may be performed as learning with or without a teacher.

It is worth emphasizing that each cognitive agent supporting decision-making must have the ability of grounding the symbols, namely assign relevant real world objects to specific symbols of the natural language. This is necessary to correctly process unstructured knowledge saved mainly by means of the natural language and thus, for instance, the clients' opinions on products. The knowledge of this type is currently becoming more and more significant for a company because it may have impact on its competitiveness level. For instance analysing the clients' opinions on a given product, the sales volume of a given product in the future may be estimated (of course with a certain level of probability).

Taking into consideration:

- the complexity of the cognitive agent's architecture (Fig 1) and functionality,

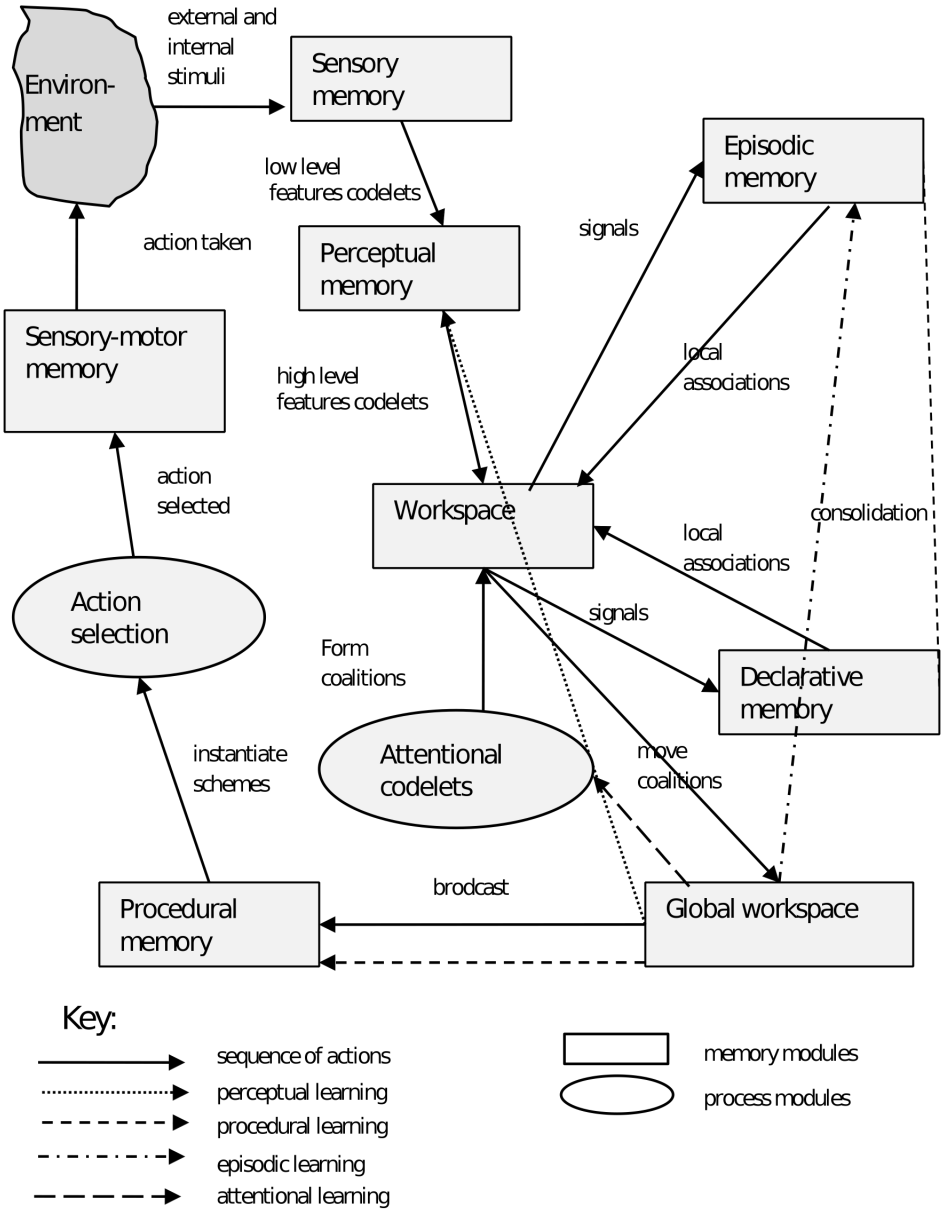


Fig. 1 The architecture of LIDA cognitive agent.
Source: Own work on the basis of [1, 2]

- asynchronous nature of the cognitive cycle (5-10 cycles per second) having impact on the contents of particular architecture models,

it may be concluded that they may constitute the reasons for the occurrence of knowledge conflicts. Potential places of these conflicts' sources may occur in the modules of the cognitive agent's architecture and be connected with:

- the domain of the value of objects stored in memory,
- the results of phenomena interpretation,
- events,
- rules,
- the perception of the current state of the environment (objects and links between them),
- the results of algorithm operation,
- the selection of the agent's actions.

Further in the article the sources of knowledge conflicts will be presented, illustrated with specific examples, connected with supporting the company decision-making process.

III. Sources of Knowledge Conflicts Connected With Supporting The Decision-Making Process

The environment of the functioning of cognitive agents supporting decision-making process, constitutes the company and its environment. The occurrence of knowledge conflicts is related to a situation when various values are assigned to the same objects, links between them, features, phenomena events and actions occurring in the environment of the party to a conflict. Generating various decisions by agents, at the same time, may serve as an example [17].

For the needs of considerations made in the article, the sources of the conflicts of knowledge will be presented at the example of an agent, the objective of which is customer relation management. This agent constantly obtains stimuli from the environment which relate to the characteristics of sales such as sales dynamics indexes in break-up into particular clients (delivered for instance from the agent/agents supporting logistic processes), the clients' opinions of products (being for instance in the company on-line store database), characteristics of products offered by the competition, actions taken by the competition (delivered for instance by the agent monitoring competition). The further part of article describes modules of agent's architecture from the point of view of the occurrence of conflicts of knowledge.

A. Sensory memory

As sales characteristics are stored, on regular basis, in the agent's sensor memory, the contents of this memory may constitute the source of knowledge conflict. These conflicts are mainly connected with the domain of the value of objects stored in the memory. For instance, if it was adopted in the solution that the memory should include the users' opinions recorded in the text form, however there occurs a situation in which the opinion contains graphic elements, their interpretation may be difficult or even impossible. As a consequence, the cognitive agent may incorrectly perceive the current state of the environment.

B. Perceptual memory

Sales characteristics are sent further to the perceptual memory where they are interpreted, for instance determining whether the clients' opinions are positive or negative or determining the difference between the characteristics of products offered by the company in question and the characteristics of products offered by the competition. The knowledge conflicts occurring in the perceptive memory are thus connected with the results of its contents interpretation. For example, if an opinion contains only the product characteristics such as the color, dimensions, the function, it is difficult to determine the polarity of the opinion (state whether the opinion is positive or negative).

C. Workspace, episodic memory and declarative memory

The perception results in the form of objects or events are sent to the workspace in which knowledge conflicts relate to the perception of the current state of the environment and are connected with the creation of local links with the use of events stored in the episodic memory and the rules stored in the declarative memory. Knowledge conflicts connected with the contents of episodic memory mainly relate to contradictory events which occurred as a result of the earlier event. For example, the earlier event recorded in the episodic memory is: "two years before the competition launched two products (which are also manufactured by the company in question) with better characteristics (product 1 and product 2) and later events recorded in the episodic memory include: "in the previous year the sales of product 1 decreased" and "in the previous year the sales of product 2 increased)" The knowledge conflicts occurring in the declarative memory are connected with the occurrence of the contradiction of rules (for instance "if the users' opinions are negative, the decrease in sales will take place", and "if users' opinions are negative, the sales will remain at the same level").

Based on episodic and declarative memory a current situational model is generated, in the workspace, in the form of objects (for example sales characteristics), events (for example the actions of the competition) and links among them (for example: the competition offered a product with better characteristics and in our company a decrease in sales is observed"). Knowledge conflicts occurring in the workspace take place as a result of conflicts occurring in the episodic and declarative memory - the current situational model may contain incorrect objects or incorrect links between them.

D. Attentional codelets

In the attentional codelets module, there are significant elements of the situational model (the agent "rejects" insignificant elements of the situational model such as for instance "the drop of sales of products to the client X occurred because this client liquidated business" - this element is insignificant as no marketing actions can be taken with respect to client X any more). The conflict of knowledge in this module relates to the results of algorithm actions determining which elements of the current situational model are insignificant.

E. Procedural memory

The procedural memory, in turn, contains specific action schemes - for instance "improving the product characteristics", "lowering the product price", "launching the new product meeting the clients' expectations on the market". The conflict of knowledge relates to algorithms implemented as an action scheme, for instance determining what measures should be taken to launch a new product on the market.

F. Action selection module

The knowledge conflict in the action selection module relates to decisions which should be taken, for instance whether the action: "lowering the product price" or "launching a new product meeting the clients' expectations" should be chosen.

G. Global workspace and sensory-motor memory

In the cognitive agent's architecture there are also modules in which the sources of knowledge conflicts do not oc-

cur. They include: the global workspace (sources of knowledge conflicts do not occur in this module because there are significant elements of the situational model transferred from the module of current awareness for the purpose of initiating the phase of action selection) and sensory-motor memory (the sources of knowledge conflicts do not occur in this memory because it is a working module).

H. Discussion

It should be noticed that the occurrence of the conflict results in restrictions in the agent's learning process. For instance, implementing perceptual learning, the agent may learn the interpretation of unknown economic indexes (for instance looking for their interpretation on the Internet - learning without a teacher or using human assistance - learning with a teacher). If, however, the index interpretations found are contradictory, the process of perceptive learning is disturbed.

And implementing procedural learning (learning without a teacher may be applied here as well (the agent may use the actions defined in its own perceptive memory and assigned so far to other elements of the situational model), with a critic (for example the agent may assign particular actions implemented in connection with the decrease in the sales dynamics and a person defines whether the actions are correct) or with a teacher, (the agent may for instance learn what actions should be taken in a situation when sales is dropping in a company and the competition is launching a new product). If the action algorithms are different (a knowledge conflict occurs), the learning process will be disturbed as well.

Knowledge conflicts occurring in the episodic memory have, in turn, a negative impact on episodic learning (performed without a teacher) consisting in remembering all events occurring in the environment.

Conscious learning (performed without a teacher), on the other hand, consisting in determining which elements of a situational model are significant, may be limited by knowledge conflicts occurring both in the workspace and in the attentional codelets module.

It should be also emphasized that the sources of knowledge conflicts may occur in other symbolic, emergent and hybrid architectures of cognitive agents. As similarly as LIDA architecture, their structure consists of many modules.

IV. CONCLUSION

The use of cognitive agent for the purpose of supporting decision-making allows for the implementation of actions performed in a company by a human being so far, starting with the operation of work stations, through the diagnosis of the current economic situation to automatic decision-taking, both at the operational, tactical and strategic level. This is connected with the agents' skills in the scope of correct interpretation and associating of facts, discovering links between the objects and phenomena of the real world, learning and having experience.

For cognitive agent to be able to effectively perform their tasks, they should be created upon conducting the analysis of particular modules, with respect to knowledge conflicts. Thus the identification of the sources of knowledge conflicts, presented in the article, and its consideration at designing the decision-making process support systems will allow for automatic detection of conflicts of this kind and, as a consequence, their solving. These actions are extremely significant because, as has already been emphasized, they have a positive influence on the effectiveness of processes performed by an agent, and, in turn, the effectiveness of decisions taking place in a company.

This results in the need to perform further research works connected with, among others, the elaboration of the formal model of conflict solving and the creation of the prototype of cognitive decision-support system.

REFERENCES

[1] A. Bytniewski, Hernes M., „Wykorzystanie agentów kognitywnych w budowie zintegrowanego systemu informatycznego zarządzania", in: T. Porębska-Miąc, H. Sroka (ed.), Systemy Wspomagania Organizacji, Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach, Katowice 2013.

[2] Cognitive Computing Research Group, http://ccrg.cs.memphis.edu/, access date: 29.01.2014.

[3] M. Dastani , G. Governatori , A. Rotolo , L. Van Der Torre, "Programming cognitive agents in defeasible logic", LPAR 2005, DOI: 10.1.1.76.6973.

[4] W. Duch, Architektury kognitywne, czyli jak zbudować sztuczny umysł, in: R. Tadeusiewicz (ed.) Neurocybernetyka teoretyczna, Wydawnictwa Uniwersytetu Warszawskiego, Warszawa 2010.

[5] J. Hawkins, S. Blakeslee, "On intelligence: How a New Understanding of the Brain will Lead to the Creation of Truly Intelligent Machines", Times Books 2004

[6] R. Hecht-Nielsen, "Confabulation Theory: The Mechanism of Thought", Springer 2007.

[7] M. Hernes, N.T Nguyen, "Deriving Consensus for Hierarchical Incomplete Ordered Partitions and Coverings", Journal of Universal Computer Science, no. 13(2), pp. 317-328, 2007.

[8] M. Hernes M, J. Sobieska-Karpińska, "A comparative analysis of conflicts resolving methods in multiagent decision support systems", in: M. Pańkowska, H. Sroka, S. Stanek (ed.), Cognition and creativity support systems, Studia Ekonomiczne. Zeszyty Naukowe Wydziałowe UE w Katowicach, Wydawnictwo UE w Katowicach, Katowice 2013, s. 23-32.

[9] M. Dastani, L. Van der Torre, "A Classification of Cognitive Agents", in: W.D. Gray, C.D. Schun (ed.), Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society, Cognitive Science Society, Mahwah 2002. DOI: 10.1.1.16.8401.

[10] D. De Long, P. Seemann, " Confronting Conceptual Confusion and Conflict in Knowledge Management", Organizational Dynamics, 2000.

[11] S. Franklin, F.G. Patterson, "The LIDA architecture: Adding new modes of learning to an intelligent, autonomous, software agent". in: Proc. of the Int. Conf. on Integrated Design and Process Technology. San Diego, CA: Society for Design and Process Science, 2006.

[12] A. Hensinger, M. Thome, T. Wright, "Cougaar: A Scalable, Distributed Multi-Agent Architecture", IEEE International Conference on Systems, Man and Cybernetics, 2004. DOI: 10.1109/ICSMC.2004.1399959.

[13] K.M. Khalil, M. Abdel-Aziz, T. T. Nazmy T, A. M. Salem, "Intelligent Techniques for Resolving Conflicts of Knowledge in Multi-Agent Decision Support Systems", Sixth International Conference on Intelligence Computing and Information Systems, Cairo, Egypt, 2013, DOI: arXiv:1401.4381.

[14] W. Lorkiewicz, "An approach to resolving semantic conflicts of temporally-vague observations in artificial cognitive agent", Proceedings of the 10th international conference on Knowledge-Based Intelligent Information and Engineering Systems - Volume Part III, Springer-Verlag 2006 pp. 1004-1011. DOI: 10.1007/11893011_127.

[15] B. Mianowska, N. T. Nguyen, "Using knowledge integration techniques for user profile adaptation method in document retrieval systems", Transactions on computational collective intelligence V, Springer-Verlag Berlin, Heidelberg, 2011.

[16] NT. Nguyen, R. Katarzyniak, “Multi-agent Systems, Ontologies and Conflict Resolution”. Special issue in Journal of Intelligent & Fuzzy Systems 17(3), 2006.
[17] J. Sobieska-Karpińska, M. Hernes, “Consensus determining algorithm in multiagent decision support system with taking into consideration improving agent's knowledge”, Federated Conference Computer Science and Information Systems (FedCSIS), 2012.
[18] J. Sobieska-Karpińska, M. Hernes, “The postulates of consensus determining in financial decision support systems”, in: Annals of Computer Science and Information Systems, Proceedings of Federated Conference Computer Science and Information Systems (FedCSIS), Kraków, 2013, ISSN 2300-5963, s. 1165 - 1168.
[19] R. Tadeusiewicz, „ Systemy kognitywne - nowy wymiar informatyki ekonomicznej”, http://ryszardtadeusiewicz.natemat.pl/75001,systemy-kognitywne-nowy-wymiar-informatyki-ekonomicznej [29.12.2013].
[20] R. Yager., “Approximate reasoning and conflict resolution”, Machine Intelligence Institute, Iona College, 2000.
[21] M. Żytniewski, R. Kowal, A. Sołtysik, “The outcomes of the research in areas of application and impact of software agents societies to organizations so far. Examples of implementation in Polish companies”, in: Annals of Computer Science and Information Systems, Proceedings of Federated Conference Computer Science and Information Systems (FedCSIS), Kraków, 2013 s. 1165 - 1168.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1163–1170
DOI: 10.15439/2014F271
ACSIS, Vol. 2

On Winners and Losers in Procurement Auctions

Gregory (Grzegorz) E. Kersten
InterNeg Research Centre
Concordia University
Montreal, Canada
gregory@jmsb.concordia.ca

Tomasz Wachowicz
Department of Operations Research
University of Economics
Katowice, Poland
tomasz.wachowicz@ae.katowice.pl

Abstract—**The use of auctions in procurement results in price reduction as well as the reduction in the cost and the time required to complete transactions. In many situations, price-only auctions resulted in the violation of the contracts or even contractors' bankruptcies. This is one reason for the introduction of multi-attribute auctions as well as auctions-followed-by negotiations. Auction theory is based on two assumptions which assure that auctions are efficient mechanisms, guarantee efficient solutions, and produce the best possible results for bid-takers. In practice these assumptions are often violated. The contribution of this paper is to propose a procedure for auctions-followed-by negotiations which retains important auction features such as process transparency and efficiency while allowing for increased social welfare. The unique feature of the procedure is the introduction of the win-win phase in which the market participants may attempt to make joint improvements to efficient solutions obtained from auctions.**

I. Introduction

Government-to-business (G2B) and business-to-business (B2B) online revers auctions have been introduced in mid-1990s to complement the paper-based submissions. They have become a popular way to source products and services - a few years after their introduction, 25% of total purchasing was done using these auctions [1]. Their use resulted in price reduction as well as the reduction in costs and time required to complete transactions. Other benefits attributed to these auctions include the creation of new markets, increased participation of suppliers, increased transaction transparency and price visibility, and increased standardization and efficiency of purchasing [2, 3]. These advantages have been contrasted with such disadvantages as collusion, opportunism and coercion [4]. However, these drawbacks are not limited to the auction mechanisms. Catalogues and negotiations have been also found prone to collusion, coercion and opportunism as well as deceit and threats.

Many of the negative, often illegal, aspects of market mechanisms can be addressed through the imposition of organizational and legal frameworks. Organizational framework may be used to enforce transparency, information disclosure, and standardization. Laws can be used to make collusion and coercion difficult as investigations uncover criminal behavior, for example such as bids rigged by Montreal Mafia, which controlled most of road contracts [5].

The underlying theory of the exchange mechanisms appears to favor auctions. Bulow and Klemperer [6] in one of the first comparative studies of auctions and negotiations prove that given a choice between auctions with n+1 bid-makers and negotiations with n negotiators the bid-maker should always choose an auction. This implies that the value of increased competition in auctions exceeds the negotiator's skill and prowess.

There are three key conditions which need to be met in order for the exchange mechanisms to be considered efficient, namely:

1. MAE: *Mechanism allocative efficiency*: the requirement that the mechanism maximizes social welfare. This condition assures that the use of the mechanism is optimal in the sense that its use does not result in social welfare loss;
2. SOE: *Solution efficiency*: the requirement that the solution that is the result of the mechanism usage is efficient. This condition is the specific instance of the allocative efficiency; and
3. OCM: *The owner's criterion maximization*: the requirement that the mechanism results in the solution that is the best possible for the owner, namely the organizations that set the mechanism up and invites others to voluntarily use it in an exchange process. This condition means that the mechanism's users have incentives that lead them to propose solutions that favor the owner leading to the final (winning) solution that is the best the owner could achieve in the particular circumstances.

Auctions in which the sole criterion is price or any other single and linear measure may meet the above conditions providing that the participants are risk-neutral. However, in many B2B and G2B auctions over production goods and services the evaluation criteria are more complex. These criteria include multiple attributes, which may be different for the

This work has been supported by the grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) and Concordia University.

bid-taker and the bid-makers. In fact, surveys of procurement managers as well as field studies find that procured goods (i.e., products and services) are described by price as well as non-price attributes, which range from 2 to 30 [7-10]. In addition, one may argue that in procurement undertaken by public organizations attributes which describe changes in the society's well-being caused by contracting and its implementation should be considered. An example of such a consideration is A+B bidding discussed in Section 2.

The use of multiple attributes does not necessarily imply that price-only auctions cannot be effectively used in B2B and G2B auctions. It does however, raises a question about their appropriateness, in particular, about auctions being able to meet the above key conditions: MAE, SOC and MOC. Given that auctions are used in transactions amounting to billions, their potential inefficiency is of great importance to public and private organizations.

The purpose of this work is to discuss single and multi-attribute auctions conducted in B2G and G2B exchanges and to show that in many situations auctions may result in the winning bid which is an efficient solution (SOE condition) but they are allocative inefficient mechanisms (i.e., they violate AEM condition). Furthermore, we show that if the two assumptions underlying—which is likely in real-life situations—then the winning bid may be replaced with a solution which improves both the bid-taker's and the bid-maker's criteria.

There are three more sections in the paper. In Section 2 we formulate formal conditions that are required for multi-attribute auctions to produce better results than one-attribute auctions. In Section 3 the discussion of two underlying assumptions of auction theory is contrasted with the assumptions' implications and their limitations in real-life auctions. Section 4 concludes with a brief discussion of a process that can be used to transform auctions to hybrid mechanisms which would yield higher utility for the bid-takers (owners) and higher social welfare.

II. One-, two-, and many-attribute auctions

In many auctions, including procurement auctions, bidders submit only price. There are also auctions in which bids are vectors of multiple attributes. In this section price-only auctions are briefly discussed followed by a discussion on two- and more-attribute auctions. Two-attribute auctions are distinguished because they are simple to use and seemingly assure bid-takers that they obtain goods at the lowest price.

A. Bidding on price

Many procurement contracts are awarded to qualified bid-makers who offer the lowest price and who can meet the delivery time and other pre-specified conditions. There are also many auctions which do not result in a contract. These auctions are used to determine price and are followed by negotiations [11]. The reason for the auction-followed-by-negotiation mechanism and buyer-determined auctions is that the auction theory and the results from the experimental economics are difficult to implement in procurement [12].

Nelken [13], a spokesperson for the Polish General Directorate of National Roads and Motorways (GDNRM) observes: "If all conditions stated in the contract are met, the price is the best criterion for contractor selection." This is the case if the bid-taker uses only price as the criterion for the good's assessment. In reality, often the attributes which values are given to the bid-makers are criteria rather than bounds. For example, GDNRM prefers shorter road completion time than longer and longer warranty period than shorter. The use of price-only auction in which bidders must observe the conditions does not allow the bid-makers to compete on non-price attributes.

If there are more than one attributes that the bid-taker uses, then the imposition of constraints on the good's attributes, other than price, is insufficient to claim that the winning bid-maker offers the best contract.

B. Two-attribute auctions

Two attribute auctions increase the exchange flexibility because the bid-makers can tradeoff value of one attribute against the second attribute value. They may also be necessary when the bid-taker is obliged to or wants to obtain a contract which optimizes two rather than one criteria. In order to show that in this case two-attribute may produce results that are superior to a single attribute auction.

Let:

$x_1 \geq 0$ be the price attribute and $x_2 \geq 0$ – the second (non-price) attribute;

$u_b(\boldsymbol{x})$ (where $\boldsymbol{x} = [x_1, x_2]$) be the assessment function of the bid-taker b which b wants to minimize, i.e., it is preferable that both x_1 and x_2 take low values.

In real-life auctions $u(x)$ is often a linear function, i.e.,

$$u(\boldsymbol{x}) = x_1 + a \cdot x_2, \ (a > 0). \tag{1}$$

Let's now consider two bid-makers $s1$ and $s2$ who have different capabilities. Assume that the best bid of s_1 is $(x_1^{s1}; x_2^{s1})$ and s_2's best bid is $(x_1^{s2}; x_2^{s2})$, and $x_1^{s1} < x_1^{s2}$. Bid-maker s_1 wins the price-only auction. If

$$x_1^{s1} + a \cdot x_2^{s1} > x_1^{s2} + a \cdot x_2^{s2},$$

then s_2 wins the two-attribute auction. This means that the higher price offered by s_2 is offset by the lower value of the second attribute weighted by a, i.e.,

$$x_1^{s1} - x_1^{s2} > a \cdot (x_2^{s2} - x_2^{s1}).$$

Condition (1) corresponds to a well-known and observed in practice situation when price leaders may be unable to provide goods or deliver them in a shorter time than suppliers who charge higher price.

The gains for the bid-taker in two-attribute auctions over the price-only auctions have been confirmed experimentally [14]. Although in experimental settings these gains were found to be modest, in the comparison of real-life auctions these gains have been found significant. Lewis and Bajari [15] studied over 1300 hundred contracts awarded by the California Department of Transportation, through one- and two-attribute auctions. The two-attribute auctions where of the A+B

type, where A represents price and B represents the total number of days required to complete the contract weighted by the user cost, which is the cost incurred by the road users who have to take alternative routes. Lewis and Bajari used structural analysis to estimate the counterfactual welfare gain from switching from A (price-only) to A+B (price and social costs). They report that this gain represents almost 22% of the total contract value of $1.14 billion. In addition, the contracts were completed 30-40% faster. These are significant savings for the bid-takers; as of 2003, 38 U.S. states were using auctions with scoring functions for large projects [16].

Lewis and Bajari [15] made another important observation. In both one- and two-attribute auctions the same businesses participated and many of them won both types of auctions. Although the winners were paid about 7% more in two-attribute as opposed to one-attribute auctions, the bid-takers' savings greatly exceeded these additional costs. Another observation is that the contractors (bid-makers) are able to be significantly more efficient when they have an incentive to do so. From the social perspective this is a significant result because in this way public organizations can not only reduce their total costs but they can contribute to the overall efficiency increase of the industry.

C. Multi-attribute auctions

The superiority of the results obtained through two-attribute rather than one-attribute auctions can be extended to exchanges characterized by multiple attributes. Several multi-attribute auctions are discussed in literature. Hohner, Rich et al. [17] discuss an electronic private exchange established by Mars Inc., in which volume discount bidding and multi-attribute bidding were used most often.

The attributes that were used in the auctions included payment terms (e.g., pre-payment, payment date, and discount) as well as turnaround time, delivery schedule, product quality, type of material, and color.

Trade Extensions (TradeExtensions.com) offers a procurement software platform which includes reverse auctions. A review of four procurement case studies (i.e., Ineos, Road resurfacing, Elderly Care Services, and Cleaning Services) shows that Trade Extension's uses a full costing process in which all attributes must be expressed in monetary terms. The focus is on the minimization of the costs of procured services subject to constraints imposed on the attributes and their combinations.

It has been recognized recently that public organizations ought to base their procurement decisions on multiple attributes and that this should be made explicit to their suppliers. The European Union has adopted a new public procurement directive, which requires that the procurement authority publish ex ante relative weighting of each criterion. The E.U. directives (Article 55 in 2004/17/EC or Article 53 in 2004/18/EC) require that public contracts be allocated by competitive bidding. The buyer has to either use a scoring function, in which price and other attributes and their weights are given, or a lexicographically ordered list of attributes. In Poland, some of the road construction auctions are still awarded solely based on price [13] but many of these auctions resulted in contractors' bankruptcy [18]. This may be one reason for the introduction attributes additional to price, such as completion time and warranty period [19].

III. THREE CRITERIA

All cases discussed in the preceding sections concern products and services that will be produced and delivered in the future. This characteristic must be contrasted with situations in which auctions are over earlier produced goods. The difference is that in post-auction, manufactured products and services can be customized to meet specific needs of the bid-takers. This can be exemplified with the A+B auctions discussed in Section 2.2 in which bid-makers increased price on average by 7% in exchange for the reduction of the contract fulfilment time. This may be seen as a standard business practice as it brings increase of quality, shortening of delivery time, and addition of more features, however, production costs are increased and thus the tree conditions formulated in Section 1 are violated.

A. Two assumptions and their implications

In Section 1 the three key conditions (i.e., MAE, SOC and OCM) for efficient exchange mechanisms were formulated. Auctions meet all the conditions if the following two assumptions about the bid-makers and bid-takers are met [20, 21]:

1. The bid-maker and the bid-takers are risk neutral; and
2. The bid-maker and the bid-takers employ an evaluation functions (e.g., utility, profit, and value functions) which are quasi-linear.

Attitude towards risk influences, among others, the way payoffs are considered. For a risk-neutral person every unit of money has the same value (irrespectively if it is one unit more or less) as long as there is no difference in risk associated with getting more (less) of units. If both bid-takers and bid-makers are risk neutral, then they have the same assessment of the price. If however, their risk attitudes differ, then the same price level will be seen differently.

Violation of risk-neutrality causes that the market participants' utility is not quasi-linear however, this is not a sufficient condition. Quasi-linearity means that that the market participants' utility is the sum of price (i.e., the numeraire) and the valuation function of all non-price attributes. The valuation function is strictly convex and twice differentiable (Varian 2010). The implication of the quasi-linearity assumption is that the valuation function is can be expressed in monetary terms.

Attitude towards risk: and the price cannot be separated from valuation [21, pp. 32-41]. The assumption that participants are risk neutral is often unrealistic; risk aversion has been used to explain overbidding behavior [22]. In procurement auctions, sellers of timber and construction firms were found to be risk averse [23, 24]. Procurement managers in public organizations were found more risk averse than their counterparts in private organizations [25]. (The risk neutrality

assumption is required in both price-only and multi-attribute auctions.)

Buyers who base their purchasing decisions on the long-run price and other direct and indirect costs or on the total cost of ownership (TCO) model, consider many attributes. The middle- or long-term perspective lends itself to associating money with time, which includes future interest paid and various types of risk (e.g., delayed or not delayed payment, litigation, and change in interest). Because different participants are likely to have different financial, market and production constraints their preferences over money may also differ.

Quasi-linearity utilities: Quasi-linearity is a strong assumption and not particularly realistic [26, p. 63]. Ausubel and Milgrom [27, p. 24] note that the assumption is very restrictive and "requires that there is no effective budget limit to constrain the bidders and that the buyer, in procurement auction, does not have any overall limit on its costs of procurement. Although we have no data on how frequently these assumptions are satisfied, it appears that failures may be common in practice."

When market participants' utilities are quasi-linear, then the efficient transactions, i.e., these that lie on the contract curve of a bid-taker and a bid-maker differ in price value but not in the configuration of attribute values [28]. This may be difficult to accept when the auction is not over goods produced earlier (in which case their costs are fixed) but over goods that are produced only after the auction successfully concludes. Other limitations include such requirements as: (1) the preferential order of attribute values has to be opposite for the bid-taker and the bid-makers; there may not be constraints which bind efficient solutions; and (3) price has to be either a single attribute or a sum of price attributes with exactly the same weights for the bid-taker and the bid-makers [29].

B. Convex/concave utilities of buyers and/or sellers

Given the above, we may assume that many market participants' utilities are not quasi linear and that they are not risk-neutral. The implication of this assumption is that the auctions with risk seeking or risk averse participants do not meet at least one of the mechanism efficiency condition. One may argue that private businesses, who are bid-takers, may not be concerned with the condition violation as long as the OCM condition is met, i.e., the winning bid maximizes their utility. This argument may be questioned when the bid takers are public organizations which should be concerned with the efficient use of resources (SOE condition) and maximization of social welfare (MAE condition).

If the utilities are convex or quasi-convex or, more generally, if the efficient frontier for the bid-taker and the winning bid-maker is concave, then the OCM and MAE conditions may not be met. In such situation the winning bids may be efficient solutions but they can be improved for both the bid-taker and the winning bid-maker. This is because concave efficient solutions create an opportunity to introduce trade-offs which increase utility value of both of them. This potentially significant situation is illustrated in Figure 1.

Consider a winning bid (A); it is a winning bid for both quasi-linear utilities, its efficient frontier is shown by a broken line, and utilities for which the efficient frontier is concave. Let's assume that seller i is the winner and the winning bid is point A. If the preferences are quasi-linear, then A maximizes both the buyer's surplus and social welfare. If the efficient frontier is concave, then A does not maximize social welfare; both B and D yield a higher social welfare than A.

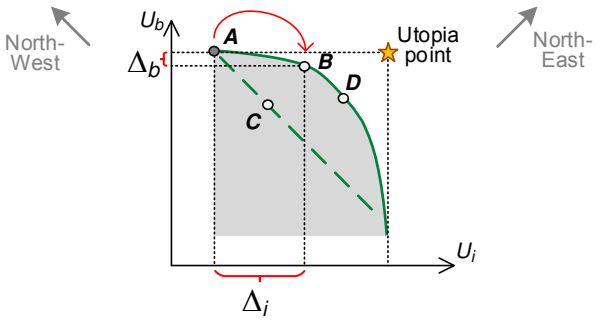


Figure 1. Improvement of the winning bid for concave efficient frontier

Market participants who want to maximize social welfare need to move in the North-East direction. Sellers, who are pushed by competition to increase the buyer's surplus, move in the North-West direction. Quasi-linear preferences together with the use of the sum of utilities as the measure of social welfare, remove the conflict in directions because the North-West moves do not change the distance from the Utopia point (max u_b; max u_i). However, market participants should be aware of the conflict as it arises when other types of preferences and/or other welfare measures are deemed more suitable.

The alternatives shown in Figure 1 have the following coordinates (u_b, u_i): A = (17; 3); B = (16; 11); C = (12; 7.5); and D = (13; 14.5). If we move from A to B, then social welfare increases from 20 to 27, i.e., by 35%. The maximum social welfare is 27.5 and it is reached at alternative D. This simple example illustrates that the difference in social welfare value may be significant.

Moreover, reaching a solution which is better than the winning bid may be possible. This, however, requires moving beyond the initial problem formulation. Let's assume that u_b and u_i are both expressed in monetary terms. We can see that the move from A to B results in buyer b's loss $(u_b(A) - u_b(B)) = \nabla_b$ = \$1 and seller i's gain $(u_i(B) - u_i(A)) = \nabla_i$ = \$8. If buyer b realizes the differences between the winning bid A and the alternative B, then she could suggest selecting alternative B under the condition that i pays her \$5 (or some other amount, which exceeds \$1).

When the utilities are not quasi-linear, then price transfer affects welfare. These utilities may also be assumed to be transferable (this assumption is often made in economics). If the amount to be transferred is positive, i.e., $\nabla_i - \nabla_b > 0$, then the winning bid A may be improved.

When the efficient frontier is concave, the move from the winning bid (*A* in Fig. 1) to a bid that increases social welfare (*B*) involves a transfer which is similar to price transfer when the frontier is linear with -1 slope, except for the following three differences:

1. Both price and configuration are included in the transfer and in social welfare calculation;
2. The transfer of value from the seller to the buyer requires a change of the configuration; and
3. The value transfer improves the buyer's and the seller's surplus as well as the social welfare.

Value θ that is transferred to the buyer has to be greater than the buyer's loss (i.e., $\theta > \nabla_b$). The difference between seller's *i* gain and buyer's *b* loss, i.e., $\nabla_i - \nabla_b > 0$, corresponds to the social welfare increase. Assuming that the concavity of the efficient frontiers for the pairs buyer$_b$-seller$_i$ ($i \in I$) is given and does not change, the size of the transferable value ($\nabla_i - \nabla_b$) can be viewed as the "value of competition". This is because the stronger the competition the greater the buyer's surplus, that is, the winning bid is further from the solution maximizing social welfare. To ascertain this let's denote the concave efficient frontier as function of u_i, i.e., $\upsilon_b(u_i)$ and assume that $\upsilon_b(u_i)$ is twice differentiable. We also assume that the buyer's utility produced by the winning bid is not smaller than the utility which maximizes social welfare.

Given concave efficient frontier, the greater the utility value of the buyer for the winning bid, the greater the transferable value ($\nabla_i - \nabla_b$). Function $\upsilon_b(u_i)$ is concave, therefore its second derivative is non-positive ($\upsilon''_b \le 0$). This means that the rate of increase of $\upsilon_b(u_i)$ decreases with the increase of u_i. Conversely, the rate of increase of $\upsilon_b(u_i)$ increases when u_i decreases. In other words, a small change in u_i causes an increasingly greater change in $\upsilon_b(u_i)$ as u_i gets smaller.

IV. Auctions and Multi-Bilateral Negotiations

A concave efficient frontier is the result of convex utilities. Also pairs of linear utilities defined over a convex set of feasible alternatives form a quasi-concave frontier under the condition that different attributes are more important for the participants [30]. In this case, the auctions which have been discussed in literature, are likely allocative inefficient mechanisms, i.e., they do not maximize social welfare. This may result in an efficient solution that maximizes the bid-taker's surplus. However, it may be possible, as the discussion given in Section 3.2 indicates, to improve this solution for both the buyer and the seller.

The competitive force pulls the bid-makers towards the maximum value of the bid-taker utility. If the efficient frontier is concave, then we should consider the winning bid as a tentative but not the final solution. Rather than conclude the exchange, the participants may continue and seek win-win solutions as shown in Figure 1 (points B and D). In this later phase negotiation may be preferable because the bid-taker needs to get involved and propose potential alternatives as well as the required compensation.

Kersten, Wachowicz et al. [31] discuss a multi-bilateral negotiation procedure with verifiable and non-verifiable offers. The procedure is implemented in the Imbins system which is a multi-bilateral negotiations support system, however, it also allows running auctions. This is because in the verifiable offer option the best bid is automatically displayed to all bid makers. This allows the bid-taker to announce the auction rules to all participants at the beginning of the process and then withhold her participation until a winning bid is obtained. After the auction concludes successfully the bid taker reviews the alternatives, namely the bids preceding the winning bid, and selects one or more of them. For every selected alternative she determines the minimum compensation. The compensation is required in order to compensate the bid-taker for the loss she would incur if the winning bid was replaced with the selected alternative.

The determination of the type of the compensation may require that the bid-taker discusses the issue with the bid-makers because it is likely that it concerns attributes which were different from those which were used in the bidding process. When the compensation is established, the bid-taker determines the values associated with each alternative and submits these pairs for the bid-maker's consideration. This may initiate an auction or a negotiation process which, if successful, results in a solution that is better for both the bid-taker and one of the bid-makers.

The Imbins system, which provides verifiable offers, was experimentally compared with a multi-attribute auction system Imaras (Kersten, Wachowicz et al. [31]. The results show that some of the experiment participants (bid-takers) initially used Imbins as an auction system and subsequently became involved and negotiated with their counterparts (see Apendix A for a screen shot). Other bid-takers may have ignored the fact that the best offer (bid) they received was displayed to every bid-maker and they used Imbins as a negotiation support system. In effect, the bid-makers who used Imaras (see Apendix B for a screen shot) achieved significantly lower profit than those who used Imbins. On the other hand, the bid-takers' profit was significantly higher in Imaras than in Imbins. This is despite the fact that in both systems the bid-makers received the same the verified information.

This result shows that the bid-taker's active participation causes them to accept solutions that are worse when they do not participate in the exchange process. The likely explanation can be found in the reciprocity theory which posits that peoples' social upbringing leads them to reciprocate to both positive and negative acts made by other people [32]. Thus, when the bid-makers make concessions, then the bid-takers reciprocate and also make concessions. We conclude that reciprocity in multi-bilateral negotiations weakens the competitive forces and they are left intact in auctions.

Other results of the Imbins and Imaras comparative studies experimentally confirm the propositions made above. Social welfare obtained in verifiable negotiations significantly exceeds welfare obtained in auctions. More importantly, the potential improvement of the negotiated contracts is not only

significantly greater than contacts obtained through auctions but it is greater than the average welfare of the efficient solutions that dominate the contracts from auction. Social welfare in auctions can reach 42 monetary units (m.u.) on average while social welfare in the negotiations can reach 120 m.u. The difference of 78 m.u. can be distributed between the bid-taker and the winning bid-maker making then both better in negotiations than in auctions. Even if the auction winner receives revenue 0 m.u. and the bid-taker gets 42 m.u., they both lose compared to the bid-maker who gets 50 m.u. and the bid-taker who gets 70 m.u.

V. Winners and losers

In many procurement exchanges, both buyers and sellers make decisions based on many attributes rather than price only. A review of recent studies on single- and multi-attribute auctions indicates that multi-attribute auctions produce better results for both the bid-takers and the bid-makers than a single-attribute auction. When, however, multi-attribute bid evaluation functions are used (e.g., utility, profit, and total cost of ownership), then these functions are likely to violate the two key auction theory assumptions. These assumptions assure that auctions are efficient mechanisms, which allow maximization of the social welfare. They also allow maximization of the bid-takers' criterion.

We showed that if the "bid-taker and bid-maker" pairs of the evaluation functions form a quasi-concave efficient frontier, then the winning bid does not maximize social welfare. This means that, in these situations, auctions are not efficient exchange mechanisms. This is an important conclusion for public organizations because their mission is not limited to achieving the best deal for them but which is detrimental to the socio-economic growth. Public organizations may present these winning deals as the most economic (e.g., the cheapest), but this would be a narrow myopic perspective. Their stakeholders, that is society, are the losers. In extreme cases, these loses may be due to the contractors' bankruptcies and foreclosures. In less severe situations, the contractors loses lead to unemployment and underutilization of resources.

When the efficient frontier is concave, the winning bid may be improved not only for the winning bid-maker but also for the bid-taker. The improvement is possible even if the bid is efficient. Given that the bid-maker compensates the bid-taker for the loss, the bid-taker may accept an efficient alternative which is significantly better for the bid-maker than the winning bid. This means that the bid-taker who accepts the auction's winning bid loses because he/she could achieve a better bid. This result is important for both public and private organizations because they forgo the possible improvement. It is also important for suppliers (bid-makers) who could get better contracts than the contracts specified in the winning bids.

Acknowledgments

We thank Margaret Kersten, Norma Paradis and Bo Yu for their contribution to the system design and experiment organization.

References

[1] Kaufmann, L. and C.R. Carter, *Deciding on the Mode of Negotiation: To Auction or not to Auction Electronically.* Journal of Supply Chain Management, 2004. **40**(2): p. 15-26. DOI: 10.1111/j.1745-493X.2004.tb00166.x

[2] Schoenherr, T. and V.A. Mabert, *Online Reverse Auctions: Common Myths versus Evolving Reality.* Business Horizons, 2007. **50**(5): p. 373-384. DOI: 10.1016/j.bushor.2007.03.003

[3] Smart, A. and A. Harrison, *Reverse Auctions as a Support Mechanism in Flexible Supply Chains.* International Journal of Logistics, 2002. **5**(3): p. 275-284. DOI:10.1080/1367556021000026718

[4] Bajari, P. and G. Summers, *Detecting collusion in procurement auctions.* Antitrust LJ, 2002. **70**: p. 143.

[5] Perreaux, L. and R. Seguin, *Montreal Mafia controls 80 per cent of road contracts, whistleblower says*, in *The Globe and Mail*2009: Toronto.

[6] Bulow, J. and P. Klemperer, *Auctions versus Negotiations.* American Economic Review, 1996. **86**(1): p. 80-194. DOI: 10.3386/w4608

[7] Ferrin, B.G. and R.E. Plank, *Total Cost of Ownership Models: An Exploratory Study.* Journal of Supply Chain Management, 2002. **38**(3): p. 18-29. DOI: 10.1111/j.1745-493X.2002.tb00132.x

[8] Plank, R.E. and B.G. Ferrin, *How manufacturers value purchase offerings: an exploratory study.* Industrial Marketing Management, 2002. **31**(5): p. 457-465. DOI: 10.1016/S0019-8501(01)00161-4

[9] Gupta, D., E.M. Snir, and Y. Chen, *Contractors' and Agency Decisions and Policy Implications in A+ B Bidding.* Production and Operations Management, 2014: p. (in print).

[10] Iimi, A., *Multidimensional auctions for public energy efficiency projects : Evidence from the Japanese ESCO market*, in *Policy Research Working Paper*2013, The World Bank. p. 47.

[11] Engelbrecht-Wiggans, R., E. Haruvy, and E. Katok, *A Comparison of Buyer-determined and Price-based Multi-attribute Mechanisms.* Marketing Science, 2007. **26**(5): p. 629–641.

[12] Aloini, D., R. Dulmin, and V. Mininno, *E-reverse Auction Design: Critical Variables in a B2B Context.* Business Process Management Journal, 2012. **18**(2): p. 3-3. DOI 10.1108/14637151211225180

[13] Nelken, U. *GDDKiA: Zarzuty NIK są bezpodstawne*. 2013 [cited 2014 Apri 15]; Available from: http://forsal.pl/artykuly/702197,gddkia_zarzuty_nik_sa_bezpodstawne.html.

[14] Bichler, M., *An Experimental Analysis of Multi-attribute Auctions.* Decision Support Systems, 2000. **29**(3): p. 249-268. DOI: 10.1016/S0167-9236(00)00075-0

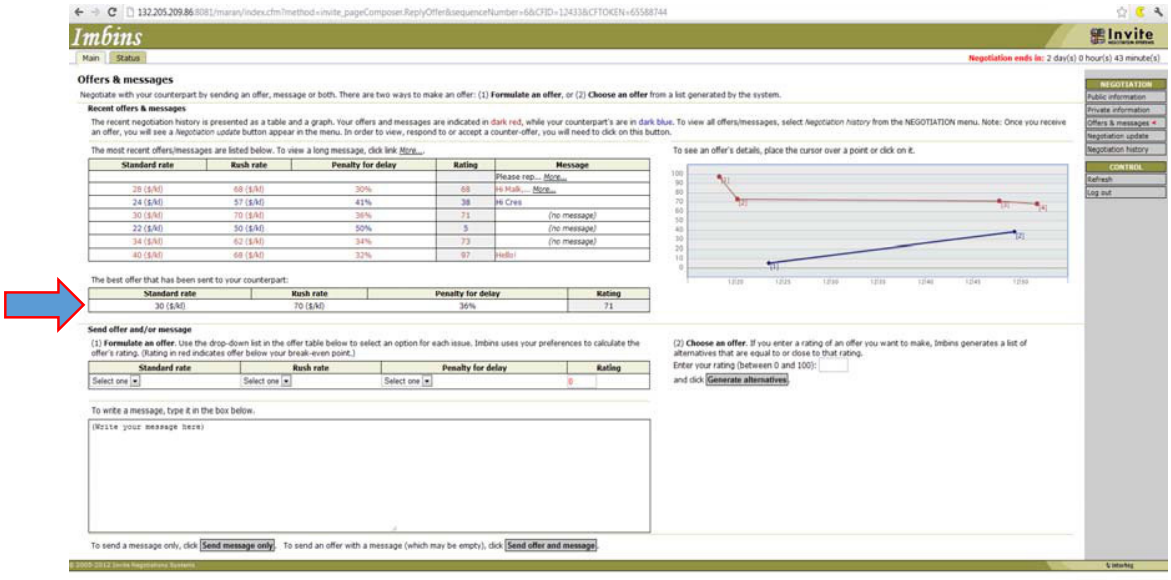
[15] Lewis, G. and P. Bajari, *Procurement Contracting with Time Incentives: Theory and Evidence.* The Quarterly Journal of Economics, 2011. **126**(3): p. 1173-1211.

[16] Asker, J. and E. Cantillon, *Procurement when Price and Quality Matter.* The RAND Journal of Economics, 2010. **41**(1): p. 1-34.

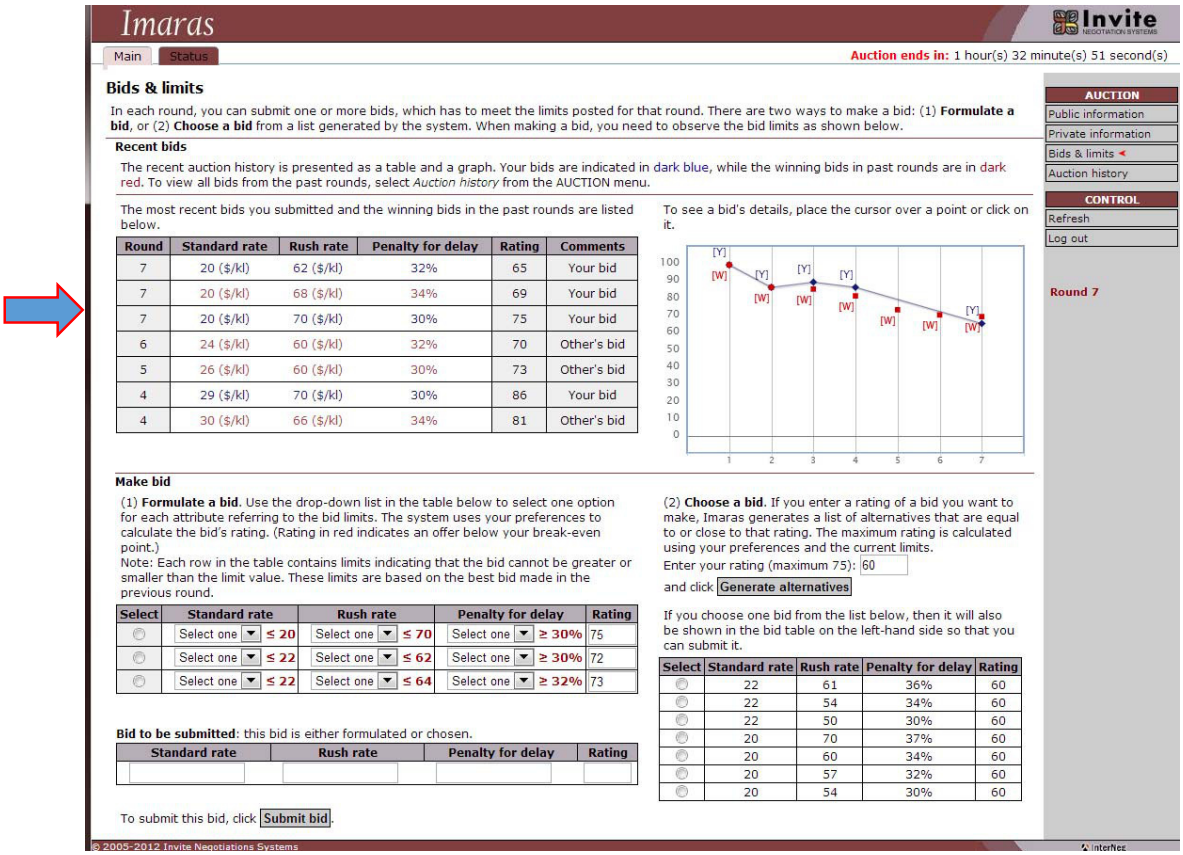
[17] Hohner, G., et al., *Combinatorial and quantity-discount procurement auctions benefit Mars, Incorporated and its suppliers.* Interfaces, 2003. **33**(1): p. 23-35.
[18] Majszyk, K. *Plaga upadłości wśród wykonawców autostrad to wina GDDKiA*. 2013 [cited 2014 Apri 15]; Available from: http://forsal.pl/artykuly/702992,plaga_upadlosci_wsrod_wykonawcow_autostrad_to_wina_gddkia.html.
[19] Majszyk, K. *Nowatorski pomysł GDDKiA: premie za autostrady przed czasem*. 2013 [cited 2014 Apri 15]; Available from: http://forsal.pl/artykuly/672280,nowatorski_pomysl_gddkia_premie_za_autostrady_przed_czasem.html.
[20] Handfield, R.B. and S.L. Straight, *What Sourcing Channel is Right for You?* Supply Chain Management Review, 2003. **7**(4): p. 63-68.
[21] Krishna, V., *Auction Theory*. 2010: Academic Press.
[22] Bajari, P. and A. Hortacsu, *Are Structural Estimates of Auction Models Reasonable? Evidence from Experimental Data.* Journal of Political Economy, 2005. **113**(4): p. 703-741. DOI: 10.1086/432138
[23] Athey, S. and J. Levin, *Information and competition in US forest service timber auctions.* Journal of Political Economy, 2001. **109**(2): p. 375-417. DOI: 10.3386/w7185
[24] Campo, S., *Risk Aversion and Asymmetry in Procurement Auctions: Identification, Estimation and Application to Construction Procurements*. Journal of Econometrics, 2012. **168**(1): p. 96-107. DOI: 10.1016/j.jeconom.2011.09.011
[25] Boyne, G.A., *Public and Private Management: What's the Difference?* Journal of Management Studies, 2002. **39**(1): p. 97-122. DOI: 10.1111/1467-6486.00284
[26] Varian, H.R., *Intermediate Economics. A Modern Approach*. 8 ed. 2010, New York: Norton.
[27] Ausubel, L.M. and P. Milgrom, *The Lovely but Lonely Vickrey Auction*, in *Combinatorial Auctions*, P. Cramton, Y. Shoham, and R. Steinberg, Editors. 2006, MIT Press: Boston. p. 17-40.
[28] Strecker, S., *Information Revelation in Multiattribute English Auctions: A Laboratory Study*. Decision Support Systems, 2010. **49**(3): p. 272-280. DOI: 10.1016/j.dss.2010.03.002
[29] Kersten, G.E., *Multi-attribute Procurement Auctions: Efficiency and Social Welfare in Theory and Practice.* INFORMS Decision Analysis Journal, 2014: (in print).
[30] Mumpower, J.L., The Judgement Policies of Negotiators and the Structure of Negotiation Problems. Management Science, 1991. **37**(10): p. 1304 - 1324. doi.org/10.1287/mnsc.37.10.1304
[31] Kersten, G.E., T. Wachowicz, and M. Kersten. Multi-attribute Reverse Auctions and Negotiations with Verifiable and Non-verifiable Offers. in Federated Conference on Computer Science and Information Systems. 2013. Kraków: IEEE.
[32] Cropanzano, R. and M.S. Mitchell, *Social Exchange Theory: An Interdisciplinary Review.* Journal of Management, 2005. **31**(6): p. 874-900. DOI: 10.1177/0149206305279602

APPENDIX

A. Screenshot of an Imbins interface with best offer shown



B. Screenshot of an Imaras interface with best offer shown



Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1171–1180
DOI: 10.15439/2014F188
ACSIS, Vol. 2

Performance evaluation of decision-making agents' in the multi-agent system

Jerzy Korczak, Marcin Hernes, Maciej Bac
Wrocław University of Economics ul. Komandorska 118/120, 53-345 Wrocław, Poland

e-mail: {jerzy.korczak, marcin.hernes, maciej.bac}ue.wroc.pl

***Abstract*—The article presents the performance analysis issues of buy-sell decisions agents' in a-Trader system. The system allows for supporting of investment decision on FOREX market. The first part of article contains a description of a a-Trader system. Next, the algorithms of the selected buy-sell decision agents is presented. In the last part of article the evaluation function of agents' performance is detailed, and the approach to performance analysis is proposed and illustrated.**

I. Introduction

SUPPORTING financial decision making process is performed with the use of methods based on mathematics, statistics, economy or artificial intelligence [2, 4, 5, 10, 12, 15, 17, 19]. The methods are often implemented as the algorithms of the functioning of software agents in multi-agent systems [22]. The paper [1] presents using a multi-agent system in the FOREX market (Foreign Exchange Market). This is one of the biggest financial foreign exchange markets in the world. Currencies are traded against one another in pairs, for instance EUR/USD, USD/PLN. Also a-Trader is the example of system, which enables to support taking investment decisions on the FOREX [13, 14]. This system used tick data, on the basis of which minute aggregates (M1, M5, M15, M30), hourly aggregates (H1, H4), daily aggregates (D1), weekly aggregates (W1) and monthly aggregates (MN1) are created.

Agents functioning in the system take buy-sell decisions with the use of diversified support methods. There arises the need of constant evaluation of the performance of the agents for the purpose of determining the agents giving advice, in the current market situation, regarding the best decisions. As a consequence, the agents' decisions which are given the highest evaluation may constitute the basis for the performance of the buy-sell transaction by the investor. Return on investment cannot be assumed as the only evaluation criterion because other aspects having influence on the effectiveness of the buy-sell decision taken, such as for instance investment risk [9] or transaction costs should also be taken into consideration.

The purpose of this article is to perform the analysis of the performance of selected agents functioning in the a-Trader system with the use of various measures and to elaborate the method of its measurement (evaluation).

In the first part of the article, the a-Trader system is shortly characterized. The algorithms of three selected agents are then presented. In the final part of the article the results of the performance evaluation of these agents is described.

II. a-Trader multi-agent system

The a-Trade platform is of the nature of a multi-agent solution supporting the analysis of high frequency time series, such as e.g. listing of currency pairs on the FOREX market. The basic features of this system include openness, enabling the integration and development of new functionalities of the system and ensuring appropriate communication between particular agents. The system operates in the real time, processing data from the currency market maker live, provided with the use of the MetaTrader or JForex software. After the processing of data provided, the a-Trader system returns the information on making the transaction of the change of the item parameters to the broker (stop loss, take profit). Detailed description of the architecture and its elements as well as information about the agents operating within the framework of the platform may be found in the previous works describing the a-Trader system [13, 14]. This study describes a sample information flow within the platform. The solution diagram regarding the problem of the transfer of information generated by over one thousand agents with the frequency of up to 100 signals per minute will be presented.

A sample route of the signal delivered by the broker (Data Provider 1) is shown in Fig 1. The information processed goes through all system components:

a. Notification Agent (NA),
b. Historical Data Agent (HDA),

978-83-60810-58-3/$25.00

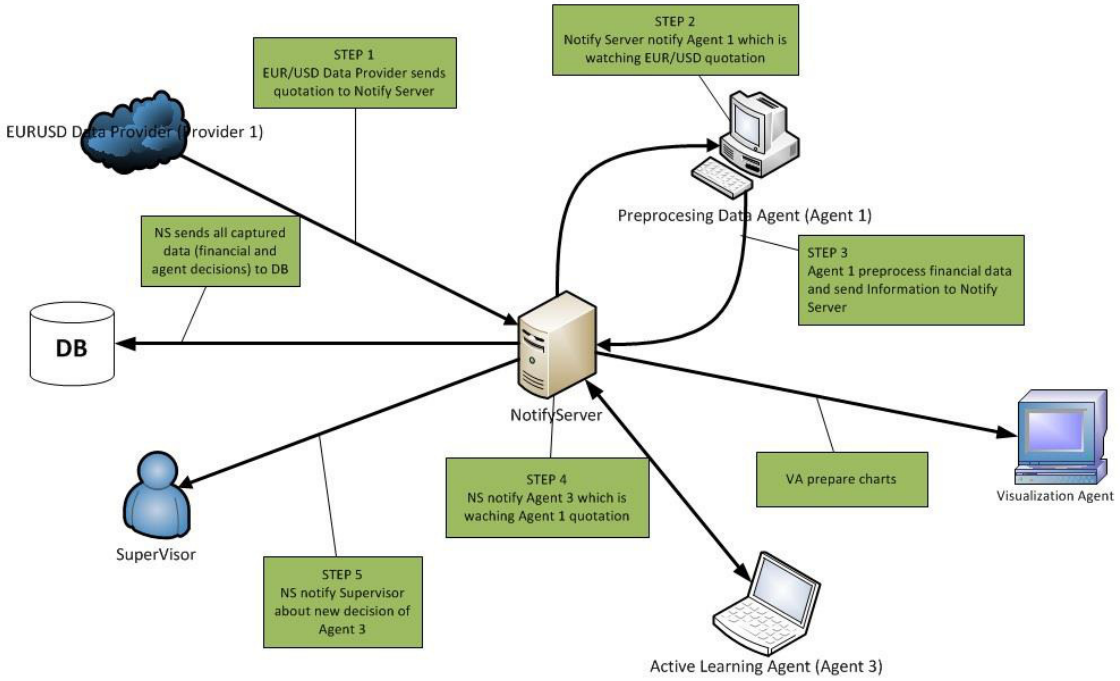


Fig. 1 Signal flow in the A-Trader system.
Source: Own work.

c. Cloud of Computing Agents (CCA),
d. Market Communication Agent (MCA),
e. User Communication Agent (UCA),
f. Supervisor (S),
g. System Database (SD),

Information about the change of quotation value goes directly to the Notification Agent (NA) – STEP 1. This agent decides of the further information flow sending it to all agents that listen to a given signal. Independently, the NA sends the signal to the historic data base. In the analysed case the signal is sent to the Data Processing Agent – STEP 2. The Data Processing Agent checks whether the signal is correct and may be subject to analysis by further agents. It sends the verified signal back to the Notification Agent (NA) – STEP 3. NA notifies Agent 3 (Active Learning Agent) about the signal received from Agent 1 – STEP 4. The Active Learning Agent processes information and sends it to the Supervisor through NA – STEP 5. The Supervisor Agent, on the basis of the decision of Active Learning Agent and the decisions of other agents, takes the final decision concerning the transaction. It sends it back to NA which saves it in the data base and sends to the market through the Data Provider Agent 1.

The Notification Agent (NA) ensures efficient communication inside the system. It is the intermediary agent in sending signals between agents as per declared indications (see Fig. 1). Each agent, the status of which changes, notifies its notification agent. The notification agents forwards the information about the change of the status of a given agent to all agents which are recorded in the notification register as clients/observers of its signals. The notification takes place by calling an appropriate web method (SOAP) at all agents from the list of the ones listening to the indicated signals. Then it records the information about the change of the status of the notification agent in the data base. The functionality of the notification agent elaborated this way makes the system flexible and scalable, gives the possibility of simple adding and removing agents and ensures making the system independent of the agent's location.

In order to be able to efficiently manage communication, the Notification Agent operates in a multi-threaded way. Information concerning the message flow, so which agent awaits signals from which agent, is read during the Notification Agent initialization to the Routing Table. Paralelly threads sending information to particular agents are created (Sending Threads Table). The sending threads are responsible for sending information to listening agents. After the creation they check whether the agent to which they are supposed to deliver the information is active and whether it is listening at a given addresses and is ready for processing data. Fig 2 shows the exemplary data flow inside the Notification Agent. The NA listens at the indicated port XXXX. After receiving information from Agent 4 it finds in the Routing Table which agents listen to the signals of Agent

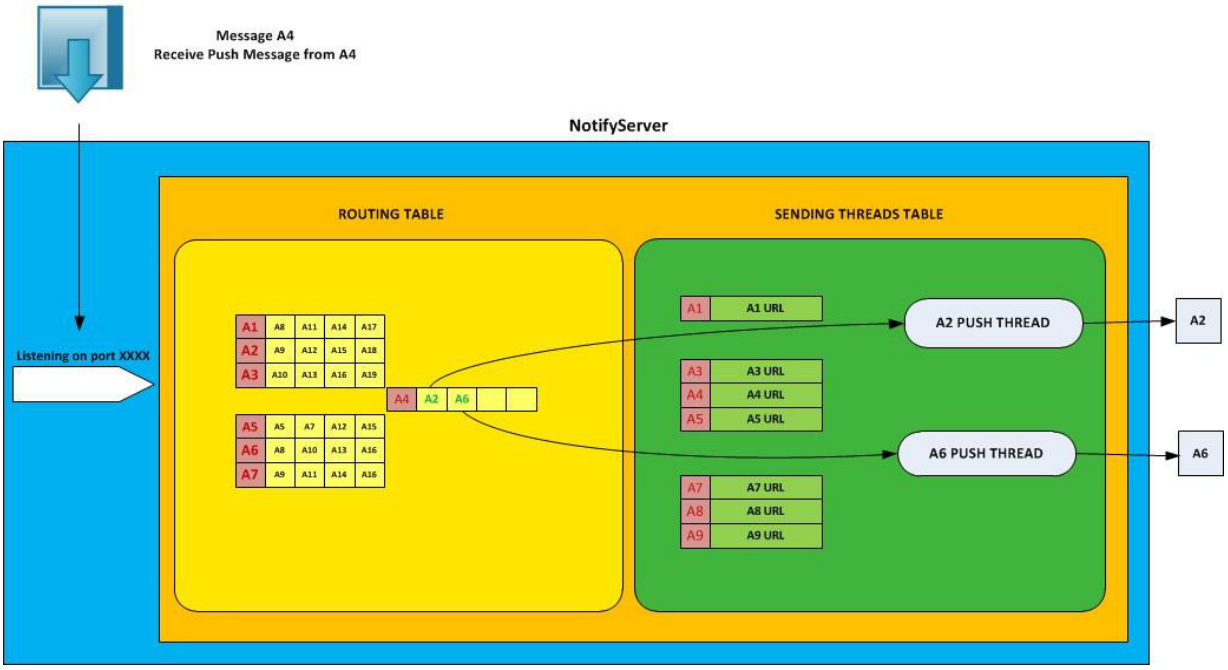


Fig. 2 Data flow inside the Notification Agent.
Source: Own work.

4. In the analysed case these are Agent 2 and 6. Then, the Notification Agent finds the threads sending to Agents 2 and 6 and it forwards the information received from Agent 4 to them, sending threads forward the received information from agents with which they communicate. In case when a problem with communication with agents occurs, the sending thread goes into inactive mode not to load the system. The information about the agent's inactivity is recorded in the event log. The inactive sending thread checks if the communication with the agent to which the information is sent can be established. In the case if the inactive agent starts working again, the communication is resumed.

The presented solution enables complete scalability of the platform. In case of excessive loading of the Notification Agent, another instance is activated. The Agents are notified about being assigned a new Notification Agent and they send their signals to it. One NA instance may receive signals from the group of agents, and the received signal must be sent to all agents awaiting given information.

The information received by the Notification Agent may be divided into three groups. The information division is presented in fig 3. The first one is standard information containing the signal generated by any agent working within the framework of the a-Trader system. The second group of information is the control commands. The third group is the warnings and errors sent by agents. Information flow from the first group was already described. Control information is used for managing the Notification Agent. With the use of this information the agents may demand sending information from other agents or demand that selected agents cease sending information. With the use of the control information an agent may resume its activity, the Notification Agent will then start sending signals to it. In case when an agent is being switched off, it should send information to the NA that it wants to stop listening. The sending thread will go into an inactive mode. The third group of information is used for forwarding the information about errors and warnings of particular agents to the central data base. It is sent to NA in the event of the occurrence of exceptional situations in the agent's operation. They include, e.g. too heavy load of an agent, receiving information which it should not receive and other exceptional situations.

The presented flow of the signal inside the a-Trader system allows for better understanding of processes occurring in the system. The agents described in the next part of the article operate in accordance with the presented convention. They accept other agents' signals generated in real time, process them and take financial decisions. The presented technological grounds of the system operation guarantee its scalability. The control messages, messages about errors and warnings increase the system reliability. This enables to process, almost in real time, thousands of signals generated by agents.

From the point of view of the user (investor) the most important components of the a-Trader system are the agents setting the buy-sell decisions (belonging to the cloud of computing agents) and the Supervisor agent. The functioning of these agents enables the investor to make transactions at

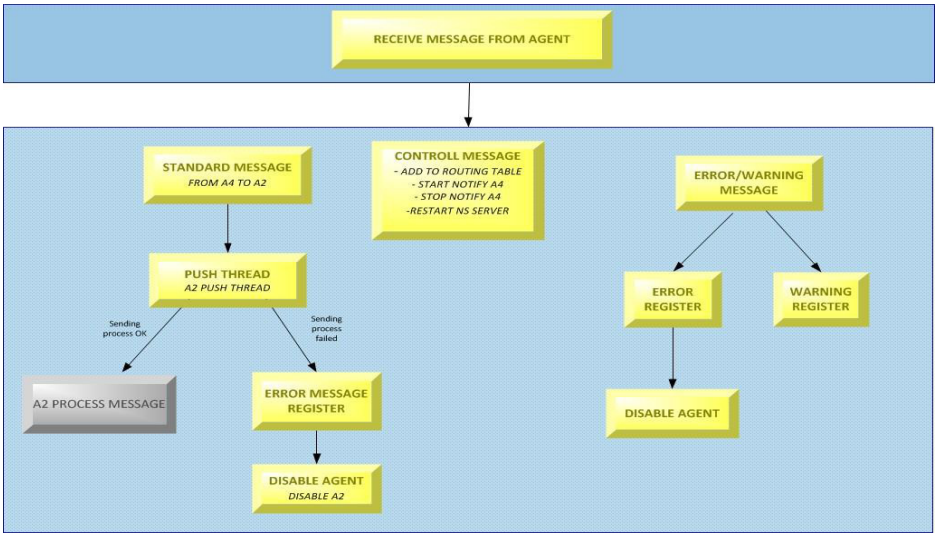


Fig. 3 Types of information received by the Notification Agent.
Source: Own work.

the FOREX market, in accordance with signals generated by them. The characteristics of selected agents setting purchase-sale transactions will be presented.

III. DESCRIPTION OF THE BUY-SELL DECISION AGENTS

Approximately 1000 agents function in the a-Trader system, including approx. 800 agents processing data concerning quotations at the FOREX market (for instance they calculate trend indicators, oscillators) and 200 agents (functioning in all time periods) setting the buy-sell decision. For the needs of this article, in order to perform the efficiency analysis, three agents were selected: TrendLinearReg, MultiTrendSignal and Consensus. These are agents taking decisions on the basis of more complex algorithms than the algorithms of typical technical analysis indicators. The specific nature of the functioning of selected agents will be presented in the further part of the article.

A. . The TrendLinearReg agent

The agent functions on the basis of the assumption that the trend of a certain number M of quotations is approximated with the straight line with the equation: y = ax + b. The straight line inclination depends on the value of the "a" parameter or the tangent value of the inclination angle with the use of linear regression [11, 21]. The agent generates the purchase signal when the coefficient value changes from positive to negative and the sales signal is generated when the coefficient changes value from negative to positive. The change of the agent's decision is made with the use of hysteresis, the level of which is defined by means of the coefficient Δ, the value of which should be higher than transaction costs.

The TrendLinearReg agent functioning algorithm is as follows:

Data: The vector of quotation value of the currency pair $w=<w_1, w_2, ... w_M>$ consisting of M quotations and the previous value the a coefficient marked as *preva.*

Result: The D decision (value 1 denotes „buy" decision, value -1 denotes „sell" decision, value 0 denote s „ leave unchanged") with respect to w and *preva* value.

BEGIN

1: Let i:=1;*sumy*:=0; *sumx*:=0.0; *sumxy:*=0; *sumx2*=0. *II* where: *sumy* means the sum of the value of M quotations, *sumx* means the sum of the quotation number in the vector, *sumxy* means the sum of the products of the quotation value and quotation number in the vector, and *sumx2* means the sum of the squares of quotation numbers in the vector.

2: $sumy:= sumy+w_i$; $sumxy:= sumxy+w_i*i$; $sumx=: sumx+i$; $sumx2:= sumx2+i*i$; $i:=i+1$;

3: If $i<M$ then go to 2. If $i \geq N$ then go to: 4.

4: $c:= sumx2*M-sumx*sumx$. If c=0 then c:=0,1.

5: $a:=(sumxy*M-sumx*sumy)/c$;

6: If $(a(+/-\Delta)=preva=0)$ or $((a+\Delta)<0$ and $preva<0)$ or $((a-\Delta)>0$ and $preva>0)$ then $D:=0$;
If $((a+\Delta)<0$ and $preva>0)$ then $D:=1$; If $((a-\Delta)>0$ and $preva<0)$ then $D:=-1$;

7: $preva:=a$;

END

The complexity of the algorithm, significant due to minimizing of the agent reaction time, amounts to O(M), where M means the number of quotations.

B. . The MultiTrendSignal agent

Agent MultiTrendSignal generates the purchase-sales decision on the basis of the decision of agents functioning with the use of most often used technical analysis ratios [3,

16]. The decisions of the following nine base agents are analysed:

- Average Directional Index (ADX),
- Relative Strength Index (RSI),
- Rate of Change (ROC),
- Commodity Channel Index (CCI),
- Moving Average of Oscillator (OsMA),
- Moving Average Convergence Divergence (MACD),
- Stop and Reverse (SAR),
- Williams %R,
- Moving Average (MA).

The agent considers four time periods (M1, M5, M15, M30) in such a way that the buy/sell decision in case of the period M30 is taken when the same decision is taken by most base agents also within the periods M1, M5, M15, M30.

The structure of the investment decision was defined in the study [7]. This decision is taken on the basis of financial instrument quotation such as, e.g. currency pairs EUR/USD, USD/GBP and it is defined as follows:

Definition 1.

Decision D about finite set of financial instruments $E=\{e_1,e_2,\ldots,e_N\}$ is defined as a set

$$D=\left\langle\{EW^+\},\{EW^\pm\},\{EW^-\},Z,SP,DT\right\rangle, \quad (1)$$

where:

1) $EW^+=\left\langle e_o,pe_o\right\rangle,\left\langle e_q,pe_q\right\rangle,\ldots,\left\langle e_p,pe_p\right\rangle$ - a positive set; in other words, it is a set of financial instruments about which the agent knows the decisions to buy, and the volume of this buying.

2) $EW^\pm=\left\langle e_r,pe_r\right\rangle,\left\langle e_s,pe_s\right\rangle,\ldots,\left\langle e_t,pe_t\right\rangle$ - a neutral set, in other words, it is a set of financial instruments, about which the agent does not know that buy or sell. If these instruments are held by an investor, that they should not be sold, or if they are not in possession of the investor, should not be bought by them.

3) $EW^-=\left\langle e_u,pe_u\right\rangle,\left\langle e_v,pe_v\right\rangle,\ldots,\left\langle e_w,pe_w\right\rangle$ - a negative set; in other words it is a set of financial instruments of which the agent knows that these elements should sell. Couple $\left\langle e_x,pe_x\right\rangle$, where: $e_x\in E$ and $pe_x\in[0,1]$, denote financial instrument and this instrument's participation in set EW^+, $EW^\pm$, EW^-.

Financial instrument $e_x\in EW^+$ is denoted as: e_x^+.

Financial instrument $e_x\in EW^\pm$ is denoted as: $e_x^\pm$.

Financial instrument $e_x\in EW^-$ is denoted as: e_x^-.

4) $Z\in[0,1]$ - predicted rate of return.

5) $SP\in[0,1]$ - degree of certainty of rate Z. It can be calculated on the basis of the level of risk related with the decision.

6) DT - date of decision.

The set of agents' decisions, on the basis of which the MultiTrendSignal agent sets decisions is called a profile. .

The agent functioning algorithm is as follows;

Data: Profiles $AM1=\{AM1^{(1)},AM1^{(2)},\ldots.AM1^{(9)}\}$
$AM5=\{AM5^{(1)},AM5^{(2)},\ldots.AM5^{(9)}\}$
$AM15=\{AM15^{(1)},AM15^{(2)},\ldots.AM15^{(9)}\}$
$AM30=\{AM30^{(1)},AM30^{(2)},\ldots.AM30^{(9)}\}$
consist of *9* agents' decisions.

Result: Decision $DEC=\left\langle DEC_+,DEC_\pm,DEC_-,DEC_Z,DEC_{SP},DEC_{DT}\right\rangle$ according the profiles.

BEGIN

1: Let $DEC_+=DEC_\pm=DEC_-=\varnothing,DEC_Z=DEC_{SP}=DEC_{DT}=0$.

2: *j*:=1.

3: *i*:=+.

4: If ($tM1_i(j)>4$) and ($tM5_i(j)>4$) *and* ($tM15_i(j)>4$) *and* ($tM30_i(j)>4$) then $DEC_i:=DEC_i\cup\{e_j\}$. Go to:6. // $tMxX_i(j)$ – the number of occurrences of the financial instrument in a positive, neutral or negative set in a given time period.

5: If *i*= + then *i*:=±. Ifi *i*=± then *i*:=- . if *i*=- then go to: 6.
Go to: 4

6: If *j*<*N* then *j*:=*j*+1 go to:3.
If *j*≥*N* then go to: 7.

7: *i*:=*Z*.

8: Determine *pr*(*i*). //ascending order

9: $k_i^1=(9+1)/2$, $k_i^2=(9+2)/2$.

10: $k_i^1\le DEC_i\le k_i^2$

11: If *i*=*Z* then *i*:=*SP*. If *i*:=*SP* then *i*:=*DT*. Ifi *i*:=*DT* then END.
Go to: 8.

END.

The algorithm complexity amounts to O(18N), where N means the number of currency pairs.

C. The Consensus agent

In the a-Trader system the agents take buy-sell decisions independently of one another. Thus a conflict situation may occur, in which (at a given moment) these decisions are mutually contradictory (for instance some agents suggest the purchase decision and other agents – the sale decision). In order to solve this conflict a few strategies were implemented into the system: the strategy of dominating decisions, the strategy based on moving average, the consensus strategy, the evolution strategy. The Consensus method [8, 18, 20] is described in the article.

The Consensus agent (characterized in detail in the work of [14] determines the decisions on the basis of the set of decisions generated by other agents functioning in the system.

The agent functioning algorithm is as follows:

Data: The profile $A=\{A^{(1)},A^{(2)},\ldots.A^{(M)}\}$ consist of *M* agents' decisions. //$A^{(1)},A^{(2)},\ldots.A^{(M)}$ – decisions of particular agents

Result: Consensus
$CON=\left\langle CON_+,CON_\pm,CON_-,CON_Z,CON_{SP},CON_{DT}\right\rangle$ according *A*.

BEGIN

1: Let $CON_{+}=CON_{\pm}=CON_{-}=\emptyset, CON_{Z}=CON_{SP}=CON_{DT}=0$.

2: j:=1.

3: i:=+.

4: If $t_i(j) > M$ then $CON_i := CON_i \cup \{e_j\}$. Go to:6.

// $t_i(j)$ – the number of occurrences of the financial instrument in the positive, neutral or negative set

5: If i=+ then i:=±
If i=± then i:=-
If i=-, then Go to:6
Go to:4.

6: If $j<N$ then j:=j+1 Go to:3
If $j \geq N$ then Go to:7.

7: i:=Z.

8: Determine $pr(i)$. // ascending order

9: $k_i^1 = (M+1)/2$, $k_i^2 = (M+2)/2$.

10: $k_i^1 \leq CON_i \leq k_i^2$.

11: If i=Z then i:=SP.
If i:=SP then i:=DT.
If i:=DT then END.
Go to: 8.

END.

The complexity of the algorithm amounts to O(3NM), where N means the number of currency pairs and M means the number of agents belonging to the profile (in the research experiment conducted in the next part of the article, M=25, N=1).

Computational complexities of the algorithms of the agents' functioning have impact on the performance of the whole a-Trader system. Taking into consideration the fact that the system is processing tick signals and a large number of agents function in it (approx. 1000), short time of computation made by particular agents is very significant.

In general, agents functioning in the system for the purpose of determining the decisions use the methods of technical analysis, fundamental analysis, neural networks, evolution algorithms, behavioural models.

The Supervisor agent also functions in the system and its major purpose is to maximize the rate of return and reduce the investment risk. The Supervisor's task is to coordinate the functioning of agents setting the buy-sell decisions and presenting the final decision to the investor. This agent uses various strategies, analyses them and evaluates the agents' performance.

A case study relating to the method of the measurement of the performance of selected agents taking buy-sell decisions is presented further in the article.

IV. The Agent Performance Evaluation Method – Case Study

The agents performance analysis is performed for data within the M5 range of quotations from the FOREX market. For the purpose of this analysis, a test was performed in which the following assumptions were made:

1. EUR/USD pairs, out of four randomly selected pairs of the following periods, were used:
   - 31-03-2014, 0:00 am to 31-03-2014, 23:59 pm,
   - 03-04-2014, 0:00 am to 03-03-2014, 23:59 pm,
   - 04-04-2014, 0:00 am to 04-04-2014, 23:59 pm,
   - 05-04-2014, 0:00 am to 05-04-2014, 17:00 pm,
2. At the verification the decisions (signals buy-value 1, sell-value -1, leave unchanged - value 0) generated by the agents TrendLinearReg, MultiTrendSignal are used (the example is presented in the Fig 4, where the green arrow means the decision "buy", the red one - "sell"), and Consensus.
3. It was assumed that the initial capital held by the investor amounts to USD 1000 and the difference between this amount and the amount which the investor will have after the last sales transaction in a given period is considered the rate of return. The rate of return is expressed in nominal units (USD).
4. The transaction costs are directly proportional to the number of transactions.
5. The capital management - it was assumed that in each

Fig. 4 The MultiTrendSignal agent decisions.
Source: Own work.

transaction the investor engages 100% of the capital held. The capital management strategy may be determined by the user. The investor every time invests 1000 at the leverage 10:1 and invests all the capital held.

6. The performance analysis was performed with the use of the following measures (ratios):
   - rate of return (ratio x_1),
   - the number of transaction,
   - gross profit (ratio x_2),
   - gross loss (ratio x_3),
   - total profit (ratio x_4),
   - the number of profitable transactions (ratio x_5),
   - the number of profitable transactions in a row (ratio x_6),
   - the number of unprofitable transactions in a row (ratio x_7),
   - Sharpe ratio (ratio x_8)

$$S = \frac{E(r) - E(f)}{|O(r)|} \cdot 100\% \qquad (2)$$

     where:
     E(r) – arithmetic average of the rate of return,
     E(f) – arithmetic average of the risk-free rate of return,
     O(r) – standard deviation of rates of return.
   - the average coefficient of variation (ratio x_9)is the ratio of the average deviation of the arithmetic average multiplied by 100% and is expressed:

$$V = \frac{s}{|E(r)|} \cdot 100\% . \qquad (3)$$

     where:
     V – average coefficient of variation,
     s – average deviation of the rates of return,
     E(r) – arithmetic average of the rates of return.
   - Value at Risk (ratio x_{10}) – the measure known as value exposed to the risk - that is the maximum loss of the market value of the financial instrument possible to bear in a specific timeframe and at a given confidence level [3].

$$VaR=P*O*k \qquad (4)$$

     where:
     P – the initial capital,
     O – volatility - standard deviation of rates of return during the period ,
     k –the inverse of the standard normal cumulative distribution (assumed confidence level 95%, the value of k is 1,65),
   - the average rate of return per transaction (ratio x_{11}), counted as the quotient of the rate of return and the number of transactions.

7. For the purpose of the comparison of the agents' performance, the following evaluation function was elaborated:

$$y = (a_1 x_1 + a_2 x_2 + a_3 (1 - x_3) + a_4 x_4 + a_5 x_5 + a_6 x_6 + \ldots$$
$$+ a_7 (1 - x_7) + a_8 x_8 + a_9 (1 - x_9) + a_{10} (1 - x_{10}) + a_{11} x_{11}) \qquad (5)$$

where x_i denote the normalized values of ratios mentioned in item 6 from x_1 to x_{11}. It was adopted in the test that coefficients a_1 to a_{11}=1/11.

It should be mentioned that these coefficients may be modified with the use of, for instance, an evolution method or determined by the user (investor) in accordance with his/her preference (for instance the user may determine whether he/she is interested in the higher rate of return with simultaneous higher risk level or lower risk level but simultaneously agrees to a lower rate of return).

The function is given the values from the range [0..1], and the agent's efficiency is directly proportional to the function value.

8. The results obtained by the tested agents were compared with the results of the Buy-and-Hold strategy and the strategies using EMA..

The agent efficiency tests were performed in the following manner:

1. On the basis of data from the first period each agent defined when to buy and when to sell EUR/USD currency.
2. In the next step, on the basis of the results of particular agents and Buy-and-Hold and EMA, for each purchase-sale operation the value of capital held and the rate of return in USD were determined.
3. At the final stage the value of performance ratios was calculated with respect to the rates of return resulting from all decisions generated by the analysed agents and the Buy-and-Hold as well as EMA strategies (not only from the final rates of return but from all rates of return calculated after each sales decision). The evaluation functions were also calculated.
4. Then the steps from 1 to 3 were repeated with the use of the data from the successive periods.

Table 1 presents the results obtained in the particular periods.

Generalizing the agent efficiency analysis results, it may be noticed that in the periods in question their decisions generated both profit and loss. Thus, in the efficiency analysis not only the rate of return should be taken into consideration but also other ratios, including the level of risk involved in the investment, which is enabled by the evaluation function elaborated in the article.

In fig 5 the diagram of the value of ratios and evaluation function of particular agents (and the B&H method as well as EMA) in the periods in question are presented. To illustrate relationships between ratios and agents the parallel

TABLE I.
PERFORMANCE ANALYSIS RESULTS

Ratio	TrendLinearReg				MultiTrendSignal				Consensus				B&H				EMA			
	Period 1	Period 2	Period 3	Period 4	Period 1	Period 2	Period 3	Period 4	Period 1	Period 2	Period 3	Period 4	Period 1	Period 2	Period 3	Period 4	Period 1	Period 2	Period 3	Period 4
Rate of return [USD]	16,07	15,09	25,77	14,83	18,51	14,98	24,49	9,83	19,80	14,45	26,34	14,89	17,97	13,94	-19,65	-38,79	15,60	-15,82	2,39	-8,77
The number of transactions	9	10	12	12	4	5	9	10	5	7	9	12	1	1	1	1	18	19	17	16
Gross profit [USD]	13,60	8,63	9,21	4,56	10,27	7,66	5,28	6,56	6,89	5,91	9,22	3,91	17,97	0,00	0,00	0,00	26,36	10,74	20,03	4,75
Gross loss [USD]	-4,65	-3,27	-1,82	-2,65	1,32	-3,47	-3,02	-4,65	1,60	-1,69	-1,67	-1,51	0,00	13,94	-19,65	-38,79	-4,00	-13,50	-19,02	-5,45
Total profit [USD]	30,90	24,16	28,53	19,63	18,51	18,45	27,51	14,63	19,80	16,14	28,01	16,4	17,97	0,00	0,00	0,00	32,07	27,42	40,50	7,57
The number of profitable transactions	4	5	9	10	4	4	8	8	5	6	8	11	1	0	0	0	7	4	4	3
The number of profitable transactions in a row	3	4	3	5	4	4	5	5	5	5	7	6	1	1	0	0	2	2	2	2
The number of unprofitable transactions in a row	2	4	1	1	0	1	1	1	0	1	1	1	0	0	1	1	5	5	5	6
Sharpe ratio	0,27	0,35	0,74	0,58	1,18	0,72	1,04	0,35	1,72	0,85	0,96	0,95	0,00	0,00	0,00	0,00	0,13	-0,15	0,01	-0,25
The average coefficient of variation [%]	3,01	2,30	0,89	1,24	0,61	0,99	0,68	1,62	0,49	0,82	0,67	0,71	0,00	0,00	0,00	0,00	3,43	4,15	33,42	2,59
Value at Risk [USD]	107,50	69,55	47,69	34,59	64,34	68,15	42,92	45,06	37,87	39,61	50,09	21,24	0,00	0,00	0,00	0,00	107,89	92,89	144,90	36,87
The average rate of return per transaction	1,79	1,51	2,15	1,24	4,63	3,00	2,72	0,98	3,96	2,06	2,93	1,24	17,97	13,94	-19,65	-38,79	0,87	-0,83	0,14	-0,55
Value of evaluation function (y)	0,45	0,43	0,53	0,50	0,55	0,50	0,54	0,48	0,59	0,52	0,55	0,46	0,54	0,41	0,42	0,40	0,44	0,34	0,30	0,34

Source: Own work.

coordinates are used, which is a common way of visualizing and analyzing multivariate data. An agent in n-dimensional space is represented as a polyline with vertices on the parallel axes; the position of the vertex on the ith ratio corresponds to the ith coordinate of the agent.

It may be noticed that the values of efficiency ratios of particular agents differ in each period and get the values (after normalization) in the range from 0 to 1. The values of such ratios as x_2,x_3 and x_{11} are approximate in case of all agents and the values of ratios x_5, x_6, x_8 are characterized by significant distribution in case of particular agents. It may also be noticed that in case of the agents TrendLinearReg, MultitrendSignal and Consensus the values of ratios x_1, x_2, x_3, x_9 and x_{11} are similar in each of the periods in question and the values of ratios x_4, x_5, x_6, x_7, x_8, x_{10} are characterized by much variability in particular periods. A large scope of changes of ratios significantly hinders the analysis by the user and, as a consequence, prevents taking decisions in time close to real time. And the application of the evaluation function allows for immediate appointment of the agent with the best efficiency. It may be noticed that the evaluation function values oscillate in the range from 0.03 – 0.59 thus despite large deviations in the values of particular ratios, the agents are evaluated in the range characterized by a smaller value deviation. The results of the experiment performed allow to state that the ranking of agents' evaluation differs in particular periods. In the first period the Consensus agent turned out to be the best agent and the MultiTrendSignal agent was ranked higher and the TrendLinearReg agent was ranked lower than the B&H benchmark evaluation. The EMA benchmark was ranked the lowest in this period. In the second period the MultiTrendSignal and TrendLinearReg agents and the Consensus agent were ranked higher than the EMA and B&H benchmarks. Considering the third period it may be noticed that the evaluation ranking is similar to the one in the second period. And in the fourth period the MultiTrendSignal was ranked the highest and the TrendLinearReg and Consensus Agents were ranked higher than the B&H benchmarks. The EMA benchmark was ranked the lowest in this period.

Taking into consideration all the periods in question it may be stated that the Consensus agent was ranked highest most often (3 out of 4 periods) although the rate of return of this agent was not always the highest. This evaluation results, however, from the low level of risk connected with investing on the basis of the Consensus agent decision. And, on the other hand, the TrendLinearReg agent was ranked low most often (3 out of 4 periods) because at a relatively high risk level it generated little rate of return. Rates of return obtained by the MultiTrendSignal lower than rates of return of the other two agents may result from the fact that this agent is characterized by a low number of transactions because it takes decisions with the use of a few quotation time slots. It may also be noticed that low evaluation of the EMA benchmark in all periods results not from the level of the rate of return but from a high risk level and a large number of loss transactions in a row.

Referring to the evaluation analysis performed in other systems (e.g. in the MetaTrader system) it should be emphasized that it is, in most cases, performed "manually" by the investor. Due to its time consumption, its utility in the systems operating in real time is very limited. Besides, these systems only offer the functions calculating the basic ratios (rate of return, number of transactions, highest profit, highest loss, total profit, number of profitable transactions, number of profitable transactions in a row, number of loss transactions in a row), and in the a-Trader system also additional ratios are calculated, such as the risk measures (Sharp ratio, average value coefficient, risk exposed value), or the average rate of return from transaction.

The evaluation function, elaborated in this article, enables the measurement and the performance evaluation of particular agents taking the buy-sell decision in the system. These operations are made automatically, in time close to real time, by the Supervisor agent which may then suggest the investor taking final decisions on the basis of decisions generated by the agent with the highest level of performance. In addition, enabling the user to change a_i and x_i parameters

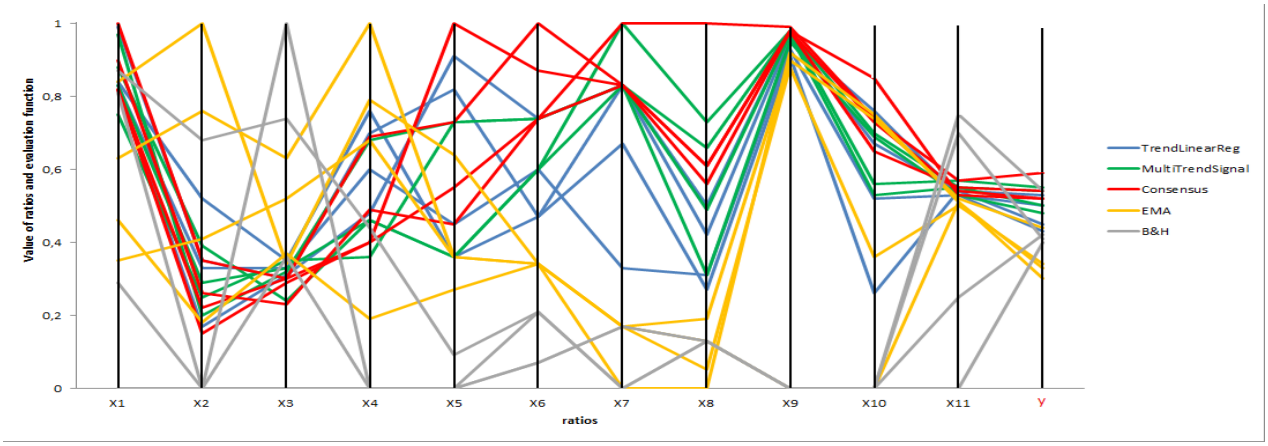


Fig. 5 Values of ratios and evaluation function of particular agents in the periods in question.
Source: Own work.

of the evaluation function allows for considering his/her preference concerning the criterion of importance of particular evaluation ratios. The evaluation value also considers the transaction costs with the assumption that the dependency between the number of transactions and the average rate of return from the transaction is reflected. However this simple principle cannot be adopted because a large number of transactions has impact on the reduction of the agent's efficiency level, especially for the transactions with a high rate of return.

The elaborated evaluation function may be extended with other ratios which do not have directly or (reversely) proportional impact on the function value. For example, this may be the correlation between the rate of return and the ratios defining the risk.

V. CONCLUSION

The agents in the a-Trader system take independent buy-sell decisions using various methods for this purpose. The functioning of these agents involves, however, the need to perform constant analysis of their performance, which should be performed by the Supervisor agent. As a consequence, this enables the investor to present decisions generated by the best agents. The analysis results presented in this article allow to draw conclusions that, depending on the current situation on the FOREX market, the level of performance of particular agent changes. There is no agent which definitely dominates over the other ones. And the use of this performance evaluation function allows for automatic setting of the best agent in time close to real time, which has, in turn, a positive influence on investment effectiveness.

Agents based on artificial intelligence methods also function in the a-Trader system. Neural networks recognize the models or sequences of changes of agent signals and on this basis they take a decision. Evolution algorithms are developed, which are able to calculate most effective combinations of agents over a few hundred seconds. Owing to this they adjust to the variable situations dynamically. intelligent methods will be described in the successive articles,

Currently works are being performed on the implementation of the "directional change algorithm" [6], the evolution method of determining a_i coefficients into the a-Trader system and the implementation of cognitive agents, performing the fundamental analysis and analysing experts' opinions in the scope of forecasts referring to quotations on the FOREX market.

REFERENCES

[1] M. Aloud, E.P.K. Tsang, R. Olsen, "Modelling the FX Market Traders' Behaviour: An Agent-based Approach", [in] Alexandrova-Kabadjova B., S. Martinez-Jaramillo, A. L. Garcia-Almanza & E. Tsang (eds.), Simulation in Computational Finance and Economics: Tools and Emerging Applications, IGI Global, 2012, pp. 202-228.

[2] R.P. Barbosa, O. Belo, "Multi-Agent Forex Trading System", [in] Agent and Multi-agent Technology for Internet and Enterprise Systems, Studies in Computational Intelligence Volume 289, 2010, pp. 91-118.

[3] J. Bollinger, "Bollinger on Bollinger Bands", McGraw Hill, 2001.

[4] L. Chan, A. Wk Wong, "Automated Trading with Genetic-Algorithm Neural-Network Risk Cybernet-ics: An Application on FX Markets", Finamatrix, January 2011, pp.1-28.

[5] M. Dempster, C. Jones, "A Real Time Adaptive Trading System using Genetic Programming", Quantitative Finance, 1, 2001, pp. 397-413. DOI: doi:10.1088/1469-7688/1/4/301.

[6] J. B. Glattfelder, A. Dupuis, R. Olsen, "Patterns in High-Frequency FX Data: Discovery of 12 Empirical Scaling Laws", Quantitative Finance, Volume 11 (4), 2011, pp. 599-614,

[7] M. Hernes, Metody consensusu w rozwiązywaniu konfliktów wiedzy w wieloagentowym systemie wspomagania decyzji, Praca dokt., Uniwersytet Ekonomiczny we Wrocławiu, 2011.

[8] M. Hernes M., N.T. Nguyen, "Deriving Consensus for Hierarchical Incomplete Ordered Partitions and Coverings", Journal of Universal Computer Science 13(2) /2007, pp. 317-328.

[9] K. Jajuga, T. Jajuga, Inwestycje: Instrumenty finansowe, ryzyko finansowe, inżynieria finansowa, PWN, Warszawa 2000.

[10] R. Karjalainen, "Using Genetic Algorithms to Find Technical Trading Rules", Journ. of Financial Econ., 51, 1999, pp. 245-271.

[11] C. D. Kirkpatric, J. Dahlquist, :Technical Analysis: The Complete Resource for Financial Market Technicians", Financial Times Press, 2006.

[12] J. Korczak, P. Lipinski, „Systemy agentowe we wspomaganiu decyzji na rynku papierów wartościo-wych", [in] Rozwój informatycznych systemów wieloagentowych w środowiskach społeczno-gospodarczych, ed. S. Stanek et al., Placet, 2008, pp. 289-301.

[13] J. Korczak, M. Bac, K. Drelczuk, A. Fafuła, "A-Trader - Consulting Agent Platform for Stock Ex-change Gamblers", [in] Proc. FedCSIS, Wrocław, 2012, pp.963-968.

[14] J. Korczak, M. Hernes, M. Bac, "Risk avoiding strategy in multi-agent trading system", [in] Proceedings of Federated Conference Computer Science and Information Systems (FedCSIS), Kraków, 2013.

[15] B. LeBaron, "Active and Passive Learning in Agent-based Financial Markets", Eastern Economic Journal, vol. 37, 2011, pp. 35-43.

[16] C. Lento, "A Combined Signal Approach to Technical Analysis on the S&P 500". Journal of Business & Economics Research 6 (8), 2008, pp. 41–51.

[17] S. Martinez-Jaramillo, E.P.K. Tsang, "An Heterogeneous, Endogenous and Co-evolutionary GP-based Financial Market", IEEE Transactions on Evolutionary Computation, Vol.13, No.1, 2009, pp.33-55.

[18] N. T. Nguyen, "Using Consensus Methodology in Processing Inconsistency of Knowledge", [in] Last M. et al. (Eds): Advances in Web Intelligence and Data Mining, series Studies in Computational Intelligence, Springer-Verlag, 2006, pp. 161-170.

[19] I. Palit, S. Phelps, W. L. Ng, "Can a Zero-Intelligence Plus Model Explain the Stylized Facts of Financial Time Series Data?", [in] Proceedings of the Eleventh International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS) 2012 - Volume 2. Valencia, Spain: International Foundation for Autonomous Agents and Multiagent Systems, pp. 653–660.

[20] J. Sobieska-Karpińska, M. Hernes, "Consensus determining algorithm in multiagent decision support system with taking into consideration improving agent's knowledge", [in] Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS), 2012, pp. 1035-1040.

[21] TrendLinearReg, http://forexwikitrading.com/forex-indicator/trendlinearreg/ [access: 2014.02.02].

[22] M. Żytniewski, R. Kowal, A. Sołtysik, "The Outcomes of the Research in Areas of Application and Impact of Software Agents Societies to Organizations so far. Examples of Implementation in Polish Companies", [in] Annals of Computer Science and Information Systems, Proceedings of Federated Conference Computer Science and Information Systems (FedCSIS), Kraków, 2013 pp. 1165 – 1168.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1181–1190
DOI: 10.15439/2014F243
ACSIS, Vol. 2

Critical Success Factors for ERP Projects in Small and Medium-sized Enterprises – The Perspective of Selected German SMEs

Christian Leyh
Technische Universität Dresden
Chair of Information Systems, esp. IS in Manufacturing and Commerce
Helmholtzstr. 10, 01069 Dresden, Germany
Email: Christian.Leyh@tu-dresden.de

***Abstract*—The aim of our study was to provide a contribution to the research field of the critical success factors (CSFs) of ERP projects, with a specific focus on smaller enterprises (SMEs). Therefore, we conducted a systematic literature review in order to update the existing reviews of CSFs. On the basis of that review, we led several interviews within German SMEs that have implemented ERP systems. As a result, we showed that all factors found in the literature also affected the success of ERP projects in SMEs. However, within those projects, technological factors gained much more importance compared to those factors that most influence the success of larger ERP projects. For SMEs, factors such as the Organizational fit of the ERP system as well as ERP system tests are even more important than Top management support or Project management, which were the most important factors for large-scale companies.**

I. Introduction

TODAY'S enterprises are faced with the globalization of markets and rapidly changing economies. In order to cope with these fluid conditions, the use of technology, as well as information and communication systems, is almost mandatory. Specifically, the adoption of enterprise resource planning (ERP) systems as standardized systems that encompass the activities of entire enterprises has become an important factor for today's businesses. The demand for ERP applications has increased for several reasons, including competitive pressure to become low-cost producers, expectations of revenue growth, and the desire to re-engineer businesses to respond to market challenges. A properly selected and implemented ERP system offers several benefits including considerable reductions in inventory costs, raw material costs, lead time for customers, production time, and production [1]-[4]. Therefore, the majority of enterprises around the world now use ERP systems. For example, according to a survey conducted in Germany from 2010 to 2011, ERP systems are used in more than 92 percent of all German industrial enterprises [5].

Due to the saturation of ERP markets targeting large-scale enterprises, ERP system manufacturers today are also concentrating on the growing market of small and medium-sized enterprises (SMEs) [3], [6]. This has resulted in a highly fragmented ERP market and a great diffusion of ERP systems throughout enterprises in nearly every industry and of every size [7]-[9]. Due to the strong demand and the high fragmentation of the market, there are many ERP systems with different technologies and philosophies currently available on the market. This multitude of software manufacturers, vendors, and systems implies that enterprises that use or want to use ERP systems must strive to find the "right" software as well as to be aware of the factors that influence the success of the implementation project.

The implementation of an information system (e.g., an ERP system) is a complex and time-consuming project, and in the process, companies face not only great opportunities but also enormous risks. To take advantage of potential opportunities without being affected by the risks of these implementation projects, it is essential to focus on those factors that support the successful implementation of an information system [10], [11]. Being aware of these factors, a company can positively influence the success of the implementation project and effectively minimize the project's risks [10]. Recalling these so-called critical success factors (CSFs) is of high importance whenever a new system is to be adopted and implemented or a current system needs to be upgraded or replaced. Errors during the selection, implementation, or maintenance of ERP systems, incorrect implementation approaches, and ERP systems that do not fit the requirements of the enterprise can all cause financial disadvantages or disasters, perhaps even leading to insolvencies. Several examples of such negative scenarios can be found in the literature (e.g., [12], [13]). SMEs must be especially aware of CSFs, since they lack the financial, material, and personnel resources of larger companies. Thus, they are under greater pressure to implement and run ERP systems without failure and as smoothly as possible.

These critical success factors have already been considered in numerous scientific publications. Several case studies, surveys, and literature reviews have previously been conducted by different researchers (e.g., [4], [14]-[16]). However, the existing ERP system success-factor research has focused mostly on the selection and implementation of ERP systems in large enterprises. Less attention has been paid to implementation projects in SMEs, despite the fact that research focusing on CSFs in smaller companies has been recommended by the research community for several years (e.g., [17], [18]).

Therefore, the aim of our study was to focus on the implementation of ERP systems in SMEs, focusing in

978-83-60810-58-3/$25.00

particular on the differences in CSFs between larger ERP projects and smaller projects. Prior to this study, we conducted a systematic literature review in order to update the existing reviews of CSFs. On the basis of the CSFs identified, we conducted multiple interviews within German SMEs that have already implemented ERP systems in order to obtain insights into the similarities and differences in CSFs for ERP system implementation in SMEs. Overall, our study was driven by the following research question:

Q1: What similarities and differences exist between critical success factors for ERP implementation projects in larger and smaller enterprises?

Therefore, the paper is structured as follows. The next section gives a short overview of the subsequently discussed and important CSFs before the section that follows deals with the results of our literature review. Here, we will point out which factors are the most important and which factors seem to have little influence on the success of an ERP implementation project. Next, our data collection methodology is described before the results of the interviews are presented and the research question is answered. Finally, the paper concludes with a summary of the results and discusses the limitations of our study.

II. Critical Success Factors Identified

A CSF for ERP projects has been defined by [15] as a reference to any condition or element that was deemed necessary in order for the ERP implementation to be successful. To provide a comprehensive understanding of the different CSFs and their concepts, these are described in this section before presenting the research methodology and discussing the results. However, only the most-important and subsequently discussed factors are described. The detailed definitions of the other CSFs can be found in [7] and [19].

ERP system configuration: Since the initial ERP system version is based on best practices, a configuration or adaptation of the system according to business processes is necessary in every ERP implementation project. Hence, as closely as possible, the company should attempt to adopt the processes and options built into the ERP rather than seeking to modify the ERP [20]. Following [21], the more the original ERP software is modified (e.g., even beyond the "normal" configuration), the smaller the chance is for a successful implementation project. Hence, a clear business vision is helpful because it reduces the effort of capturing the functionality of the ERP business model and therefore minimizes the effort needed for the configuration [20]. Again, extensive system modifications will not only cause implementation problems but also interfere with system maintenance. Therefore, the need for fewer adjustments reduces the effort of integrating new versions, releases, or updates [22].

ERP system tests: In ERP implementation, "go live" on the system without adequate and planned system testing may lead to an organizational disaster. Tests and validation of an ERP system are necessary to ensure that the system works correctly from a technical perspective, and that the business process configurations have been done the right way [23]. Therefore, testing and simulation exercises for both, the whole system and its separate parts and functions, have to be performed during and in the final stages of the implementation process [15], [24].

Organizational fit of the ERP system: The fact that the organizational fit of an ERP system should be examined and considered comprehensively before its implementation sounds logical. Nevertheless, ERP manufacturers often attempt to create blind confidence in their ERP packages, even if the organizational fit is obviously low. In [21], the extent to which the implementation success of an ERP system depends on the fit between the company and the ERP system was empirically examined, and it was found that the adaptation and configuration effort negatively correlate with implementation success. Therefore, the careful selection of an ERP system with consideration of its company-specific organizational fit, such as company size or industry sector, is essential, and appropriate ERP system selection is an important factor in the effort to ensure a good fit between the company and the ERP system.

Project management: Project management refers to the ongoing management of the implementation plan [15]. The implementation of an ERP system is a unique procedure that requires enterprise-wide project management. Therefore, it involves planning stages, the allocation of responsibilities, definitions of milestones and critical paths, training and human resource planning, and the determination of measures of success [25], [26]. This enables timely decisions and guarantees that such decisions are made by the "right" company members. Furthermore, continuous project management allows the focus to remain on the important aspects of ERP implementation and ensures that timelines and schedules are met [25]. Within project management, comprehensive documentation of the tasks, responsibilities, and goals is indispensable for the success of an ERP implementation [17].

Top management support and involvement: Top management support and involvement is one of the most-important success factors for ERP implementation [14]. Committed leadership at the top management level is the basis for the continuous accomplishment of every project [15]. Thus, innovations and, in particular, new technologies are more quickly accepted by employees if these innovations are promoted by top management. Before the project starts, top management has to identify the peculiarities and challenges of the planned ERP implementation. Since many decisions that must be made during the project can affect the whole enterprise, these decisions will require the acceptance and commitment of senior managers and often can only be made by them [27]. The commitment of top management is important for the allocation of necessary resources, for quick and effective decision making, for solutions to conflicts that need enterprise-wide acceptance, and for enlisting the cooperation of all departments [24].

User training: Missing or inadequate end user training is often a reason for failures in the implementation of new software. The main goal of end user training is to provide an effective understanding of the new business processes and applications as well as the new workflows that are created by the ERP implementation. Therefore, establishing a suitable plan for training employees is important [24]. Furthermore, during the implementation of such an extensive project, management must determine which employee is the best fit for which position or for which application of the new software. This depends strongly on the employee's previously acquired knowledge and/or additional training courses [28].

III. Literature Review of Critical Success Factors

In order to identify the factors that affect the success or failure of ERP projects, several case studies, surveys, and literature reviews have already been conducted by a number of researchers (e.g., [4], [15], [16], [20]-[22]).

However, most of the literature reviews cannot be reproduced because descriptions of the review methods and procedures are lacking. Some researchers have pointed out the limitations of literature review articles, noting specifically that they lack methodological rigor [29]. Therefore, in order to update the existing reviews by including current ERP literature, we conducted a literature review by systematically reviewing articles in five different databases as well as papers drawn from several international conference proceedings. More specifically, we conducted two separate literature reviews according to the same search procedure and steps. The first was performed in mid-2010 [7], [19]. Since we identified 20 or more papers published each year, it was essential for us to update this review every two or three years. Therefore, we conducted the second review in mid-2013. In the Appendix (see Tables 6 and 7) we provide an overview of the identified databases and conferences as well of the used search terms. The overall procedure for the literature review will not be part of this paper. It is described in detail in [7], [19], [30].

We identified 320 papers that referred to CSFs of ERP projects. These papers were reviewed again in depth in order to determine the various concepts associated with CSFs. For each paper, the CSFs were captured, along with the publication year, the type of data collection used, and the companies (i.e., the number and size) from which the CSFs were derived.

All 320 papers were published between 1998 and mid-2013. Table 1 shows the distribution of the papers by publication year. As is shown, most of the papers were published between 2006 and 2013.

Every year since 2004, approximately 20 papers about CSFs have been published, and since 2009, 30 or more papers a year have been published. Therefore, it can be argued that a review every two or three years is reasonable in order to update the results of previously performed literature reviews, especially when considering the rapidly evolving technology and changing system availability such as the concept of "Software-as-a-Service"-concept or ERP systems provided in the cloud.

TABLE 1. PAPER DISTRIBUTION BY YEAR

Year	Papers	Year	Papers
2013	30	2005	15
2012	31	2004	20
2011	39	2003	11
2010	37	2002	11
2009	42	2001	5
2008	22	2000	5
2007	24	1999	3
2006	24	1998	1

Overall, 31 factors influencing the success of ERP system implementation were identified. Figure 1 shows the results of our review, i.e., the CSFs identified, their ranks, and each factor's total number of occurrences in the reviewed papers. The factors *Top management support and involvement*, *Project management*, and *User training* are the three most-named factors, with each being mentioned in more than 160 articles. The factor *Top management support and involvement* stands out; it is ranked #1 and referred to in more than 200 papers. As mentioned above, we will not describe each factor and its concepts in detail in this paper. However, to provide a fuller understanding of the different CSFs and their concepts, we described all 31 factors in [19] and the top eight factors again in greater detail in [7].

Regarding the form of data collection, it must be stated that the papers consist of 144 single or multiple case studies, 118 surveys, and 58 literature reviews or articles where CSFs are derived from the chosen literature.

In most previous literature reviews, the CSFs were grouped without as much attention to detail; therefore, the number of CSFs used was lower (e.g., [4], [15]). However, we took a different approach in our review. For the 31 factors, we used a larger number of categories than other researchers because we expected the resulting distribution to offer more insight. If broader definitions for some CSFs might be needed at a later time, further aggregation of the categories is still possible. Comparing these results with other literature reviews (e.g., [15]), the top five factors are obviously similar, with only the ranked positions differing. Due to our large literature base, the total numbers of observed mentions are much higher. Therefore, the differences in the CSF frequencies are much higher as well, making the distinctions in the significance of the factors clearer.

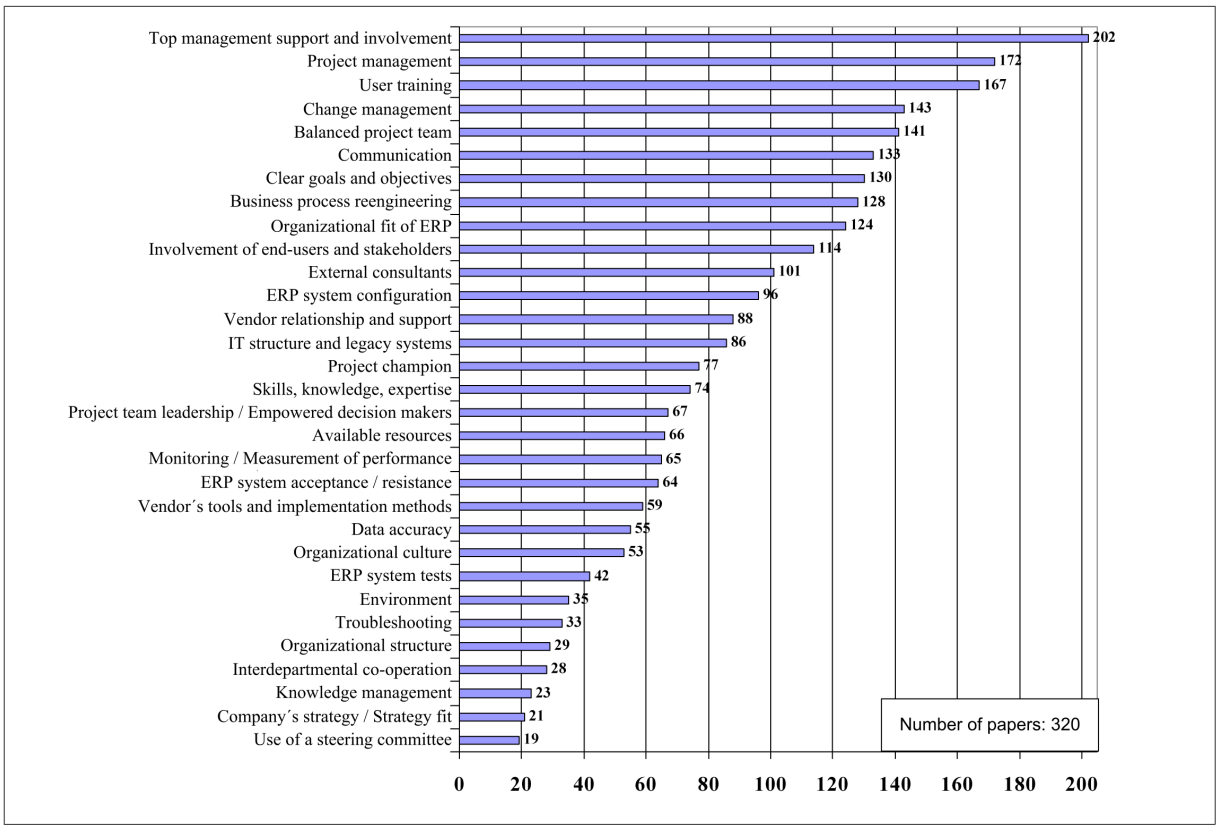


Figure 1. ERP Project CSFs in Rank Order Based on Frequency of Appearance in Analyzed Literature

Concerning company size during review 1 (conducted in mid-2010), only 12 papers explicitly focus on small and medium-sized enterprises (SMEs), mostly within single or multiple case studies. In the review update (conducted in mid-2013), 25 articles dealt with SMEs explicitly. In some surveys, SMEs are included and analyzed as well, but they are a minority in these surveys. Therefore, deriving CSFs that are important for SMEs is difficult and can be seen as still lacking in the CSF research.

Within these 37 papers focusing on SMEs, *Top management support and involvement* (mentioned in 25 articles), *Project management* (mentioned in 25 articles), and *User training* (mentioned in 22 articles) are again the most-frequently named factors for ERP projects in smaller enterprises. However, the differences in the CSF frequencies are only minimal and may be related to the small number of identified papers (see [30]). Therefore, deriving CSFs that are important for SMEs is difficult due to the small number of studies focusing solely on them. This is clearly a research gap in the research area of CSFs for ERP projects.

Therefore, our study focuses on this gap. We investigated these CSFs by interviewing SMEs that have implemented ERP systems. The results will be part of the following sections.

IV. Critical Success Factors for SMEs' ERP Projects

I. Study Design – Data Collection Methodology

To gain an understanding of the differences in the CSFs for ERP system projects in large-scale enterprises and SMEs, we used a qualitative exploratory approach within German small and medium-sized enterprises.

The units of analysis in our study are the implementation projects carried out within the SMEs. For the data collection, we conducted several interviews with members of the ERP implementation project teams, the department managers, or with the directors of the companies in order to identify the factors that they determined to be relevant for the success of their projects.

We interviewed employees of nine small and medium-sized enterprises located in Germany. The SMEs operate in different industry sectors and have implemented different ERP systems. Table 5 in the Appendix gives an overview of the companies and the interviewees.

Within these enterprises, there is a broad range of ERP systems (which cannot be named directly within this paper due to data protection). Only SME 1 and SME 3 have implemented the same ERP system; all other companies have implemented different systems. Some systems are quite small and industry specific and others are more widespread systems on the German ERP market. Most of the

implementation projects took place within the mid-2000s. All of the interviewees (except one of the department managers from SME 2) were directly involved in the ERP system implementation projects.

To gain an in-depth and detailed view of the enterprises and their structures, we chose direct structured interviews as our method of data collection. The interviews were conducted in retrospect to the ERP projects between April and July 2013. The interviews were designed as partially standardized interviews using open to semi-open questions as initial starting points for the conversation. Both personal (face-to-face) interviews, as well as telephone interviews, were conducted by the author. An interview guideline was developed based on the questions of [31], who conducted a similar study, as well as on the basis of one of our previous CSF-studies that had another focus [32]. We changed the questions to align with our identified CSFs (see Figure 1) to ensure that all of the factors were discussed in the interviews. The interview guideline consisted of 52 main questions with further sub-questions that referred to the 31 identified CSFs. These questions were formulated in an open way so that it would be possible to identify "new" CSFs that were not results of the literature review. This questionnaire was sent to interviewees before the interviews took place to allow them to prepare for their interviews. The complete listing of the formulated questions and their assignment to the success factors will not be part of this paper but will be provided by the author upon request.

For a better analysis of the results, we recorded all interviews (the interviews typically took between 70 and 180 minutes) and transcribed them afterwards (resulting in about 200 pages of written text). As a first step, non-verbal and para-linguistic elements and other elements that were not relevant to the study were excluded. To evaluate the CSFs, the interviews were analyzed with reference to each CSF question block, and the statements of the interviewees were classified according to a three-tiered scale (2–very important factor; 1–important factor; 0–less/non-important factor) as well as for a finer classification according to a five-tiered scale (4–very important factor; 3–important factor; 2–factor was seen as relevant; 1–factor was mentioned but not seen as being very relevant; 0–factor was not seen as relevant or important/factor was not mentioned at all). We used these two scales to get a preliminary understanding of whether differences would occur by using a finer/more-detailed scale. Here, the five-tiered scale could be seen as more appropriate to quantify the importance of the factors mentioned during the interviews. Also, we

TABLE 2. CSFS ACCORDING THE FIVE-TIER-SCALE RATING

Rank	Factor	Factor rating (five-tier-scale)	Rank	Factor	Factor rating (five-tier-scale)
1	User training	33	**18**	ERP system acceptance / resistance	16
2	ERP system tests	32	**19**	Change management	14
3	Organizational fit of the ERP system	31		IT structure and legacy systems	14
4	Clear goals and objectives	30	**21**	Troubleshooting	12
5	ERP system configuration	28		Organizational structure	12
6	Top management support and involvement	27		Data accuracy	12
7	Project team leadership / empowered decision makers	26	**24**	Knowledge management	11
8	Balanced project team	24		Monitoring and performance measurement	11
9	Communication	23	**26**	Project champion	10
	Involvement of end-users and stakeholders	23	**27**	Environment	6
	Company's strategy / strategy fit	23		Organizational culture	6
12	Available resources	21		Interdepartmental cooperation	6
	External consultants	21	**30**	Use of a steering committee	5
14	Business process reengineering	19	**31**	Vendor's tools and implementation methods	4
	Vendor relationship and support	19			
16	Project management	17			
	Skills, knowledge, and expertise	17			

4–very important factor; 3–important factor; 2–factor was seen as relevant; 1–factor was mentioned but not seen as being very relevant; 0–factor was not seen as relevant or important/factor was not mentioned at all) / maximum possible rating on basis of 9 interviews = 36

TABLE 3. COMPARISON OF THE TOP FIVE FACTORS

Rank	Results of the literature review (all company sizes)	Results of the literature review (only SMEs (see [22]))	Factors from the interviews
1	Top management support and involvement	Top management support and involvement	User training
2	Project management	Project management	ERP system tests
3	User training	User training	Organizational fit of the ERP system
4	Change management	Balanced project team	Clear goals and objectives
		Change management	
5	Balanced project team		ERP system configuration

TABLE 4. CATEGORIZATION OF CSFS (MODEL ADAPTED FROM ([20], [33], [34])

	Strategic		Tactical	
	Critical Success Factors	**Rank**	**Critical Success Factors**	**Rank**
Organizational	**Clear goals and objectives**	**4**	**User training**	**1**
	Top management support and involvement	**6**	**Communication**	**9**
	Project team leadership / Empowered decision makers	**7**	External consultants	12
	Balanced project team	**8**	Project management	16
	Involvement of end-users and stakeholders	**9**	Skills, knowledge and expertise	16
	Company's strategy / Strategy fit	**9**	Troubleshooting	21
	Available resources	12	Monitoring / Measurement of performance	24
	Business process reengineering	14	Interdepartmental cooperation	27
	Vendor relationship and support	14		
	ERP system acceptance / resistance	18		
	Change management	19		
	Organizational structure	21		
	Knowledge management	24		
	Project champion	26		
	Organizational culture	27		
	Environment	27		
	Use of a steering committee	30		
Technological	**Organizational fit of the ERP system**	**3**	**ERP system tests**	**2**
	ERP system configuration	**5**	IT structure and legacy systems	19
			Data accuracy	21
			Vendor's tools and implementation methods	31

discussed the factor rating with other researchers in this field to reduce the subjectivity of the rating.

II. Results of the Interviews

For each interview, a ranking of the critical success factors was set up by the author. A final ranking was created including all interviews and all individual rankings (see Table 2). As shown, the three most-important factors for ERP implementation projects in small and medium-sized companies according to our study are *User training*, *ERP system tests*, and *Organizational fit of the ERP system* with more than 30 out of a possible 36 points. Also, no further factors could be identified during the interviews. Each of the 31 factors stemming from the literature review was mentioned by at least one interviewee.

With the exception of *User training*, neither of the two other factors were part of the top five, within the ranking of the literature review (see Table 3). The factor ERP system tests was not even part of the top 20 in the literature review (see Figure 1). In addition, *Organizational fit of the ERP system* has gained more importance, according to our interviewees. The importance of both factors indicates that SMEs are forced to find the right ERP system that fits their needs and to test the system properly before the Go-Live. As mentioned in the first section, due to their lack of financial, material, and personnel resources compared to larger companies, failures during and/or after the Go-Live in SMEs can easily cause financial disadvantages or disasters, perhaps even leading to the insolvency of such small companies.

This is supported by the importance of the top two factors in our study. Reasons for this can be seen in the highly fragmented ERP system market as well as in the increasing multitude of software manufacturers and ERP systems. Enterprises are facing increasingly greater difficulties in identifying the best-fitting ERP system. Therefore, more emphasis is placed upon the selection of the "right" ERP system with a high *Organizational fit of the ERP system*. Together with selecting the right software, customizing the ERP system is also seen as an important factor for ERP projects in SMEs. The factor *ERP system configuration* is ranked at #5 in the interviews, whereas this factor was not part of the top 10 within the literature review. This also supports the statement that SMEs strongly depend on ERP systems that fit their needs, even more so than large companies do. SMEs cannot afford to be restricted by stiff ERP processes; moreover, it is important that the system is adapted according to the SME's own business processes.

To categorize critical success factors, in [20] a matrix scheme is suggested. Here, the authors consider the tactical or strategic direction of the CSFs and divide them into organizational and technological factors. Thus, tactical CSFs relate to short-term aspects and goals of the system implementation project itself, whereas strategic factors aim towards long-term impacts of activities with strong connections to the development of the organization in relation to mission, vision, and core competencies of the business activity. Considering the technological and organizational character of the CSFs, the specificity and significance of technological factors are strongly dependent on the ERP systems themselves, whereas organizational factors focus on corporate culture and its environment, with its specific processes and structures [20], [33], [34].

Table 4 gives an overview of the categorization of the identified CSFs in our study with a focus on their ranking. We oriented the classification and categorization of the factors according to [33] and [34]. The top 10 factors are highlighted.

It can be seen that only a few CSFs (six out of 31) are technological factors, whereas more than 50 percent of the factors (17 out of 31) are organizational factors with a strategic characteristic. However, the top 10 factors are spread out among all four categories, although most of them are part of the organizational category. Remarkably, two of the top three factors are part of the technological view. This supports the statement above that the technological aspects of ERP projects and their impacts on enterprises are considered more important for SMEs than for larger companies. However, smaller enterprises and ERP manufacturers should consider both the organizational and technological aspects when implementing an ERP system.

V. Conclusion and Limitations

The aim of our study was to gain insight in the research field of CSFs for ERP implementation projects, with a focus on ERP projects in small and medium-sized enterprises. Research in the field of ERP system projects and their CSFs provides valuable information that may enhance the degree to which an organization's implementation project succeeds [15]. As a first step, we carried out a systematic literature review to identify CSFs and to update existing reviews. Our review found a variety of papers, i.e., case studies, surveys, and literature reviews, focusing on CSFs. All in all, we identified 320 relevant papers dealing with CSFs of ERP system projects. From these existing studies, we derived 31 different CSFs (see Figure 1). Most of the identified papers and studies focus on large companies. Small and medium-sized enterprises are – if included at all – usually underrepresented in quantitative studies. Studies exclusively focusing on SMEs are rare. We identified 37 of the 320 articles with this explicit focus. These represent only 12 percent of all published papers with a focus on CSFs. Even though research focusing on CSFs in smaller companies has been recommended by the research community for several years (e.g., [17], [18]), our reviews reveal that SMEs are still not the primary focus of CSF research. Therefore, this still can be seen as a clear lack of research.

To address this concern, we developed a study with a specific SME focus. We conducted several interviews within SMEs that have implemented ERP systems. Using a guideline consisting of 52 initial questions about CSFs, we conducted nine interviews. We found that all 31 factors found in the literature review were mentioned by at least one interviewee, and therefore, all 31 factors also in some way affect the success of the ERP system projects in SMEs. However, contrary to the ranking resulting from the literature review, we identified factors with a more-

technological focus as being important for those ERP projects. Here, the factors *ERP system tests* and *Organizational fit of the ERP system* as well as *ERP system configuration* as part of the top five factors refer to more technological aspects. Hence, factors with an organizational characteristic could also be identified as part of the top five factors in our study (*User training*, *Clear goals and objectives*). For the interviewees, *User Training* is seen as the most important factor influencing the success of ERP implementation projects.

Regarding the research question, our study was able to show that all factors that influence the success of ERP system implementation projects in large-scale enterprises also influence ERP projects in SMEs. We could not identify any additional factors that were not already referred to in the literature. However, we could show that the importance of the factors differs a lot and that SMEs and also the ERP manufacturers have to be aware of these differences in the factors' characteristics, focusing also on technological aspects of the ERP implementations rather than focusing mainly on the organizational factors, as they are more important for large-scale companies.

A few limitations of our study must be mentioned as well. For our literature review, we are aware that we cannot be certain that we have identified all relevant papers published in journals and conferences, since we made a specific selection of five databases and five international conferences (see Tables 6 and 7 in the Appendix). Therefore, journals not included in our databases and the proceedings from other conferences might also provide relevant articles. Another limitation is the coding of the CSFs. We attempted to reduce any subjectivity by formulating coding rules and by discussing the coding of the CSFs with several independent researchers. However, other researchers may code the CSFs in other ways. For the interview study, the interviews conducted and data evaluated represent only an investigation of sample ERP projects in SMEs. These results are limited to the specifics of these enterprises. In light of this, we will conduct further case studies and several larger surveys to broaden the results of this investigation.

VI. References

[1] T.H. Davenport, *Mission critical: realizing the promise of enterprise systems*. Boston, USA: Harvard Business School Press, 2000, DOI: 10.1225/9067.

[2] S.V. Grabski, and S.A. Leech, "Complementary controls and ERP implementation success," *International Journal of Accounting Information Systems*, vol. 8, no. 1, pp. 17-39, 2007, DOI: 10.1016/j.accinf.2006.12.002.

[3] S.C.L. Koh, and M. Simpson, "Change and uncertainty in SME manufacturing environments using ERP," *Journal of Manufacturing Technology Management*, vol. 16, no. 6, pp. 629-653, 2005, DOI: 10.1108/17410380510609483.

[4] T.M. Somers, and K. Nelson, "The impact of critical success factors across the stages of enterprise resource planning implementations," in *Proceedings of the 34th Hawaii International Conference on System Sciences (HICSS 2001)*, Hawaii, USA, 2001, DOI: 10.1109/HICSS.2001.927129.

[5] Konradin GmbH, *Konradin ERP-Studie 2011: Einsatz von ERP-Lösungen in der Industrie*. Leinfelden-Echterdingen, Germany: Konradin Mediengruppe, 2011.

[6] A. Deep, P. Guttridge, S. Dani, and N. Burns, "Investigating factors affecting ERP selection in the made-to-order SME sector," *Journal of Manufacturing Technology Management*, vol. 19, no. 4, pp. 430-446, 2008, DOI: 10.1108/17410380810869905.

[7] C. Leyh, "Critical Success Factors for ERP System Implementation Projects: A Literature Review," in *Advances in Enterprise Information Systems II*, C. Møller, and S. Chaudhry, Eds. Leiden, The Netherlands: CRC Press/Balkema, 2012, pp. 45-56.

[8] A. Winkelmann, and K. Klose, "Experiences while selecting, adapting and implementing ERP systems in SMEs: a case study," in *Proceedings of the 14th Americas Conference on Information Systems (AMCIS 2008)*, Toronto, Ontario, Canada, 2008.

[9] A. Winkelmann, and C. Leyh, "Teaching ERP systems: A multi-perspective view on the ERP system market," *Journal of Information Systems Education*, vol. 21, no. 2, pp. 233-240, 2010.

[10] A. Jones, J. Robinson, B. O'Toole, and D. Webb, "Implementing a bespoke supply chain management system to deliver tangible benefits," *International Journal of Advanced Manufacturing Technology*, vol. 30, no. 9/10, pp. 927–937, 2006, DOI: 10.1007/s00170-005-0065-2.

[11] E.W.T. Ngai, T.C.E. Cheng, and S.S.M. Ho, "Critical success factors of web-based supply-chain management systems: an exploratory study," *Production Planning & Control*, vol. 15, no. 6, pp. 622–630, 2004, DOI: 10.1080/09537280412331283928.

[12] T. Barker, and M.N. Frolick, "ERP Implementation Failure: A Case Study," *Information Systems Management*, vol. 20 no. 4, pp. 43-49, 2003, DOI: 10.1201/1078/43647.20.4.20030901/77292.7.

[13] K. Hsu, J. Sylvestre, and E.N. Sayed, "Avoiding ERP Pitfalls," *The Journal of Corporate Accounting & Finance*, vol. 17, no. 4, pp. 67-74, 2006, DOI: 10.1002/jcaf.20217.

[14] P. Achanga, G. Nelde, R. Roy, and E. Shehab, "Critical Success Factors for Lean Implementation within SMEs," *Journal of Manufacturing Technology Management*, vol. 17, no. 4, pp. 460–471, 2006, DOI: 10.1108/17410380610662889.

[15] S. Finney, and M. Corbett, "ERP Implementation: A Compilation and Analysis of Critical Success Factors," *Business Process Management Journal,* vol. 13, no. 3, pp. 329-347, 2007, DOI: 10.1108/14637150710752272.

[16] F.F-H. Nah, K.M. Zuckweiler, and J.L-S. Lau, "ERP Implementation: Chief Information Officers' Perceptions of Critical Success Factors," *International Journal of Human-Computer Interaction*, vol. 16, no. 1, pp. 5-22, 2003, DOI: 10.1207/S15327590IJHC1601_2.

[17] B. Snider, G.J.C. da Silveira, and J. Balakrishnan, "ERP Implementation at SMEs: Analysis of Five Canadian Cases," *International Journal of Operations & Production Management*, vol. 29, no. 1, pp. 4-29, 2009, DOI: 10.1108/01443570910925343.

[18] A.Y.T. Sun, A. Yazdani, and J.D. Overend, "Achievement assessment for enterprise resource planning (ERP) system implementations based on critical success factors (CSFs)," *International Journal of Production Economics*, vol. 98, no. 2, pp. 189-203, 2005, DOI: 10.1016/j.ijpe.2004.05.013.

[19] C. Leyh, "Critical Success Factors for ERP System Selection, Implementation and Post-Implementation," in *Readings on Enterprise Resource Planning*, P.-M. Léger, R. Pellerin, and G. Babin, Eds. Montreal: ERPsim Lab, HEC Montreal, 2011, Chapter 05, pp. 63-77.

[20] J. Esteves-Sousa, and J. Pastor-Collado, "Towards the Unification of Critical Success Factors for ERP Implementations," in *Proceedings of the 10th Annual Business Information Technology (BIT) Conference*, Manchester, UK, 2000.

[21] K.-K. Hong, and Y.-G. Kim, "The critical success factors for ERP implementation: An organizational fit perspective," *Information and Management*, vol. 40, no. 1, pp. 25-40, 2002, DOI: 10.1016/S0378-7206(01)00134-3.

[22] T.C. Loh, and S.C.L. Koh, "Critical Elements for a Successful Enterprise Resource Planning Implementation in Small-and Medium-Sized Enterprises," *International Journal of Production Research*, vol. 42, no. 17, pp. 3433–3455, 2004, DOI: 10.1080/00207540410001671679.

[23] H. Appelrath, and J. Ritter, *SAP R/3 Implementation: Method and Tools*. Berlin, Germany: Springer, 2000.

[24] M. Al-Mashari, A. Al-Mudimigh, and M. Zairi, "Enterprise resource planning: A taxonomy of critical factors," *European Journal of Operational Research*, vol. 146, no. 2, pp. 352-364, 2003, DOI: 10.1016/S0377-2217(02)00554-4.

[25] M. Al-Mashari, and A. Al-Mudimigh, "ERP implementation: Lessons from a case study," *Information Technology & People*, vol. 16, no. 1, pp. 21-33, 2003, DOI: 10.1108/09593840310463005.

[26] F.F-H. Nah, J.L-S. Lau, and J. Kuang, "Critical factors for successful implementation of enterprise systems," *Business Process Management Journal*, vol. 7, no. 3, pp. 285-296, 2001, DOI: 10.1108/14637150110392782.
[27] J. Becker, O. Vering, and A. Winkelmann, *Softwareauswahl und – einführung in Industrie und Handel – Vorgehen bei und Erfahrungen mit ERP- und Warenwirtschaftssystemen*. Berlin, Germany: Springer, 2007.
[28] I. Teich, W. Kolbenschlag, and W. Reiners, *Der richtige Weg zur Softwareauswahl*. Berlin, Germany: Springer, 2008.
[29] J. vom Brocke, A. Simons, B. Niehaves, K. Riemer, R. Plattfaut, and A. Cleven, "Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process," in *Proceedings of the 17th European Conference on Information Systems (ECIS 2009)*, Verona, Italy, 2009.
[30] C. Leyh, "Which Factors Influence ERP Implementation Projects in Small and Medium-Sized Enterprises?," in *Proceedings of the 20th Americas Conference on Information Systems (AMCIS 2014)*, Savanah, Georgia, USA, 2014.
[31] F.F.-H. Nah, and S. Delgado, "Critical Success Factors for Enterprise Resource Planning Implementation and Upgrade," *Journal of Computer Information Systems*, vol. 46, no. 29, pp. 99-113, 2006.
[32] C. Leyh, and P. Muschick, "Critical Success Factors for ERP system Upgrades - The Case of a German large-scale Enterprise," in *Proceedings of the 19th Americas Conference on Information Systems (AMCIS 2013)*, Chicago, USA, 2013.
[33] J. Esteves-Sousa, *Definition and Analysis of Critical Success Factors for ERP Implementation Projects*. Barcelona, Spain, 2004.
[34] U. Remus, "Critical Success Factors for Implementing Enterprise Portals: A Comparison with ERP Implementations," *Business Process Management Journal*, vol. 13, no. 4, pp. 538-552, 2007, DOI: 10.1108/14637150710763568.

APPENDIX

TABLE 5. OVERVIEW OF THE SMES AND INTERVIEWEES (NUMBER OF EMPLOYEES IS ROUNDED)

Company	Industry sector	Number of employees	Go-Live of the ERP system	Interviewee
SME 1	Food industry	200 in total; 30 ERP system users	2005	Member of top management
SME 2	Manufacturing of metal goods / Machine-building industry	80 in total; around 75 ERP system users	2006	Managers of the departments *purchasing* and *service*
SME 3	Food industry	40 in total; 8 ERP system users	1998	Head of the department *accounting and IT*
SME 4	Heat treatment services	45 in total; around 20 ERP system users	2007	Plant manager
SME 5	Food industry	110 in total; around 50 ERP system user	2007	Head of the IT department
SME 6	Wood / Furniture industry	90 in total; 55 ERP system users	2000, complete system re-launch 2005	Member of top management (also manager of the IT department)
SME 7	Automotive industry	80 in total; 15 ERP system users	2005	Plant manager
SME 8	Manufacturing of metal goods / Machine-building industry	55 in total; 10 ERP system users	2012	Employee of the management (also member of the ERP project team)
SME 9	Food industry	75 in total; around 20 ERP system users	Within the end-1990s (Go-Live in several steps)	Head of the IT department

TABLE 6. SOURCES FOR THE LITERATURE REVIEW

Databases	Conferences
Academic Search Complete	AMCIS
Business Source Complete	ECIS
Science Direct	HICCS
SpringerLink	ICIS
WISO	Wirtschaftsinformatik

TABLE 7. SEARCH FIELDS AND SEARCH TERMS

Database + Search fields	Search terms / Keywords
Academics Search Complete: “TI Title” or “AB Abstract or Author Supplied Abstract”	ERP + success* ERP + failure ERP + crit* ERP + CSF ERP + CFF ERP + fact* "Enterprise system*" + success* "Enterprise system*" + failure "Enterprise system*" + crit* "Enterprise system*" + CSF "Enterprise system*" + CFF "Enterprise system*" + fact*
Business Source Complete: “TI Title” or “AB Abstract or Author Supplied Abstract”	
Science Direct: “Abstract, Title, Keywords”	
SpringerLink: “Title” or “Abstract”	
WISO: “General Search Field”	

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1191–1196
DOI: 10.15439/2014F426
ACSIS, Vol. 2

Algorithms for Automating Task Delegation in Project Management

Bogdan Pop
Department of Computer Science
Babes-Bolyai University
Kogalniceanu 1, 400084, Cluj-Napoca, Romania
Email: popb@cs.ubbcluj.ro, bogdan.pop@webraptor.eu

Florian Boian
Department of Computer Science
Babes-Bolyai University
Kogalniceanu 1, 400084, Cluj-Napoca, Romania
Email: florin@cs.ubbcluj.ro

***Abstract*—Project management can be defined as a complex set of activities that are performed by project managers, individuals or larger entities, that requires proper application of skills, knowledge, tools and techniques in order to reach or exceed project requirements. Two of the most important skills or techniques that greatly influence the end result of a project are task delegation and resource allocation. Poor decisions while delegating tasks or planning resources' allocation can result in defective project results. Many project management applications and models aid project managers in proper task delegation and resource allocation. This paper presents models and task delegation algorithms that can automate task assignment, thus reducing manual delegation, reducing loss and improving projects' end results.**

I. Introduction

DURING a project's lifecycle many factors can change at any given point in time. In order to preserve the scope and objectives of the project and its sub-projects, developers have to take countermeasures swiftly. A single change during a task or an unforeseen event can trigger a chain reaction and derail greatly the project's development. Changing the terms and environment of the project can also add additional risks and the development team must be able to asses the situation quickly and make the proper adjustments in order to deliver the project successfully. The longer it takes for such countermeasures to be considered and performed the losses are likely to be higher and growing.

Therefore projects should be proactively managed by continuously improving and detailing the plan of action as more detailed and specific information and accurate estimations become available during the project's lifecycle. To achieve this, the project managers and project owners and stakeholders need to easily and fully grasp all aspects of the project. The bigger the project the harder it is to generate reports and project statuses. This is why it is recommended that developers use proper project management applications for their projects, allowing more focus to actual work than planning and calculating reports and project statuses. Furthermore, project management applications simplify a variety of tasks that project managers must perform with the help of dedicated tools and features [1][3].

Since task delegation and resource allocation are one of the most important aspects that greatly influence the outcome of a project, it is clear that simplifying, streamlining and possibly automating task delegations and resource allocations within projects would have positive impacts on overall results, costs and profits.

A concept application that automatically assigns, without any human intervention, newly created tasks within a project and its deployment model have been previously presented in [4]. The end goal of the concept application was to ease the project manager's job, to minimise costs and maximise resource usage to its fullest. The paper presented the database model, the base algorithm and a deployment scenario as software as a service of the presented concept application. The algorithm was used to reduce the role of a project manager as known today, allowing the usage of the project manager's knowledge for actual development.

This paper is structured as follows. Section 2 briefly describes the current proposed web application, its distribution model and task delegation mechanism presented in [4] which have shown promising results while being tested and studied in comparison to some other popular project management applications currently available on the market [5]. Section 3 presents enhancements and additions that can be applied to the noSQL model and to the task delegation algorithm that may improve the performance of the system and its yielded results. The 4^{th} section describes additional tests and studies that can be performed on the amended system to asses if the changes made have improved the performance of the system or not. The final two sections present future developments and conclusions, respectively.

II. Current State of the Task Delegation Model and Assignment Algorithm

Achieving automation in the task delegation process can be done by using a number of methods, most of them quite new and derived from the field of artificial intelligence. This includes, but is not limited to neural networks, evolutionary algorithms or swarm intelligence algorithms.

However, the application developed based on the proposed model [4] can potentially store vast amount of information, which would complicate the A.I. algorithms, especially with respect to computation times. Since time management is a critical part of project management and plays an important

978-83-60810-58-3/$25.00

```
{
    "task name": {
        "keywords":
        [
            { "keyword_name_1": "optional_priority" },
            { "keyword_name_2": "optional_priority" }
        ]
    },
    "task details":[
        { "username": "username value" },
        { "project": "project name" },
        { "timeToComplete": "time" },
        { "completed": "date time" },
        { "deadline": "date time" },
        { "finishedOn": "date time" },
        { "assignedOn": "date time" }
    ]
}
```

Fig. 1. JSON representation of the Tasks super column as presented in [4]

role in a project's success, A.I algorithms may not be well suited for the task.

Different approaches such as the Gale-Shapely Courtship Algorithm described in [2] have been taken into consideration as well. However, the chosen model was the usage of classic iterative algorithms that create automation with respect to task delegation by applying a set of instructions to properly stored, sorted and indexed data available on the project's backend database.

Project data is being distributed across multiple nodes via NoSQL databases, specifically Apache Cassandra [6]. A case of why NoSQL is better suited than classic relational databases is also presented in [4].

The initial proposed model used NoSQL databases to store information regarding project tasks, their types, the people that have worked on them, the time required to complete each task, personnel availability for future tasks and more. An algorithm that used this data to programmatically determine the best match for a newly added task was created. Figures 1, 2, 3 and 4 present the original schema of the most important column families in the database while the initial task delegation method is presented by Alg. 1.

Fig. 1 stores information regarding the undergoing tasks of the project, such as metadata keywords, comprising project, deadline, if it were already assigned or finished or not. Fig. 2 stores the availability times of each user. Since the chosen database system has a default lexicographic indexing, each user's availability is stored by using a date and username key with its two parts separated by a hash tag. The value stored is the time during a day when a particular user is available. The performance score and number of tasks of each user based on task metadata keywords is also stored as shown in Fig. 3.

By automating the task assignment process the project manager was no longer required to manually perform delegations and was able to have a more direct impact in the actual development of the project, instead of only leading it. Moreover, programmatically assigning tasks resulted in fewer errors compared to those made by a human project manager. Therefore, the development costs and time required to complete projects were reduced [4].

Algorithm 1 Original task delegation algorithm as presented in [4]

```
def function addNewTask(task)

  foundUsers = null
  iter = 1
  taskAssigned = false

  while ( taskAssigned==false && iter<10 )
  do

    resetOverallScore(foundUsers)

    for keyword in task.keywords do

      foundUsers.push(
        getNUsersWithBestScore_taskAssign(
                keyword,
                Start = 10*iter -9,
                End = 10 * iter)
      )

      for user in foundUsers
        userScore =
           GetKeywordScore_userScores(
                keyword,
                user)
        user.overallScore += userScore
      end

    end

    foundUsers.sort_by { overallScore }

    count = 0
    while ( taskAssigned==false &&
                count<foundUser.size)
    do
      if foundUser[count].isAvailable?
        addTaskToUser(
                foundUser[count],
                task)
        taskAssigned=true
        return true
      end
    end

    iter = iter + 1
  end

  return false

end
```

```
{
    "DATE_1#username_1":[
        { "startTime_1":"endTime_1" },
        { "startTime_2":"endTime_2" }
    ],
    "DATE_1#username_2":[
        { "startTime_1":"endTime_1" },
        { "startTime_2":"endTime_2" }
    ],
    "DATE_2#username_1":[
        { "startTime_1":"endTime_1" },
        { "startTime_2":"endTime_2" }
    ]
}
```

Fig. 2. JSON representation of the userAvailability column family as presented in [4]

```
{
    { "keyword_1#username_1": "score_1#tasksNo" },
    { "keyword_1#username_2": "score_2#tasksNo" },
    { "keyword_1#username_3": "score_3#tasksNo" },
    { "keyword_2#username_4": "score_1#tasksNo" },
    { "keyword_2#username_1": "score_2#tasksNo" },
    { "keyword_2#username_2": "score_3#tasksNo" },
    { "keyword_2#username_5": "score_4#tasksNo" },
    { "keyword_3#username_1": "score_1#tasksNo" }
}
```

Fig. 3. JSON representation of the userScores column family as presented in [4]

III. Enhancements to the Current Model and Algorithm

The initial algorithm described in [4] and presented in Alg. 1 worked in the following manner: when a new task was created the algorithm looped through the taskAssign column family (Fig. 4) for each keyword, starting from top to bottom, from the best score to the lowest possible, in order to assign it to the best suited user. The algorithm then computed each user's overall score counting zero if a user's score was null for a specific keyword. Following that, the user that got the best score and was available for work, according to its userAvailability (Fig. 2) column data, was assigned the task. If the user with the best score could not complete the task on time, the next user with the best score lower than the previous user's score was selected. Finally, if no suited users were found, the task remained unassigned and the project manager had to manually perform the delegation.

The study performed on the model and presented in [5] revealed the aforementioned un-assignment issue. On the first projects and for the first tasks within them, no users were selected since not enough data on their performance was stored in the database. The proposed modifications are designed to solve this issue such that all tasks, no matter their creation time, initial phases of the project, during the project or near its closing, are all automatically assigned by the proposed concept

```
{
    { "keyword_1#score_1#tasksNo1": "username_1" },
    { "keyword_1#score_2#tasksNo1": "username_2" },
    { "keyword_1#score_3#tasksNo1": "username_3" },
    { "keyword_2#score_1#tasksNo2": "username_4" },
    { "keyword_2#score_2#tasksNo2": "username_1" },
    { "keyword_2#score_3#tasksNo2": "username_2" },
    { "keyword_2#score_4#tasksNo2": "username_5" },
    { "keyword_3#score_1#tasksNo2": "username_1" }
}
```

Fig. 4. JSON representation of the taskAssign column family as presented in [4]

```
{
    { "skillset_1#keyword_1#username_1": "score_1" },
    { "skillset_1#keyword_1#username_2": "score_1" },
    { "skillset_1#keyword_1#username_3": "score_1" },
    { "skillset_1#keyword_2#username_2": "score_2" },
    { "skillset_2#keyword_1#username_1": "score_1" },
    { "skillset_2#keyword_1#username_3": "score_1" },
    { "skillset_2#keyword_2#username_1": "score_2" },
    { "skillset_2#keyword_2#username_3": "score_2" }
}
```

Fig. 5. JSON representation of the userSkillsetScore column family

```
{
    { "keyword_1#username_1": "preference_score" },
    { "keyword_1#username_2": "preference_score" },
    { "keyword_1#username_3": "preference_score" },
    { "keyword_2#username_1": "preference_score" },
    { "keyword_2#username_2": "preference_score" },
    { "keyword_2#username_3": "preference_score" },
    { "keyword_3#username_1": "preference_score" },
    { "keyword_3#username_3": "preference_score" }
}
```

Fig. 6. JSON representation of the userPreferenceScore column family

application and no manual input from the project manager is required.

In order to achieve the proposed scope, the cause of the problem had to be determined. The issue was in fact lack of data within the database, so the solution was to pre-populate the database with relevant information regarding each user. This could theoretically be possible for different projects within the same company or entity. However, this may not be the case for each developed project and it is therefore not a viable and feasible solution.

The proposed solution is a fallback within the task assignment algorithm that would be triggered when the original algorithm could not automatically assign a task based on data it can access from the userScores, taskAssign and userAvailability columns. This trigger would use two new column families that would store information about the users, mainly their skill-set, including all their certified and non-certified ones as well as their preference to what kind of work and tasks they prefer. Their skill-set information could be automatically computed by using predefined scores for different types of certifications. Their non-certified skills could only be scored and measured by an authority within the company or within the project, such as a HR or management representative. Fig. 5 shows the database schema for the user's skill-set score.

The mechanisms for generating a user's preferences score are trivial, each user having access to the project management application in order to set their preferences. The database schema of the preferences score is similar to that of the user's skill-set score and is presented in Fig. 6, while Alg. 2 presents the modified task assignment delegation process.

Another approach for modifying the algorithm would be to always take into account the skillset score and preference score of every user on the project. If this approach were chosen, a balance between users' past performance score, their skill score and preference score should be selected.

There are a couple of options for balancing the three different scores as follows. The first one is to allow manual

Algorithm 2 Modified task delegation algorithm with skill-set and preference fallback

```
def function addNewTask(task)
  foundUsers = null
  iter = 1
  taskAssigned = false
  while ( taskAssigned == false && iter < 10)
  do
    resetOverallScore(foundUsers)
    for keyword in task.keywords do
      foundUsers.push( getNUsersWithBestTaskAssignScore(keyword,
                        Start = 10*iter -9, End = 10 * iter) )
      for user in foundUsers
        userScore = GetScoreForKeywordFrom_userScores(keyword,user)
        user.overallScore += userScore
      end
    end
    foundUsers.sort_by { overallScore }
    count = 0
    while ( taskAssigned == false and count < foundUser.size ) do
      if foundUser[count].isAvailable?
        addTaskToUser(foundUser[count],task)
        taskAssigned=true
        return true
      end
    end
    iter = iter + 1
  end
  # trigger that takes skillset and preference into account
  while ( taskAssigned == false && iter < 10) do
    resetOverallScore(foundUsers)
    for keyword in task.keywords do
      foundUsers.push( getNUsersWithBestSkillSetScore(keyword,
                        Start = 10*iter -9, End = 10 * iter) )
      for user in foundUsers
        uSkillsetScore = GetScore_userSkillsetScore(task.skillset , keyword,user)
        uPreferenceScore = GetScore_userPreferenceScore(task.skillset , keyword,user)
        uSkillsetPreferenceScore = uSkillsetScore*0.75 + uPreferenceScore*0.25
        user.overallScore += uSkillsetPreferenceScore
      end
    end
    foundUsers.sort_by { overallScore }
    count = 0
    while ( taskAssigned == false and count < foundUser.size ) do
      if foundUser[count].isAvailable?
        addTaskToUser(foundUser[count],task)
        taskAssigned=true
        return true
      end
    end
    iter = iter + 1
  end
  return false
end
```

selections within each company and within each project. This way, management personnel could modify the ratio based on their preference and desired outcome within a project. The second option would be hardcoding the ratio for the three scores directly into the algorithm.

In order to obtain the perfect balance and ideal ratio, multiple tests with different ratios should be performed on the same set of tasks within identical projects in order to obtain relevant, comparable data. Alg. 3 shows the modifications required to use the balancing factor between users' performance, skill-set and preference scores, with their ratios being 50%, 25% and 25% respectively. The ratios were empirically determined as the algorithm allows modification for each project within different companies. Future work may include a more formal, mathematical approach to establishing the best ratios for the algorithm.

IV. Testing the Proposed Model and Algorithm Enhancements

The first step towards testing the modifications performed on the model and on the algorithm is to implement and deploy the changes into the web application developed based on the initial model [7]. The original model and its corresponding web application have already been tested and studied by using a small scale project developed by a small three-man development team [5].

The study revealed that the proposed concept is successful and overall performance improves with each new project that is managed via the application. The original study has been performed on a small scale construction project, a treehouse building process, and the results and outcomes may be completely different on larger projects and with larger teams or integrated ones. Therefore, future studies and tests of the enhanced model and task delegation algorithm should be performed by using larger, more complex projects.

The study presented in [5] also shown some drawbacks that interfered with the project workflow in the initial steps of the development, mainly because there was no information stored that could be used to automatically assign tasks to workers. The result of this test has generated the proposed enhancements presented in this paper. It is straightforward that if more tests are performed the likelihood to discover even more extensions and improvements that can be applied to the proposed concept will increase. It is therefore recommended that more diverse tests should be performed.

Additionally, tests should be performed on identical projects in order to obtain the golden ratio between the three scores: user performance, skill and preference, described in section 3 of the current paper, ratio that could be used for hardcoding the balancing factor within the task assignment algorithm.

V. Future Work

Proactive project management usually requires the following tasks be performed: identifying the requirements and objectives of the project, addressing the concerns and expectations of the project owners / stakeholders, balancing the project

Algorithm 3 Modified task delegation algorithm with performance, skill-set and preference balancing

```
def function addNewTask(task)
  foundUsers = null
  iter = 1
  taskAssigned = false
  while (taskAssigned==false && iter<10)
  do
    resetOverallScore(foundUsers)
    for keyword in task.keywords do
      foundUsers.push(
        getNUsersWithBestTaskAssignScore(
                keyword,
                Start = 10*iter -9,
                End = 10 * iter)
          )
      foundUsers.push(
        getNUsersWithBestSkillSetScore(
                keyword,
                Start = 10*iter -9,
                End = 10 * iter)
          )
      for user in foundUsers
        uPerformance =
           GetScore_userScores(keyword,user)
        uSkillset =
           GetScore_userSkillsetScore(
               task.skillset, keyword,user)
        uPreference =
           GetScore_userPreferenceScore(
               task.skillset, keyword,user)
        userScore = uPerformance * 0.5 +
           uSkillset*0.25 + uPreference*0.25
        user.overallScore += userScore
      end
    end
    foundUsers.sort_by { overallScore }
    count = 0
    while ( taskAssigned==false &&
                count<foundUser.size )
    do
      if foundUser[count].isAvailable?
        addTaskToUser(foundUser[count],task)
        taskAssigned=true
        return true
      end
    end
    iter = iter + 1
  end
  return false
end
```

constraints: scope, budget, schedule, resources, quality and risks [1]. The current proposed models and algorithms address only scheduling of personnel. An important aspect that can be improved upon is scheduling of hardware and other non-human resources.

Future developments may include enhancements of the current model such that tasks could be assigned to multiple individuals or such modifications that each task may or may not have preceding tasks to create proper dependencies. The current additions to the model allows tasks assignment based on user skill and preference. Similarly, a simple extension could be added such that more difficult tasks are assigned to individuals or teams with more experience and easier tasks assigned to the rest of the development team. This can be achieved if the model is modified in such a way that the database stores the difficulty of each task which could be assessed automatically or by the project manager. Similarly, one can also define some tasks as urgent, and these tasks should be assigned quicker based on their emergency level.

Future work may also include additional features such as estimating budget costs and automatically determining any risks before they affect the project. Another important aspect that would dramatically increase the potential of the application would be automatically identifying the requirements and objectives of the project from client communications, and translating them into tasks that developers understand and know how to perform.

VI. Conclusions

The proposed application concept reduces the role of the project manager by automatically delegating tasks and shows potential for large growth. Further more, the original application has performed up to 26.77% better than those of the competitors, as shown in [5]. The same test also shown that there is at least a 20% margin for improvement, leaving room for continuous and future developments.

The few cases when manual input by project managers was required [4] have also been reduced by the modifications performed on the original model and presented in this paper.

References

[1] *A guide to the project management body of knowledge (PMBOK Guide)*, Fourth Edition, Project Management Institute, Inc., 2008
[2] D. Gale, *The two-sided matching problem. Origin, development and current issues*, International Game Theory Review, Vol 3, Nos. 2 & 3, p. 237-252, 2001
[3] H. Kerzner, *Project Management a Systems Approach to Planning, Scheduling, and Controlling*, Tenth Edition, John Wiley & Sons, 2009
[4] B. Pop, *Building an Automated Task Delegation Algorithm for Project Management and Deploying It As Saas*, Studia Univ. Babes-Bolyai, Informatica, Volume LVIII, Number 2, June, 2013
[5] B. Pop, F. Boian, *Comparative Study of Task Delegation Models in Software As a Service Project Management Applications*, Knowledge Engineering: Principles and Techniques Conference, KEPT, Cluj-Napoca, July 5-7, 2013
[6] http://cassandra.apache.org
[7] http://automated.pm

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1197–1206
DOI: 10.15439/2014F79
ACSIS, Vol. 2

Development of the Organizational Agility Maturity Model

Roy Wendler
TU Dresden
Faculty of Business Management and Economics
Chair of Information Systems, esp. IS in Manufacturing and Commerce
Helmholtzstr. 10, 01069 Dresden, Germany
Email: roy.wendler@tu-dresden.de

***Abstract*—The importance of organizational agility in a competitive environment is nowadays widely recognized and accepted. However, despite this awareness, the availability of tools and methods that support an organization in assessing and improving their organizational agility is scarce. Therefore, this study introduces the Organizational Agility Maturity Model in order to provide an easy-to-use yet powerful assessment tool for organizations in the software and IT service industry. Based on a design science research approach with a comprehensive literature review and an empirical investigation utilizing factor analysis, both scientific rigor as well as practical relevance is ensured. The applicability is further demonstrated by a cluster analysis identifying patterns of organizational agility that fit to the maturity model. The Organizational Agility Maturity Model further contributes to the field by providing a theoretically and empirically grounded structure of organizational agility supporting the efforts of developing a common understanding of the concept.**

I. Introduction

ORGANIZATIONAL agility is an important and relevant concept for more and more organizations in today's competitive and fast-changing environment. Especially in the software and IT service industry, organizations are faced with an environment of rapid technological changes, which are accompanied by just as much change in customers' expectations and requirements [1], [2]. In addition, the fact that software and IT have become essential components of many other products – consumer electronics, automotive products, etc. – has increased the competitive pressure further [3].

However, despite an increasing awareness that organizational agility is a key concept in coping with this competitive pressure, the term "agility" is nowadays often inflated by many organizations without reasonable seriousness. Agility is nothing that can simply be put into practice. The management of an organization has to understand that the organization itself cannot be agile, but its employees can be. However, people are not independent from their environment, and they have to share appropriate skills in order to work under agile conditions and with suitable technologies [4], [5]. Hence, the path to an agile organization is a development process affecting all parts of an organization from workforce through organizational structures and processes to technologies used and the overall organizational culture [6], [7]. This process shows that managing the transition to an agile organization is a complex and strategic task. To fulfill this task, the management of an organization has to go continuously through three steps: (1) assessing the current level of organizational agility, (2) identifying potential areas for improvement, and (3) planning, executing, and monitoring appropriate improvement actions. It becomes clear that supporting tools are necessary to accompany these steps.

Already at the first step – assessing the current level of agility – an organization is faced with a difficult challenge. The assessment of agility implies that the components of an agile organization are clearly described and that assessment tools or methods are available. However, there is a lack of a clearly defined framework for explaining agility from an organizational perspective [8], and, hence, there is no consensus about what constitutes an agile organization [9]. As a result, this missing consensus about the determinants and dimensions of organizational agility limits the applicability of research results in practice and restricts the possibility to develop useful assessment tools [10].

The aim of this study is to solve this problem by introducing a maturity model as an assessment tool for organizational agility. The core contributions are a theoretically and empirically grounded structure of the components of organizational agility supporting the efforts of developing a common understanding of organizational agility as well as a useful and practical tool that is able to actually reflect existent patterns of organizational agility in the software and IT service industry.

The remainder of the paper is structured as follows: Section II summarizes available assessment approaches of organizational agility. The research methods used for designing the maturity model are described in section III, while the maturity model itself is introduced in section IV. A first evaluation of the model based on a cluster analysis of empirical data is given in section V. The paper closes with a conclusion and a description of further research opportunities in section VI.

II. Literature Review

Due to the missing consensus about what determines organizational agility (see [9], [11] for a detailed discussion), a universal definition is also missing. The literature contains a huge variety of more or less comprehensive definitions, each heavily influenced by context and application domain. While a

978-83-60810-58-3/$25.00

detailed discussion of these definitions is beyond the scope of this paper, a number of authors have analyzed the differences in definitions (e.g. [7], [8], [12], [13]).

For this work, two definitions have been selected as a basis that fits well into the software and IT service context and complement each other in terms of content. First, the definition of Yusuf, Sarhadi, and Gunasekaran [14] generally describes the prevalent situation in the industry under consideration and emphasizes the role of customers as well as the importance of internal capabilities, structures, and people. They define agility as "the successful exploration of competitive bases (speed, flexibility, innovation proactivity, quality and profitability) through the integration of reconfigurable resources and best practices in a knowledge-rich environment to provide customer-driven products and services in a fast changing market environment" [14, p. 37]. This definition can be further extended to explain agility as "an effective integration of response ability and knowledge management in order to rapidly, efficiently and accurately adapt to any unexpected (or unpredictable) change in both proactive and reactive business / customer needs and opportunities without compromising with the cost or the quality of the product / process" [15, p. 411]. Here, the often unpredictable nature of change is further underscored. In addition, both definitions point out the essential role of knowledge in coping with these changes.

Taking these definitions as a foundation, the literature was analyzed according to existing assessment approaches of organizational agility prior to developing the maturity model. A useful summary is given in [16]. In general, these approaches can be roughly categorized into three groups: The first group consists of approaches assessing agility by various metrics. However, these approaches only focus on capabilities, omitting drivers or enablers of agility [16], and are often focused on specific subareas of an organization, for instance market-related activities [15] or the supply chain [17]. The second group utilizes methods like the analytic hierarchy process (AHP) to determine overall agility (e.g. [18]), while the third group is based on fuzzy logic (e.g., [16], [19]).

The available approaches suffer from some limitations regarding their applicability to determine the level of organizational agility in practice. This weakness stems either from a too specialized orientation and, hence, an insufficient reflection of the whole organization with its interaction of people, structures, process, and technologies, as outlined above, or from the utilization of relatively complex algorithms, limiting an intuitive and ad hoc usage by management. In addition, although the available approaches are able to determine the current state of agility, they normally do not support management in suggesting further actions for improvement or development.

The identified requirements of a comprehensive representation of the whole organization, an intuitive tool that is easy to use, a determination of the current state of organizational agility, and directions for further improvement can be fulfilled with a maturity model. A maturity model describes and determines the state of perfection or completeness (i.e., the maturity) of certain objects. The progress in maturity can be observed and managed by the definition of maturity stages or levels that measure the completeness of the analyzed objects via different sets of (multidimensional) criteria [20]. This explanation is well reflected in the definition by Becker, Knackstedt, and Pöppelbuß: "A maturity model consists of a sequence of maturity levels for a class of objects. It represents an anticipated, desired, or typical evolution path of these objects shaped as discrete stages. Typically, these objects are organizations or processes" [21, p. 213].

A systematic mapping study analyzing the field of maturity model research shows that there are no maturity models available tackling the field of organizational agility or agility in general [20]. This lack underscores the assumption that the introduced maturity model is able to contribute to the field of assessing organizational agility by proposing a new approach that has not been available yet.

III. Method

This study follows a design science research approach [22]–[24], with the purpose of developing a maturity model to assess organizational agility. Hereby, the main goal is to build a maturity model that is applicable in the software and IT service industry. To fulfill this aim, the development is based on theoretical work as well as empirical evidence.

Following [24] a typical design science project includes four basic phases: analysis, design, evaluation, and diffusion. The focus of this study is in the design and evaluation phases by describing the elements and the development of the maturity model, as well as a first proof of concept, by discussing the results of a cluster analysis in the targeted industry. The analysis phase has already been conducted prior to the work presented here and includes stating the problem relevance, formulating the research objective, and searching for existing solutions. The need for a comprehensive organizational view on agility is justified in detail in [9], [25], and an extensive literature review on maturity models is given in [20]. Both aspects have been briefly summarized in sections I and II.

Many maturity models are developed on a purely conceptual basis, and their utility is evaluated afterward, mostly by case studies. A portion of available maturity models is completely lacking any kind of empirical evaluation [20]. For the maturity model developed in this study, a different approach has been chosen to ensure its grounding in theory as well as in empirical evidence. In addition to a careful review of agility frameworks in the analysis phase [9], an exploratory quantitative survey with the overall aim of identifying the elements of an agile organization has been carried out worldwide among organizations of the software and IT service industry. This procedure prior to the design phase allows the author to include empirical evidence already in the initial design of the maturity model and, therefore, enhances its quality and applicability from the very start.

With the survey 437 valid and complete responses were collected. The sample is summarized in table I. The complete questionnaire and comprehensive descriptive results are found

TABLE I
SURVEY SAMPLE CHARACTERISTICS

Characteristic	Total	Ratio
Role within the organization:		
Chief Executive Officer	127	29.1 %
Chief Information / Technology Manager	36	8.2 %
IT / ICT Manager	59	13.5 %
Enterprise / IT Architect	155	35.5 %
Other (e.g. Managerial Board Members, other Senior Managers, ...)	60	13.7 %
	437	**100 %**
Location of the organization:		
Europe	259	59.3 %
North America	104	23.8 %
Asia	39	8.9 %
Other (e.g. Columbia, South Africa, Brazil, Australia, ...)	35	8.0 %
	437	**100 %**
Size (no. of employees) of the organization:		
less than 10	95	21.7 %
10 to 49	87	19.9 %
50 to 249	87	19.9 %
250 or more	167	38.2 %
n.a.	1	0.2 %
	437	**100 %**

in [26]. The survey results are used to identify possible dimensions of organizational agility by exploratory factor analysis [27]. In addition, a cluster analysis has been carried out to identify patterns of agile organizations. These clusters are used for a first proof-of-concept evaluation by comparing them to the structure of the developed model in section V.

IV. DESIGN OF THE ORGANIZATIONAL AGILITY MATURITY MODEL

This section introduces the Organizational Agility Maturity Model. Fig. 1 illustrates the structure consisting of three dimensions and four maturity stages. The development of the dimensions and stages is described subsequently. This structure is already the result of the second development iteration. The first version of the model had five maturity stages. However, the cluster analysis of the obtained survey data revealed that this structure of stages is not suitable to represent the empirical patterns of organizational agility appropriately (see section V).

Although the highest maturity stage is always the best one theoretically, many maturity models state that the highest stage should not automatically be the goal for every organization using the model. This is consistent with a perspective on maturity models where every stage yields a set of potential improvements. So, every organization has to individually assess its "optimal" stage. This has disadvantages for the practical applicability. Different interpretations and viewpoints may lead to difficulties for organizations in finding this optimal degree of maturity [20], [28], [29].

For the Organizational Agility Maturity Model, another approach has been chosen that is more related to the life cycle of an organization [29] that is going to become agile. The difference is that the single stages are not only seen as desirable improvement, but rather a representation of steps while evolving over time. Hence, the highest maturity stage is always the "final" goal. Although these differences are only nuances, they are important for interpretation and usage of the model [20], [29]. Therefore, the maturity model should be used by organizations that have the clear objective of achieving organizational agility and want to use the maturity model as a roadmap in accompanying this transition.

A. Dimensions

The proposed maturity model consists of three dimensions, each of them further detailed into two sub-dimensions. This structure was obtained by conducting an exploratory factor analysis on survey data from the software and IT service industry (see section III). Due to space restrictions, the whole factor analysis cannot be presented here in detail. Tables II and III contain a summary of the results and illustrate the obtained factor structure. In addition, table VI in the appendix lists the assessment items for every dimension. For more information about the survey, please refer to [26], [27].

The conducted survey includes agility-related elements extracted from 28 frameworks describing agility [9]. Hence, the dimensions of the maturity model incorporate a structure of organizational agility that is grounded in theory and based on the empirical investigation also existent in practice. The content of every dimension is described below:

Agility Prerequisites are the degree to which the people of an organization share agile values (mental prerequisites) and to what extent the organization establishes the required technological prerequisites to support agility.

- **Agile Values** include the establishment of an organizational culture following agile values like proactivity, responsiveness, trust, support of proposals and decisions of employees, and the handling of change as opportunity and chance. This culture is measured by the degree to which the agile values have disseminated throughout the organization.
- **Technology** represents the technological prerequisites supporting organizational agility by enabling efficient communication across all levels and departments; the sharing of information; and the utilization of standardized, comparable, and integrated technologies and information systems. Technology is measured by the dissemination of appropriate technological support across the whole organization.

Agility of People summarizes all necessary capabilities of the members of an organization to translate the agile values into actions. It is further distinguished into the capabilities of the workforce and the capabilities of managers to cope with change.

- **Workforce** is a very important sub-dimension and comprises mainly the capabilities of employees. They have to be multiskilled to reorganize themselves under changing

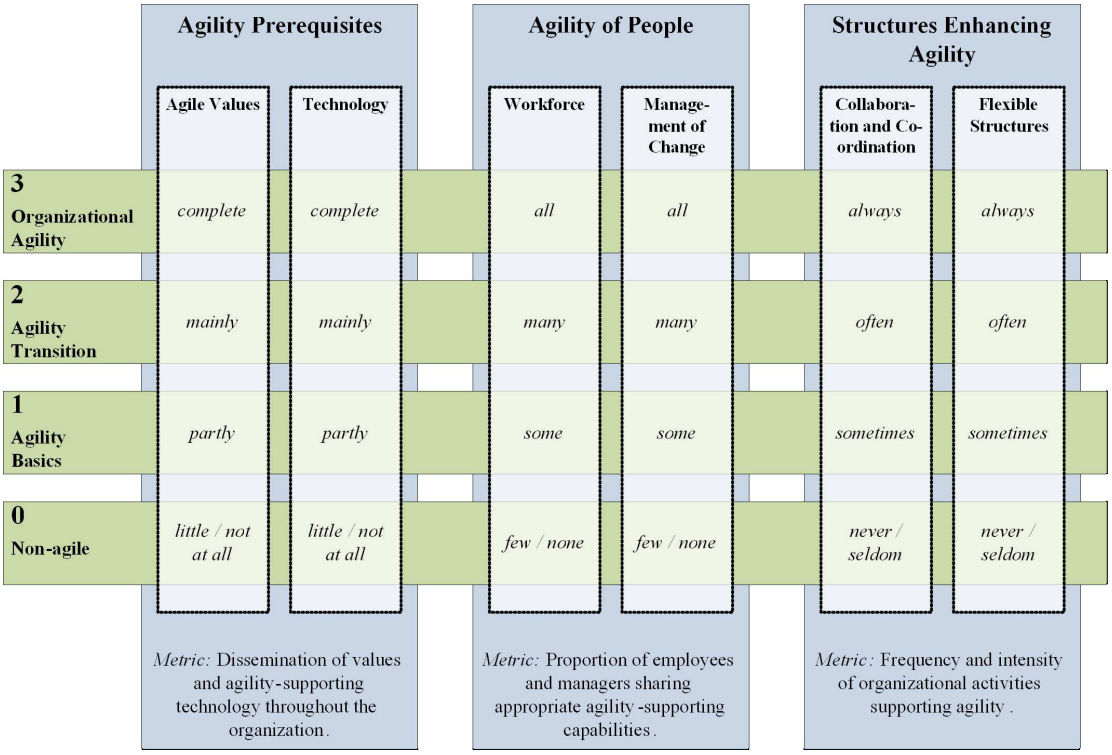


Fig. 1. Structure of the Organizational Agility Maturity Model

conditions. In addition, they should be able and willing to learn from each other to improve themselves continuously, communicate in a trustful way with each other, and take responsibility. Furthermore, they have to be able to think and act with quality and the market in mind. Workforce is measured by the proportion of employees of an organization sharing these capabilities.

- **Management of Change** involves capabilities, mainly of managers, to cope with changes appropriately and quickly (e. g., changing customer requirements, new markets, innovations, etc.). Managers have to inform the people of the organization accordingly and to inspire them to welcome these changes. In addition, they should be able to act with long-term vision and conduct IT investments strategically. Management of Change is measured by the proportion of managers of an organization sharing these capabilities.

Structures Enhancing Agility describes the ability of an organization to flexibly adopt and change itself combined with an organizational culture that supports collaboration and cooperation on every level.

- **Collaboration and Cooperation** summarizes activities of internal collaboration between departments and functions of the organization for decision making, new product/service development, etc. In addition, external cooperation with partners and customers focusing on quality, feedback, and intensive information sharing is covered by this sub-dimension. It is measured by the frequency of organizational activities enabling and supporting collaboration and cooperation.
- **Flexible Structures** describes the ability of the organization to quickly adapt organizational structures and processes to implement changes and stay competitive. Furthermore, it includes activities that enable quick decisions and a change of authorities when needed. Flexible Structures is measured by the frequency of organizational activities in establishing and incorporating flexibility.

B. Maturity Stages

The proposed maturity model consists of four distinct maturity stages that are assessed independently for every sub-dimension (see table VI). So, it may happen that an organization holds different maturity stages in the single sub-dimensions at a certain time. This difference is intended because the approach reflects the real state of the transition towards an agile organization, and it is unlikely that an organization is able to improve every aspect simultaneously and at the same pace. In addition, this approach enables an organization to determine further actions for a suitable path of improvement (see section V).

All six sub-dimensions are treated as equally important and the overall maturity score is simply the average of the single maturity stages. This unweighted and equal treatment is justified because the exploratory factor analysis revealed relatively equal proportions of explained variance between 0.09 and 0.17 for every factor (see Table III).

TABLE II
SUMMARY OF OBLIMIN ROTATED FACTOR ANALYSIS RESULTS

Factor	Item	Loading	Comm.
F1: Workforce	capemp6	0.92	0.87
	capemp7	0.84	0.80
	capemp9	0.77	0.65
	capemp5	0.74	0.68
	capemp8	0.73	0.75
	capemp4	0.73	0.67
	capemp10	0.71	0.65
	capemp11	0.60	0.62
	capemp3	0.59	0.61
	capemp2	0.57	0.66
	capemp1	0.51	0.68
F2: Technology	tech5	0.93	0.78
	tech6	0.78	0.63
	tech1	0.74	0.67
	tech3	0.62	0.75
	tech2	0.57	0.72
	tech4	0.55	0.58
F3: Management of Change	capman3	0.74	0.85
	capman5	0.72	0.73
	capman4	0.67	0.78
	capman1	0.59	0.73
	capman7	0.59	0.76
	capman2	0.57	0.65
	capman6	0.53	0.71
F4: Collaboration and Cooperation	actorggen12	0.75	0.63
	actorggen13	0.66	0.66
	actorggen16	0.58	0.69
	actorggen14	0.50	0.60
	actorggen10	0.45	0.67
	actorggen9	0.44	0.57
	actorggen6	0.37	0.67
	actorggen15	0.36	0.62
	actorggen7	0.36	0.62
F5: Agile Values	val1	0.69	0.59
	val5	0.68	0.61
	val4	0.64	0.67
	pref5	0.51	0.52
	pref1	0.47	0.59
	val2	0.46	0.51
	val3	0.45	0.61
F6: Flexible Structures	actorggen2	0.81	0.81
	actorggen3	0.78	0.76
	actorggen1	0.50	0.59
	actorggen5	0.43	0.51
	actorggen4	0.43	0.68

TABLE III
EIGENVALUE, CUMULATIVE EXPLAINED VARIANCE AND CRONBACH'S ALPHA OF FACTORS OBTAINED

	F1	F2	F3	F4	F5	F6
Eigenvalue	7.77	4.93	5.38	4.43	4.02	3.60
Cum. var. explained	0.17	0.28	0.40	0.50	0.59	0.67
Cronbach's Alpha	0.96	0.92	0.95	0.93	0.90	0.90

TABLE IV
DETERMINATION OF MATURITY STAGES REGARDING AVERAGE ASSESSMENT SCORE

Average score	Maturity stage
[1, 2.5)	0: Non-agile
[2.5, 3.5)	1: Agility Basics
[3.5, 4.5)	2: Agility Transition
[4.5, 5]	3: Organizational Agility

To determine the maturity stage of an organization, the assessment questions in table VI are used. Then, the average score is calculated for every sub-dimension. Finally, the organization is categorized to one of the maturity stages per sub-dimension according to the respective average score as outlined in Table IV.

The four maturity stages are:

0 – Non-agile: Organizations at maturity stage 0 show no or only rare properties of organizational agility. Agile values are principally unknown, and the technological basis is fragmented and unable to support communication processes effectively. Only a minority of employees and managers share capabilities necessary to implement agile values and actions. Hence, organizational activities for improving collaboration and cooperation and implementing flexible structures do not take place or only happen by chance. It may occur that single sub-dimensions show a higher score, but overall, these organizations are non-agile.

1 – Agility Basics: Organizations at maturity stage 1 share basic properties of organizational agility. Agile values and technological prerequisites underscoring agility are partly implemented in some but not the majority of departments, business areas, teams, or structural levels of the organization. Likewise, some but not the majority of employees share agile capabilities regarding communication, learning, responsibility, and customer-orientation, and some managers in the organization are able to manage change in an appropriate way. Often, these employees and managers are "concentrated" in single teams or departments. Activities to enhance collaboration, cooperation, and flexibility only take place sometimes, either by selective activities showing some "goodwill" or with a higher frequency but limited to a few agile departments or teams. These organizations have already realized and experienced the benefits of organizational agility, but in most cases only in some departments, teams, or situations, and therefore, the organizations only show some agility basics.

2 – Agility Transition: Organizations at maturity stage 2 manage to disseminate agile values and to establish an appropriate technological basis in most parts of the organization. Many employees and managers share the idea of agility and possess corresponding capabilities. Change is mostly welcomed and handled accordingly. In many instances, the organization carries out activities to support and promote teamwork and establishes organizational structures that are flexible enough to cope with upcoming changes. However, organizations at this maturity stage are characterized by weak-

nesses in one or two sub-dimensions of the model while others are already on a relatively high agility level (see section V for details). Hence, they are still in a transition phase towards a complete agile organization.

3 – Organizational Agility: Organizations at maturity stage 3 score high in every sub-dimension of the model and have overcome the partial weaknesses of the transition phase. They manage to establish a sufficient technological basis throughout the complete organization, and agile values are shared and accepted completely, too. All employees and managers have the capabilities to successfully work in an agile and changing environment. Collaboration and Cooperation are important aspects of everyday work and the structure is flexible enough to quickly and constantly react to upcoming changes. If any, there are only insignificant exceptions from the described agile attitude and behavior of the whole organization. Therefore, these organizations achieve complete organizational agility.

With this in mind, every maturity stage implies a specific goal while becoming organizationally agile. It is important for an organization to create an awareness of agility as an essential issue for staying competitive [1]. However, a particular solution affecting only one dimension of the maturity model is not sufficient, and the goal for maturity stage 1 is to get a basic understanding of agility with first transfers into practice in every dimension. This will create the foundation to generate agile solutions from the organization's own capacities [1]. Furthermore, the organizational changes that are needed have to be focused and appropriate to the characteristics of the organization, incorporating individual as well as organizational agility-related characteristics [1], [5]. Hence, the goal of maturity stage 2 is to get a clear vision of how organizational agility can be achieved, and based on this vision, a roadmap of the necessary actions has to be developed. Finally, the goal of maturity stage 3 is an equally matched interplay of all dimensions affecting agility: people, organization, and technology [5], [6].

V. Evaluation: Cluster Analysis

As a first step to evaluate the applicability of the Organizational Agility Maturity Model, a cluster analysis on the survey data has been performed to assess if the maturity stages are able to represent real-life configurations of organizations. As mentioned in section IV, the initial model had five maturity stages, due to the scales of the assessment questions used (see Table VI). However, after performing the cluster analysis, the lower two maturity stages were united to the form the maturity stage 0 (non-agile, see Fig. 1).

A. Clustering Method

To perform the cluster analysis, the dimensions of organizational agility have to be represented in the data. As described above, the sub-dimensions of the maturity model emerged from an exploratory factor analysis. Hence, average summed scales above a cut-off value of a factor loading of 0.3 were calculated for every factor [30] and used for cluster analysis. This procedure allows for the computation of a factor score for every sub-dimension, which is easily interpretable. The usage of the average allows that the original scale is retained [30], and the cut-off value of 0.3 ensures that only the variables that are included in the respective factor affect the resulting factor score (see Table II). The summed scales approach is especially suitable in this context because it is an exploratory research approach [30], [31].

Two important decisions in cluster analysis include the distance measure and the clustering method [31]. The cluster variables are the sub-dimensions of the maturity model. These sub-dimensions were extracted using an oblique rotation method (oblimin) in factor analysis and are, therefore, correlated to each other. To avoid distorted results because of correlated clustering variables, the Mahalanobis Distance has been used as the distance measure [31].

For clustering, the following procedure is recommended: First, a hierarchical approach should be used to determine the number of clusters, and second, a non-hierarchical approach should be selected to calculate the final cluster solution [31].

Here, a hierarchical approach using the Ward method was selected to estimate the number of clusters. This method is known to maximize in-cluster homogeneity by building clusters with a minimal increase of variance [31]. After that, the final cluster solution was calculated using a non-hierarchical approach, particularly fuzzy clustering [32]. Fuzzy clustering has been chosen because the clusters obtained by crisp non-hierarchical clustering methods turned out to be unstable when choosing randomized starting objects. However, the author recognized that a part of the data was stable. Such situations are well suited for fuzzy clustering because it bases the clustering on a membership function of every object to all determined clusters. In addition, it is possible to extract the "core objects" for further analysis. These objects have a high membership in one specific cluster (e. g., greater or equal to 0.7) and represent the most stable part of the data regarding the cluster solution [32]. The computation was carried out using the statistical software R [33].

B. Results and Discussion

Based on hierarchical clustering with the Ward method, a number of five clusters was the most appropriate solution. The results of the following fuzzy clustering approach are summarized in Table V. It shows the number of objects with different membership thresholds per cluster. Objects with a membership greater or equal to 0.7 have been considered as core objects and are used for further analysis.

To get an understanding of the clusters, Fig. 2 illustrates the mean values of the five clusters for every sub-dimension of the maturity model. Additional boxplots are given in Fig. 3. The first conspicuous aspect of the figure is that the lower part of the graphic is quite empty. This space does not mean that there were no respondents answering at the lower end of the scale used (see table VI) as the boxplots, particularly for cluster 3, show. But as cluster analysis reveals, they do not form a distinct cluster of their own. For this reason, the number of

TABLE V
RESULTS OF CLUSTER ANALYSIS (NUMBER OF OBJECTS) FOR DIFFERENT MEMBERSHIP THRESHOLDS ($n_{total} = 437$)

Cluster	Memb. $\geq$ 0.5	Memb. $\geq$ 0.7	Memb. $\geq$ 0.9
1	69	33	11
2	65	40	11
3	69	45	15
4	70	47	28
5	55	33	13
Total	**328**	**198**	**78**

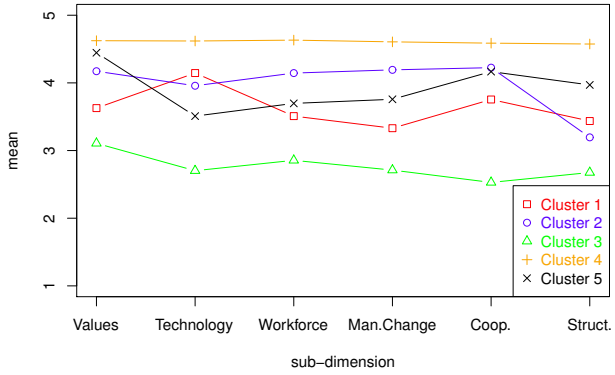


Fig. 2. Mean values of the clusters per sub-dimension

maturity stages has been reduced to four, combining the lower two scale values with no or nearly no agile attributes in the organizations (stage "0 – non-agile" in Fig. 1).

Furthermore, clusters 3 and 4 are easily distinguishable from the others. The organizations in cluster 3 score at a medium level on average regarding every sub-dimension. This cluster also includes organizations that have a lower agility assessment, but as a closer look into the data reveals, only in some of the sub-dimensions of the maturity model. Therefore, cluster 3 represents stage 1 (Agility Basics) of the maturity model where the covered organizations show initial initiatives and, hence, a basic development towards an agile organization.

Cluster 4, however, represents the opposite side of the scale. The organizations forming cluster 4 show a very high average score for every sub-dimension. As the boxplots in Fig. 3 illustrate, this cluster is the only one where 50 % of the objects score above an average of 4.0 in every sub-dimension, and, hence, describes the proportion of organizations that are most agile in the whole sample. This is represented by stage 3 (Organizational Agility) of the maturity model.

For clusters 1, 2, and 5 the interpretation is more complex. All have differing average values between 3.5 and 4.5 approximately. Therefore, the represented organizations are closer to each other but, nevertheless, show some distinguishing characteristics.

First, for cluster 2, we notice that most of the sub-dimensions show a relatively equal average score above 4.0 among the covered organizations and indicate a good advancement towards an agile organization. However, the score for the sub-dimension "Flexible Structures" clearly falls behind. This means that these organizations are characterized by a situation where a lot of agile potential (values, technology, capabilities, etc.) is lost due to structural obstacles. Their structures do not allow a fast adoption of processes, strategies, authorities, etc. to changing circumstances, and the agile potential has the risk of sticking to the team level [25]. This result is also consistent with literature where appropriate organizational structures are one central element to achieve organizational agility [6], [34].

Comparing the last two clusters (1 and 5) to each other, we recognize that they share an identical pattern for the dimensions "Agility of People" and "Structures Enhancing Agility" (the four sub-dimensions on the right in Fig. 2), with a slightly better average score for cluster 5. However, for the dimension "Agility Prerequisites," they show an opposite trend regarding "Agile Values" and "Technologies." While the organizations covered by cluster 1 score relatively high on "Technology," they score lower on "Agile Values." The opposite occurs in cluster 5.

From an interpretative perspective, this opposition means that the organizations in cluster 1 focus on the dissemination of agility-enhancing technologies. Technology is important because it is generally regarded as an essential enabler or driver of agility [9], [35], [36]. However, a pure concentration on technology also implies some risks. Increased IT spending, for instance, does not automatically lead to greater agility, and other elements represented in the maturity model have to be aligned with technology to achieve organizational agility [37].

The opposite situation is prevalent for the organizations in cluster 5. They score higher in every sub-dimension with the exception of "Technology." This lack may imply that these organizations are not yet aware of the mentioned role of technology as an enabler of organizational agility. However, in contrast to the organizations of cluster 1, they already manage to implement a culture based on agile values nearly completely. In addition, this is the only cluster, besides cluster 4, with an average score of 4.0 or above for the sub-dimension "Flexible Structures."

The organizations in clusters 1, 2, and 5 all would be assigned to stage 2 (Agility Transition) of the maturity model. They support the assumption of a transition phase that applies to the majority of organizations from the analyzed sample. This phase underscores that there are different approaches in becoming organizationally agile by concentrating on different dimensions or sub-dimensions of the maturity model. However, as cluster 4 clearly illustrates, it is important to achieve a balance between every dimension of the model. The considerations above show examples where an unbalanced or too focused improvement path may lead to risks instead of benefits.

The insights gained by conducting the cluster analysis above help to improve the structure of the developed maturity model. The main implications include the reduction of maturity stages

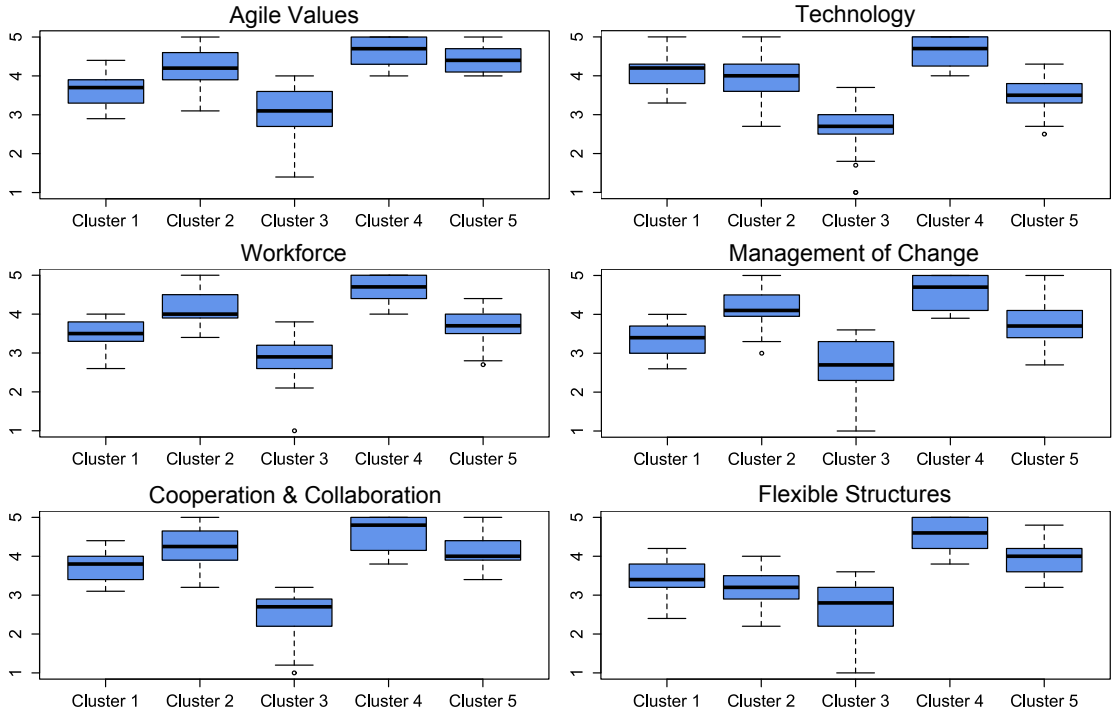


Fig. 3. Boxplots of the clusters per sub-dimension

from five to four to better reflect the empirically identified patterns of agility, a more precise naming of the maturity stages, and a better understanding of potential ways to improve agility by the three clusters assigned to the stage of Agility Transition. Furthermore, the fact that the identified clusters fit to the maturity model and are practically interpretable supports the principal structure and the applicability of the proposed Organizational Agility Maturity Model. Of course, further empirical investigation, for instance by case studies and expert interviews, is still necessary. Nevertheless, the proof-of-concept based on a cluster analysis of empirical data and the theoretically and empirically grounded development are already sufficient to confirm that the Organizational Agility Maturity Model is suitable to assess and describe the current state of organizational agility and to assist organizations from the software and IT service industry in taking further actions on their path to organizational agility.

VI. Summary and Conclusion

This study introduces the Organizational Agility Maturity Model as a new approach to assess organizational agility in the software and IT service industry. To fulfill the aim of achieving practical applicability and simultaneous theoretical grounding and rigorous development, a design-science research approach, including an extensive literature review [9], [11], [20] and an exploratory empirical investigation [26], [27], has been used. The maturity model is structured into three dimensions, each with two sub-dimensions, deduced from exploratory factor analysis, and four distinct maturity stages verified by cluster analysis.

The application of the maturity model creates useful benefits for organizations and underscores the strategic character of organizational agility. First of all, it generates an awareness of what constitutes organizational agility and creates an understanding about the complexity of organizational agility. Furthermore, it may serve as a reference frame to implement a systematic and well-directed approach for improvements and continuous assessment of actions taken.

The empirical investigation and the cluster analysis show that the industry under consideration is actually aware of the benefits of an increased organizational agility. Only a few organizations are classified as “Non-agile” in some dimensions, and they do not even form a separate cluster. Nearly all of the participating organizations have at least reached the stage “Agility Basics” and the majority is situated in stage “Agility Transition,” advancing towards “Organizational Agility.”

Interestingly, further analysis did not deliver any significant relationship of the clusters to the describing characteristics of the organizations like size, location, role, or customers. This lack indicates that organizational agility can be achieved by every organization that is really willing to take the actions necessary and that the maturity model is generally applicable to organizations in the analyzed industry.

Although an initial evaluation confirmed the applicability of the maturity model, further research should strive for additional validation. Of importance would be qualitative in-depth analyses, for instance by case studies or action research approaches, to validate the proposed stages as able to deliver helpful information for individual cases. In addition, the survey

used to identify the structure of organizational agility, and hence the structure of the maturity model, could be replicated with a different sample in other industries to check if the model is also applicable to other domains.

APPENDIX

Table VI lists the items that are used to assess the actual maturity stage. They are taken from the survey questionnaire (see section III) and represent the structure of organizational agility that was obtained by exploratory factor analysis [27].

REFERENCES

[1] J. Bessant, D. Knowles, G. Briffa, and D. Francis, "Developing the agile enterprise," *International Journal of Technology Management*, vol. 24, no. 5, pp. 484–497, 2002. doi: http://dx.doi.org/10.1504/IJTM.2002.003066

[2] P. P. Tallon and A. Pinsonneault, "Competing perspectives on the link between strategic information technology alignment and organizational agility: insights from a mediation model," *MIS Quarterly*, vol. 35, no. 2, pp. 463–486, 2011.

[3] K. Petersen and C. Wohlin, "A comparison of issues and advantages in agile and incremental development between state of the art and an industrial case," *Journal of Systems and Software*, vol. 82, no. 9, pp. 1479–1490, 2009. doi: http://dx.doi.org/10.1016/j.jss.2009.03.036

[4] K. Breu, C. J. Hemingway, M. Strathern, and D. Bridger, "Workforce agility: the new employee strategy for the knowledge economy," *Journal of Information Technology*, vol. 17, no. 1, pp. 21–31, 2001. doi: http://dx.doi.org/10.1080/02683960110132070

[5] D. Seo and A. I. La Paz, "Exploring the dark side of IS in achieving organizational agility," *Communications of the ACM*, vol. 51, no. 11, pp. 136–139, 2008. doi: http://dx.doi.org/10.1145/1400214.1400242

[6] S. L. Goldman, R. N. Nagel, and K. Preiss, *Agile competitors and virtual organizations: strategies for enriching the customer*. New York: Van Nostrand Reinhold, 1995.

[7] P. Kettunen, "Adopting key lessons from agile manufacturing to agile software product development - a comparative study," *Technovation*, vol. 29, no. 6-7, pp. 408–422, 2009. doi: http://dx.doi.org/10.1016/j.technovation.2008.10.003

[8] B. Sherehiy, W. Karwowski, and J. K. Layer, "A review of enterprise agility: concepts, frameworks, and attributes," *International Journal of Industrial Ergonomics*, vol. 37, no. 5, pp. 445–460, 2007. doi: http://dx.doi.org/10.1016/j.ergon.2007.01.007

[9] R. Wendler, "The structure of agility from different perspectives," in *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems*, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., Kraków, Poland, 2013, pp. 1177–1184.

[10] A. Charbonnier-Voirin, "The development and partial testing of the psychometric properties of a measurement scale of organizational agility," *M@n@gement*, vol. 14, no. 2, pp. 120–155, 2011.

[11] R. Wendler, "The structure and components of agility - a multi-perspective view," *Informatyka Ekonomiczna / Business Informatics*, vol. 2, no. 28, pp. 148–169, 2013.

[12] E. S. Bernardes and M. D. Hanna, "A theoretical review of flexibility, agility and responsiveness in the operations management literature: toward a conceptual definition of customer responsiveness," *International Journal of Operations & Production Management*, vol. 29, no. 1, pp. 30–53, 2009. doi: http://dx.doi.org/10.1108/01443570910925352

[13] A. Gunasekaran and Y. Y. Yusuf, "Agile manufacturing: a taxonomy of strategic and technological imperatives," *International Journal of Production Research*, vol. 40, no. 6, pp. 1357–1385, 2002. doi: http://dx.doi.org/10.1080/00207540110118370

[14] Y. Y. Yusuf, M. Sarhadi, and A. Gunasekaran, "Agile manufacturing: the drivers, concepts and attributes," *International Journal of Production Economics*, vol. 62, pp. 33–43, 1999. doi: http://dx.doi.org/10.1016/S0925-5273(98)00219-9

[15] A. Ganguly, R. Nilchiani, and J. V. Farr, "Evaluating agility in corporate enterprises," *International Journal of Production Economics*, vol. 118, no. 2, pp. 410–423, 2009. doi: http://dx.doi.org/10.1016/j.ijpe.2008.12.009

[16] Y.-H. Tseng and C.-T. Lin, "Enhancing enterprise agility by deploying agile drivers, capabilities and providers," *Information Sciences*, vol. 181, no. 17, pp. 3693–3708, 2011. doi: http://dx.doi.org/10.1016/j.ins.2011.04.034

[17] M. M. Weber, "Measuring supply chain agility in the virtual organization," *International Journal of Physical Distribution & Logistics Management*, vol. 32, no. 7, pp. 577–590, 2002. doi: http://dx.doi.org/10.1108/09600030210442595

[18] J. Ren, Y. Y. Yusuf, and N. D. Burns, "A prototype of measurement system for agile enterprise," in *Proceedings of the 3rd. International Conference on Quality, Reliability and Maintenance*, G. J. McNulty, Ed. Oxford: University of Oxford, 2000, pp. 247–251.

[19] N. C. Tsourveloudis and K. P. Valavanis, "On the measurement of enterprise agility," *Journal of Intelligent and Robotic Systems*, vol. 33, pp. 329–342, 2002. doi: http://dx.doi.org/10.1023/A:1015096909316

[20] R. Wendler, "The maturity of maturity model research: a systematic mapping study," *Information and Software Technology*, vol. 54, no. 12, pp. 1317–1339, 2012. doi: http://dx.doi.org/10.1016/j.infsof.2012.07.007

[21] J. Becker, R. Knackstedt, and J. Pöppelbuß, "Developing maturity models for IT management," *Business & Information Systems Engineering*, vol. 1, no. 3, pp. 213–222, 2009. doi: http://dx.doi.org/10.1007/s12599-009-0044-5

[22] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004.

[23] K. Peffers, T. Tuunanen, M. a. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007. doi: http://dx.doi.org/10.2753/MIS0742-1222240302

[24] H. Österle, J. Becker, U. Frank, T. Hess, D. Karagiannis, H. Krcmar, P. Loos, P. Mertens, A. Oberweis, and E. J. Sinz, "Memorandum on design-oriented information systems research," *European Journal of Information Systems*, vol. 20, no. 1, pp. 7–10, 2010. doi: http://dx.doi.org/10.1057/ejis.2010.55

[25] R. Wendler and A. Gräning, "How agile are you thinking? - an exploratory case study," in *Proceedings of the 10th International Conference on Wirtschaftsinformatik, WI 2.011*, A. Bernstein and G. Schwabe, Eds., no. February, Zurich, Switzerland, 2011, pp. 818–827.

[26] R. Wendler and T. Stahlke, "What constitutes an agile organization? - descriptive results of an empirical investigation," TU Dresden, Dresden, Tech. Rep. 68, 2014. [Online]. Available: "http://nbn-resolving.de/urn:nbn:de:bsz:14-qucosa-130916"

[27] R. Wendler, "Dimensions of organizational agility in the software and IT service industry - insights from an empirical investigation," *unpublished, currently under journal review at the time of submission*, 2014.

[28] H. J. Kohoutek, "Reflections on the capability and maturity models of engineering processes," *Quality and Reliability Engineering International*, vol. 12, pp. 147–155, 1996.

[29] T. McBride, "Organisational theory perspective on process capability measurement scales," *Journal of Software Maintenance and Evolution: Research and Practice*, vol. 22, pp. 243–254, 2010. doi: http://dx.doi.org/10.1002/smr.v22:4

[30] C. DiStefano, M. Zhu, and D. Mîndrila, "Understanding and using factor scores: considerations for the applied researcher," *Practical Assessment, Research & Evaluation*, vol. 14, no. 20, pp. 1–11, 2009.

[31] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate data analysis*, 7th ed. Harlow: Pearson Education, 2014.

[32] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. Hoboken, NJ: John Wiley & Sons, 2005.

[33] R Core Team, "R: A language and environment for statistical computing," Vienna, Austria, 2013. [Online]. Available: "http://www.r-project.org/"

[34] S. Nerur, R. Mahapatra, and G. Mangalaraj, "Challenges of migrating to agile methodologies," *Communications of the ACM*, vol. 48, no. 5, pp. 73–79, 2005. doi: http://dx.doi.org/10.1145/1060710.1060712

[35] D. Vázquez-Bustelo, L. Avella, and E. Fernández, "Agility drivers, enablers and outcomes: empirical test of an integrated agile manufacturing model," *International Journal of Operations & Production Management*, vol. 27, no. 12, pp. 1303–1332, 2007. doi: http://dx.doi.org/10.1108/01443570710835633

[36] Z. Zhang and H. Sharifi, "A methodology for achieving agility in manufacturing organisations," *International Journal of Operations & Production Management*, vol. 20, no. 4, pp. 496–512, 2000. doi: http://dx.doi.org/10.1108/01443570010314818

[37] Y. Lu and K. Ramamurthy, "Understanding the link between information technology capability and organizational agility: an empirical examination," *MIS Quarterly*, vol. 35, no. 4, pp. 931–954, 2011.

TABLE VI
ITEMS TO ASSESS THE ORGANIZATIONAL AGILITY PER DIMENSION (TAKEN FROM SURVEY QUESTIONNAIRE; SEE SECTIONS III, IV)

Dimension	**Assessment items**	**Scale**
Agility Prerequisites: *Agile Values* [val1-5, pref1,5]	Our organization values a culture that... ...harnesses change for competitive advantages. ...considers team work as integral part. ...accepts and supports decisions and proposals of employees. ...is supportive of experimentation and the use of innovative ideas. ...considers changing customer-related requirements as opportunities. Our organization prefers... ...a proactive continuous improvement rather than reacting to crisis or "fire-fighting". ...market-related changes (e. g. new competitors, preferences) to generate news opportunities.	1: not at all 2: little 3: partly 4: mainly 5: completely
Agility Prerequisites: *Technology* [tech1-6]	Our organization has Information Systems and Technologies that... ...make organizational information easily accessible to all employees. ...provide information helping our employees to quickly respond to changes. ...are appropriate to our needs and allow us to be competitive in the marketplace. ...enable decentralization in decision making. ...are integrated amongst different departments and/or business units. ...are standardized or comparable amongst different departments and/or business units.	1: not at all 2: little 3: partly 4: mainly 5: completely
Agility of People: *Workforce* [capemp1-11]	Our employees... ...are able to act with a view to continuous improvement of our products, services, processes, and/or working methods. ...are able to sense, perceive, or anticipate the best opportunities which come up in our environment. ...are able to meet the levels of product and/or service quality demanded by our customers. ...use a broad range of skills and can be applied to other tasks when needed. ...communicate with each other with trust, goodwill, and esteem. ...are ready to learn and are prepared to constantly access, apply and update knowledge. ...are in general always willing to continuously learn from one another and to pass their knowledge to others. ...obtain and develop appropriate technological capabilities purposeful. ...can re-organize continuously in different team configurations to meet changing requirements and the newly arising challenges. ...are self-motivated. ...take responsibility and think in a business-like manner.	1: none 2: few 3: some 4: many 5: all
Agility of People: *Management of Change* [capman1-7]	Our managers... ...maintain an informal management style with focus on coaching and inspiring people. ...understand the value of IT investments from a company-wide perspective. ...have the knowledge and skills necessary to manage change. ...are able to quickly implement changes in products and/or services. ...are able to recognize future competitive advantages that may result from innovations in products, services, and/or processes. ...are able to flexibly deploy their resources (material, financial, human, ...) to make use of opportunities and minimize threats. ...manage the sharing of information, know-how, and knowledge among employees appropriately.	1: none 2: few 3: some 4: many 5: all
Structures Enhancing Agility: *Collaboration and Cooperation* [actorggen6,7, 9,10,12-16]	In our organization, we... ...jointly and intensively operate throughout different functions and/or departments for strategic decision making. ...encourage early involvement of several departments and/or functions in new product and/or service development. ...inform ourselves systematically about information technology innovations. ...strategically invest in appropriate technologies and have a clear vision how IT contributes to business value. ...monitor the performance of our partners and subcontractors very closely. ...select our partners and subcontractors by quality criteria (rather than pure cost-based decisions). ...align all our activities to customer requirements and needs. ...encourage compilation and internal dissemination of information on customers needs. ...closely collaborate with and encourage fast feedback from our customers.	1: never 2: seldom 3: sometimes 4: often 5: always
Structures Enhancing Agility: *Flexible Structures* [actorggen1-5]	In our organization, we... ...scan and examine our environment systematically to anticipate change. ...react to approaching changes by immediately updating our business strategy. ...react to approaching changes by immediately updating our processes. ...are quick to make appropriate decisions in the face of market- and/or customer-related changes. ...change authorities when tasks change.	1: never 2: seldom 3: sometimes 4: often 5: always

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1207–1214
DOI: 10.15439/2014F324
ACSIS, Vol. 2

Comparison of architectures for service management in IoT and sensor networks by means of OSGi and REST services

Jarogniew Rykowski, Daniel Wilusz
Poznań University of Economics, Niepodległości 10, 61-875 Poznań, Poland
e-mail: {rykowski, wilusz}@kti.ue.poznan.pl

Abstract—In this paper two alternative architectures for service management in IoT and sensor networks are discussed. The first one is based on Open Service Gateway (OSGi) framework and Remote Services for OSGi (R-OSGi) bundle. The second architecture extends the notion of REST (Representational State Transfer) paradigm. There were few purposes of the extension. First, efficient, dynamic searching for devices capable of fulfilling certain requests within actual context was enabled. Second, both the devices and controlling services were distributed. Next, the devices were orchestrated in order to provide complex functionality. Finally, the access to the devices' functionality was standardized. OSGi-based solution was found simpler and better suited for homogeneous sensor networks, while more complex REST-based framework appeared as better suited for heterogeneous and widely distributed IoT devices and services.

I. Introduction

The Internet of Things (IoT) is the continuously evolving concept, which influences the business processes and even society on a global scale. Historically, the IoT was perceived as the intelligent network connecting objects, information and human beings to enable remote coordination of resources by people and machines [1]. First proposals of architectures of IoT networks were based on their natural predecessor – sensor networks. Nowadays IoT is perceived as cutting edge phenomenon, no longer limited to electronic identification of objects, but defined as technology integrating devices with information network, where these devices act as active participants in business processes [2].

Multiple applications of Internet of Things have been identified and implemented in such economy areas, as manufacturing, supply chains, energy, healthcare, automotive industry, insurance, financial services or research laboratories, to mention a few [2][3][4]. In the nearest future, the expansion of Internet of Things outside the internal infrastructures of companies is not only expected, but becomes the reality. The increasing popularity of Internet of Things causes the need for proper architecture, which will meet the requirements of IoT environment.

This work was supported by the National Centre for Research and Development under grant POLLUX-II-/1/2014

As IoT is very dynamic and heterogeneous, efficient management system for this environment should address these features. In this paper we discuss two alternative architectures for service management in IoT and sensor networks. The remainder of the text is structured as follows. Section II presents basic characteristics of IoT and sensor networks, which allows an enumeration of basic requirements of these environments. The Open Service Gateway (OSGi) framework is briefly described in Section III. Section IV focuses on the introduction of the Representational State Transfer (REST) methodology. OSGi-based architecture for IoT management is proposed in Section V. Next, similar architecture based on extended REST services is described in Section VI. Finally, Section VII concludes the paper.

II. IoT and sensor networks – basic features

At the first view sensor networks and Internet of Things are similar. Both networks are composed of small hardware nodes with limited resources, such as memory and CPU capabilities, both are physically distributed over certain area, both communicate by means of certain standards related with TCP/IP protocol. However, this similarity is seeming. Below, a comparison of basic features of these two network types , as well generic requirements for data acquisition and overall architecture are provided.

A typical sensor network is composed of many nodes having similar purpose, common goal and fixed functionality. Usually, one single point of interaction is assured, to contact the network as a whole, rather than particular nodes directly. As the nodes are usually small battery-operated hardware devices, several techniques are applied for energy saving and optimization of information routing.

Sensor network acts as single entity, thus it is usually controlled in the centralized manner by a single owner/administrator. The network is accessed as a "black box" of certain functionality, with strict access rights to particular global functions. As there is no need for individual addressing of devices and their functions, external access and control of individual nodes is usually blocked, and the

978-83-60810-58-3/$25.00

network is self-manageable. For example, for energy-saving reasons, some nodes are temporary deactivated, gathered information is pre-processed to minimize network transfer, etc. [5]

On the contrary, a typical IoT network is composed of heterogeneous nodes of different purpose and functionality. There is no common goal defined for the network as a whole except for very abstract goals such as "user comfort" or "energy savings". Instead, each networked IoT device, via continuous environment monitoring and information exchange with other devices, tries to act as a "good servant" [6] – invisible, however useful to maximum extent.

IoT networks are composed ad-hoc and as such they have no centralized management, single owner/administrator, global access rights, etc. The interaction with humans is incidental, sometimes even transparent to users – they are not necessary conscious about the functions provided, even if these functions rise the level of "user comfort" (such as automatic heating, air-conditioning, etc.), "energy savings" (automatic switching off of some devices once no human is detected in the neighborhood), "safety" (face scanning allows for transparent control of visitors of an office, rising "silent alarm" once an unknown person is detected inside), etc. This interaction depends on local and global context, based on such parameters as geo-location and environment features (temperature, lightness, etc.).

As the main stress is put on efficient ad-hoc interaction among humans and IoT devices, portability is the key. Interaction should abstract of such details as local addresses of IoT devices, communication means, data format, etc. Instead, an efficient searching/filtering mechanism should be provided, allowing to choose "the right" device to serve the request in given context. More precisely, portable interaction should be characterized by:

- a need for individual addressing of given network functions abstracting the implementation of these functions, including strict device addressing;
- a need for filtering "the best" device to be activated to realize certain request for function; similar – a need for searching for and choosing the "right" device(s);
- a need to distinguish several searching modes: "exactly one device capable of action X", "at least one", "everyone", etc.;
- a need for monitoring overall activity of devices (and accessibility of their functions) ;
- a need for portability due to ad-hoc nature of interactions at several places and for different situations: at least portability of requests – a need for common semantics/communication language to formulate the requests despite devices' specificity;
- a need for scalability, both for number of requests as well as the devices.

As may be drawn from the above enumeration of the needs, there must be a global (thus centralized) management mechanism for controlling the set of IoT devices, similar to a typical way of the management of the sensor networks. This is somehow contradictory to the requirement for the autonomy of the IoT devices and their ad-hoc composition and usage. However, without such centralized mechanism it is not possible to group the devices into bigger conglomerates (to orchestrate and thus multiply their functions), to search for the device which is optimal to serve given request, to balance the system load/energy efforts/network traffic, etc. Having in mind this trade-off, one may find that the optimum solution is to provide a local catalogue of devices' possibilities, extended by some statistical operations such as current load for each device, information about its accessibility and possibly temporal unavailability, number of served requests, last activation time, etc. The catalogue may also play a role of the request broker, mapping syntax and semantics of the incoming requests to the syntax and semantics used to contact the (heterogeneous by their nature) devices.

If starting the discussion on choosing the technology to implement the above-mentioned catalogue of the devices and their functions, we may point out two frameworks for centralized management of distributed resources: OSGi platform for Java programming language implementing a dynamic service registry, and extended REST services capable of centralized management of distributed REST resources and servers. In the next chapters we are going to present and compare these two frameworks, taking into account the following operations:

- registering the device/function;
- providing individual proxy for interaction/communication mapping – starting, stopping, and suspending/resuming the service;
- monitoring real-device state and providing information about device accessibility;
- searching for the device(s) to serve given request;
- if a request is to be possibly served by several devices, choosing one device based on context and statistical information collected during past activations;
- orchestrating device functions taking into account local context;
- administrating device parameters (individual GUI for each device).

III. OSGi FRAMEWORK

Open Service Gateway (OSGi) is a popular framework, which provides specification for dynamic module system for Java [10]. The core function of OSGi implementation is related with efficient management of modules (bundles) lifecycle, which may be dynamically installed, started, stopped, uninstalled, etc. Bundles are building block of OSGi-based modular systems, which are able to mutually interact. Following interactions among bundles may be distinguished:

- sharing Java packages with classes and interfaces;

- registering and calling services;
- managing the lifecycle of other bundles;
- sending and handling events to trigger specified actions.

OSGi implementations provides service registry to control functions provided by the bundles. There is no reference needed to invoke the service, as OSGi environment provides services based on their interface and additional metadata describing the service. The OSGi service registry may return zero, exactly one or multiple services compliant with the request. If there is no service of specified interface or meeting the additional constraints, null value is returned and must be properly handled. If the caller tends to invoke only one service, but in the registry there are few possible services meeting the requirements, OSGi environment returns the service of the highest rank (specified during service registration) and the lowest identifier (i.e., the oldest one) [11] [12].

Event based communication is a useful feature of the OSGi framework. Primarily it allowed handling data changes caused by dynamic behavior of the bundles or framework itself. Each bundle may be programmed to individually react to specified events, e.g., to stop if a dependent bundle is stopped or uninstalled. Since OSGi v.4, the mechanism of sending and handling the events defined by software developers has been introduced. The event-based communication allows the definition of events' topics and accompanied metadata, which are handled by dedicated `EventHandler` services [11][13][14].

Dynamic modularity, service registry and event-based communication seem to be out of the box solution for IoT systems. However, OSGi does not meet all the requirements listed in Section II that is the reason why some adaptations are needed. We propose such extensions further in Section V.

IV. REST

The Representational State Transfer (REST) style is an abstraction of the architectural elements within a distributed hypermedia system. REST ignores the details of component implementation and protocol syntax in order to focus on the roles of components, the constraints upon their interaction with other components, and their interpretation of significant data elements. It encompasses the fundamental constraints upon components, connectors, and data that define the basis of the Web architecture, and thus the essence of its behavior as a network-based application [7].

Key addressable entity of REST environment is a resource. Resources are any named pieces of information, being a target of hypertext links. Uniform Resource Locator (URL address) is used to get the value of a resource from the server: either static (if the resource is a file, piece of text, an image, etc.), or dynamic, being a result of an invocation of a piece of program code. In the latter case, the resource is treated as a "black box" from the point of view of its caller.

REST resources are frequently used to access the functionality of IoT devices, as this lightweight technology is well suited to limited hardware and software resources of the devices. Also, REST servers are used to proxy such access for very limited and non-standard devices [8]. Once the efficiency and simplicity of implementation rather than security is of primary concern, REST resources seem to be much better base than classical SOAP-based SOA services [8].

Although REST is a very useful proposition for the implementation of IoT-based framework, this technology must be substantially extended to meet the requirements presented in Section II. Thus, in Section VI we propose such extensions as a generic REST-based architecture for the Internet-of-Things environment.

V. OSGi-BASED MANAGEMENT

The IoT systems are characterized by their dynamic nature, as new devices appear, change their status including the availability, disappear or even suddenly break down. The OSGi framework with its support for dynamic modules seems to be the framework of the first choice to build on the IoT management system. Additionally, the OSGi service registry enables discovery of IoT specific features. Moreover, event-based communication allows ignoring the implementation details of services and substantially reducing the need for the standardization of interfaces. However, the IoT relies on network communication, which is not supported by pure OSGi specification. Another trait of IoT, which is difficult to be realized in OSGi framework, is a support for distributed vendors of IoT services. The next specific of IoT system is temporal unavailability of physical devices, which may be busy, broke down or temporary unplugged.

The requirement of choosing "the best" IoT service cannot be fully realized in pure OSGi framework because such choice is limited to fixed service ranking and age (e.g., a moment of first registration of the service). The last but not least important feature of OSGi-based IoT management system is the possibility to manage the devices manually by checking and modifying their status. The above described limitations were the main reasons to propose by us some extensions for OSGi. The architecture of extended OSGi framework for IoT management system is presented in Fig. 1

First, to deal with network communication among IoT services, the OSGi platform is extended by installing Remote Services for OSGi (R-OSGi) bundle developed in ETH Zürich [15]. R-OSGi provides dynamic proxy generation for remote invocation of services and register remote services, discovered in distributed registry, in local OSGi service registry.

Next, to handle modules provided by device vendors, the universal semantic (device descriptions) must be proposed, but this problem is out of the scope of the paper. Here, we

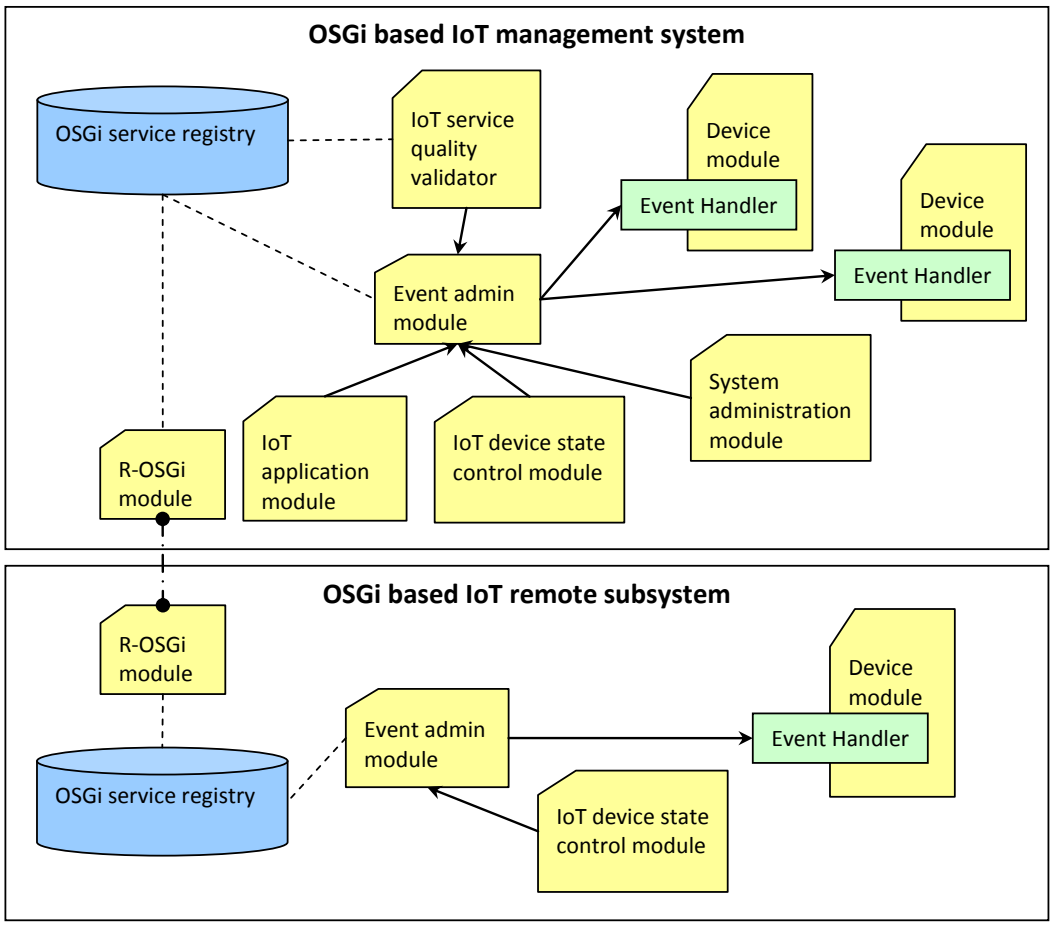


Fig. 1. OSGi based architecture for IoT environment

only point out the fact that OSGi enables specifying service metadata during service registration. The semantic information may be included in service metadata in the form of `java.util.Dictionary` object, containing OSGi event topics or additional properties.

In order to reflect the status of the IoT device related with given OSGi service, the metadata of the service may be dynamically changed, e.g., to restrict an access to unavailable devices. However, to continuously monitor the device status, such a function should be provided by the provider of the device module (corresponding OSGi bundle). IoT device-state control module, extending basic OSGi platform, should be able to send auditing events and get responses from related dependent services interested in the notifications of the device status.

The IoT management system should enable the user to invoke the most suitable service. As the capabilities of OSGi in this aspect are limited, the dedicated module needs to be provided. We propose to implement IoT statistics and validation module, which enable to validate the properties of the service and provide proper statistics on service performance. Such module should utilize the aspect oriented programming techniques to measure and store service properties such as performance time, moment of last invocation, number of invocations, number of generated exceptions (errors), etc.

For administration purposes, the OSGi console may be extended by the methods allowing manual management of available devices. The system administration module is responsible for the discovery of devices based on service metadata and allowing checking or changing the state of these devices.

We may list both the advantages as well as the disadvantages of the OSGi-based architecture, which are presented below.

The main disadvantages of OSGi based architecture are the following:

- Java dependent modules – the OSGi framework was developed for Java platform and in conclusion all module providers need to implement bundles in Java. As IoT is heterogeneous from its nature, the support for many programming languages should also be possible;
- limited built-in support for distributed services – OSGi was design to foster development of modular software running on one device (host), without taking into account distributed environments. The R-OSGi initiative mostly solves the problem, however the control over remote

services is limited and there is no built-in control over distributed bundles;
- lack of shared semantics – there is a need to propose standard semantic to enable unequivocal communication among system and the services – especially the ones provided by external entities;
- limited capabilities of service registry – the registry may provide additional information on services only in the form of `java.util.Dictionary` metadata. This solution is troublesome when extending the registry features to provide e.g., service quality information or device state control.

Despite of noticeable disadvantages, the OSGi framework is still prospective for building IoT management systems, as it has numerous advantages, which are listed below:
- support for dynamic modules – the bundle management allows to change the system capabilities in runtime. This feature is very useful, as in IoT system devices may be dynamically (dis)connected with the system at any time;
- event based communication – this trait of OSGi framework enables to separate the service invocation from its implementation. The OSGi events may carry orders in natural language, which are interpreted by event handlers, instead of directly invoking methods. What is more, requester does not need to know the location of a service, as event administrator sends events to all event handling services;
- code sharing and encapsulation – the OSGi bundles may share their code with one another, which prevents memory overhead. Additionally, each bundle directly specifies the code to share, thus enabling additional encapsulation by separating service interfaces from their implementations;
- compound services – the OSGi framework allows to compose compound services based on simple ones. The orchestration may be done just for calling specific feature, and the OSGi framework will provide the proper service and, after careful event handling implementation, even the most suitable for given request;
- providing interface for manual administration – OSGi framework provide possibilities to extend the console commands by a set of user-defined ones. This capability forms an easy way to implement IoT device administration interface.

VI. REST-BASED MANAGEMENT

The key requirement for the IoT system is an efficient method for the selection of device(s) for the incoming request in a given context. To this goal, the system must (1) know the devices' possibilities and characteristics, including their geo-locations (range of impact), (2) be able to search for the optimum device(s) to fulfill given request, and (3) activate the just-searched device in order to provide certain functionality. As already mentioned, these are basic tasks for a centralized catalogue. However, such catalogue is not a part of REST technology. Thus, we propose a uniform REST-based architecture with a dedicated catalogue (also REST-based) as the base for IoT system (Fig. 2). The architecture is based on using administrational extensions of REST servers to be observed as REST resources by the catalogue.

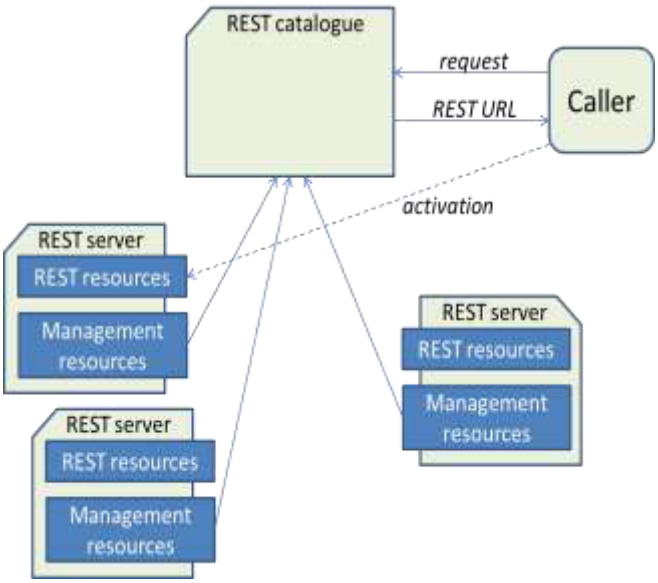


Fig. 2. REST-based architecture for IoT environment

The architecture is based on two principles: extending REST servers by some management resources, and providing mapping of semantically-expressed requests to URLs of REST resources.

To implement the first goal, we assume that each REST server is equipped with certain (predefined) resources to control and supervise the server, including:
- "management" resource to provide an access to some basic commands such as quit/suspend/resume/restart,
- "statistics" resource to provide some information about server usage (both divided to particular resources related with this server, as well as the server as the whole), for example: average service timings, number of requests served, number and description of invocation errors, etc.; this resource also provides an information about the real device (if any), connected to the resource and using this resource as its own proxy; in such way, the catalogue may be informed weather the device is accessible or temporary unavailable,
- "functionality" resource to provide some knowledge for the functions possibly served by this server (namely, by its resources); such function descriptions are defined according to common semantics for the whole system (c.f., interpretation of the request below).

Each REST server registers itself in the catalogue, providing its own URL locator. The locator (more precisely

– the resources mentioned above) may be periodically inspected by the catalogue to collect up-to-date information about server state. This information is used to search for ready-to-use devices (c.f., a description of request serving later on).

Once the internal state of the REST resource is changed, e.g., according to respective change of the state of real device connected to this resource, the server re-registers to the catalogue with the updated information.

Imagine one wants to activate certain function of the system. To this goal, he/she must address the catalogue with the semantic description of the desired action. This description is compared with the possibilities of the devices (however, only those declaring their state as currently accessible), and a device is chosen to meet the criteria. The caller obtains the locator of the resource linked with desired activity/device, to directly address respective REST service and, indirectly, the device. Note that the called URL was not known by the caller in advance, as it was (possibly dynamically) generated and sent by the catalogue. Once the situation is changed, some other resource/device may be activated according to the same request. Note also that the semantics of the URL locator of the resource to activate is not known to the caller, thus the details of the activation may be hidden towards the users of devices' functionality. This approach greatly improves portability of the system usage, on condition the semantics of the requests is common for all the callers and catalogues.

If given request is to be possibly served by several devices, the catalogue may choose one device based on context and statistical information collected from the management/statistical resources of the corresponding REST service. For example, one may address the less-overload resource, last-activated or most-unused device, the one with the shortest response time, etc.

We may also propose, instead of accessing a single resource for a single request, activating a set of resources – i.e., an orchestration of resources. To this goal, a mechanism is needed to map the request semantics to some program code, in turn responsible for the pipelining of the resources. The final result is provided as if all the activated resources are a single resource, thus the whole orchestration mechanism is transparent to the caller.

In most of the applications, connecting all the devices to a single host is not possible, due to (1) limited number of external connectors (such as USB), and (2) natural need for the distribution of the devices across a wider area. Thus, it is desirable to distribute not only the devices and hosts (proxies), but also the parts of the controlling framework. So far we assumed that there is only one central point for the control of the whole network. However, due to unrestricted distribution of REST resources it is possible to part this centralized point to a hierarchy of interconnected sub-parts, each one providing the control over certain network part (Fig. 3). To keep the control on the network as the whole, we propose to connect the sub-controllers as a graph and to span the controlling in the same manner as it is used to synchronize the resources of any peer-to-peer (P2P) network, with arbitrary restricted nesting level.

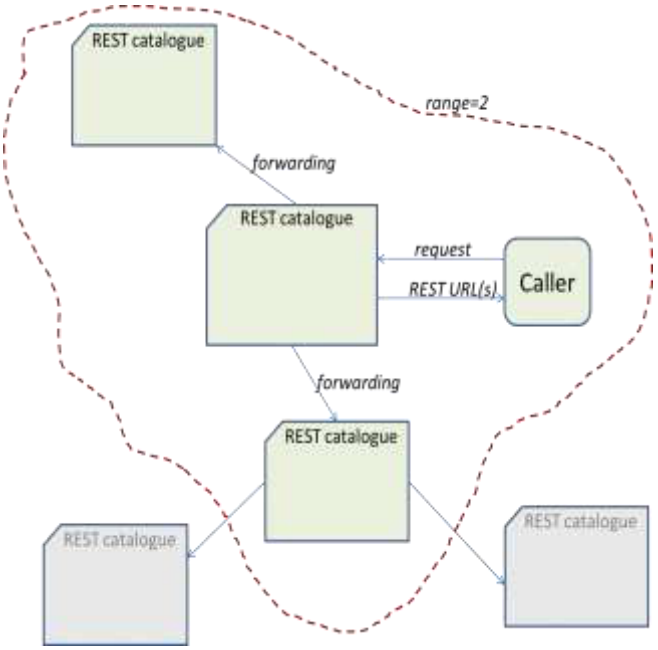


Fig. 3. Hierarchy of REST-based distributed management services

A device may choose any of the sub-controllers to register with. Then, this device is manageable locally by this sub-controlled in the direct way described above. Similar, all the local requests served by this sub-controller towards all its devices are processed locally. However, if there is a need to access remote (from the point of view of this controller) devices, then the sub-controller forwards the request to all its neighborhoods. In turn, if a neighbor is not able to process the request, forwards it to all its neighbors except the one which initiate the request, and so on. For each forwarding step, nesting level of the request (a range) is increased. While this level reaches certain value, the forwarding is stopped. Returning to Fig. 3, the request marked as "level=2" is forwarded only to three sub-controllers, while two "far" ones remain untouched. The stopping value is declared for each sub-controller by its administrator, and for the request by the initiator of this request – each time smaller of these two values is taken into consideration.

All the responses are collected by the forwarder of the request, and, if the request level is still greater than one, send as a common response to the caller. Finally, the originating node collects all the responses of all its neighbors, acting as a "global" response to the initial request.

To limit the possible cycles in the forwarding of the request, each request is identified, and the past-request identifiers are collected for some time in each of the sub-controllers. Once a request is coming already served in the past, this call is disregarded and no more forwarded. Thus,

even if the graph of interconnected sub-controllers contains cycles, these cycles are detected and never block the system.

In the same way we may obtain some global information about the network (statistics, information for certain-device of function availability, etc.), more precisely – about the local neighborhood ("no longer than N connections from the selected node"). More global is the request, more we must wait for the response, similar to typical P2P behavior.

As a typical IoT environment usually covers rather small geographical area (such as a room, a building or a public place – a market, a shop, a museum, etc.) – by restricting the level of spreading the requests to reasonable value one also limits the overall network traffic to the reasonable level, and the response delay is counting in parts of the second. We may also imagine restricting the bigger levels to those with special access rights, such as system administrators – for most of the requests, these will be addressed to local devices (level equal to 1). Then, both the possible delays and increased network traffic are not a sharp problem.

We may enumerate both the advantages and disadvantages of the proposed architecture, these are presented below.

The disadvantages are mainly related with two global observations – independence of the program code for the REST servers, and the need for shared semantics:

- there is no shared code even if identical parts of the program code (i.e., the libraries) are used by several REST servers. This feature results in large memory usage and may be relaxed by some shared-code techniques such as Dynamic Library Linking DLL;
- all the resources must know in advance the address of the catalogue, to register with. This restriction may be relaxed with non-standard usage of DHCP broadcasting;
- additional network traffic is observed to monitor device availability at real-time. However, with reasonable polling interval (for most of the system such timing as 5-10 seconds is completely enough) this restriction may be bypassed;
- one must provide strict definition of the semantics of requests and device activities (functions), shared for the whole system. This restriction is related with a need for the strict format of URL locators for additional REST resources (management/statistics). However, as this traffic is not observed by the end-users, this is a problem only for system designers;
- the activations of devices' functions are based on URL locators of the corresponding REST resources. Thus, only limited parameterization of such calls is possible – all the details must be coded and thus somehow hidden as URL locator parts. However, this is also mainly the problem of system designers, as end-users have no knowledge about the semantics of the device activations. Moreover, sometimes such information should be intentionally hidden for the end-users to limit the direct access to the devices (i.e., the access not controlled by the catalogue).

The advantages outweigh the above-presented restrictions and problems:

- it is a uniform approach – the whole traffic is realized as REST-compliant calls, which strongly facilitates the implementation of the system;
- small resources and servers dominate across the system, thus the consumption of computer resources such as CPU time is reasonably small. Our experiments showed that even hundreds of such servers on a single PC is not a problem, as a single resource typically consumes less than1% of computer resources;
- there is no problem with the synchronization of some resources "shared" among many REST servers, for example real devices, communication ports, etc. Such synchronization is needed only for a single REST server, however, as this server is programmed as a single entity and probably by a single programmer or small group of designers, such synchronization is easy to achieve;
- all the system parts (resources, devices) may be arbitrary distributed even in a local- or even wide-area network – on condition the distributed servers know and may access the catalogue host;
- REST servers may operate even with very limited hardware, also built-in to the networked devices,
- there is theoretically unlimited possibility of the orchestration of devices – providing "virtual devices" acting as real ones, however, possibly much more complex and powerful;
- the catalogue represents up-to-date information for device availability – continuous monitoring is undertaken not only for device state, but also some statistics for its usage;
- single- and group-based management for devices and their corresponding resources is possibly achieved, including server start, quit, restart, suspending/resuming, also GUI-based individual administration.

VII. CONCLUSIONS

As may be drawn from comparison of OSGi and REST based systems presented in Table I and discussed above, both approaches – OSGi-based and REST catalogue provide similar functionality and may be applied to implement an Internet-of-Things middleware. The amount of work for both approaches is similar – substantial extensions are needed to adapt the environment to the specific requirements of IoT applications. However, OSGi-based approach is better suited for sensor networks, i.e., the applications covering homogeneous devices and fixed, predefined system functionality, while the REST-based framework is more useful in ad-hoc, dynamic environment achieving heterogeneous devices and services. For the first, if we deal with shared device functionality and program code, we substantially limit memory usage. For the latter, we may

TABLE I.
COMPARING OSGI AND REST-BASED APPROACHES

IoT requirement	OSGi-based management	REST-based management
registering the device/function	X	X
providing individual proxy for interaction/communication mapping – starting, stopping, and suspending/resuming the service	X	X
monitoring real-device state and providing information about device accessibility	-	X
searching for the device(s) to serve given request	X	X
if a request is to be possibly served by several devices, choosing one device based on context and statistical information collected during past activations	X	X
orchestrating device functions taking into account local context (complex or virtual devices)	-	X
administrating device parameters (individual GUI for each device)	-	X
sharing basic libraries (functionality) for similar devices	X	-
adjusting scope of search for a device in the distributed access, managing a hierarchy of proxy servers	-	X

more easily provide distribution of proxies and devices as well as device/service orchestration, also in ad-hoc mode and based on some statistical information about past activations. If the amount of the shared code is low, due to the heterogeneity of devices/proxies, we do not observe the possible savings resulting from shared code, while still having the possibility of centralized management of the system as the whole as well as particular devices/resources.

REFERENCES

[1] D. L. Brock: The Electronic Product Code (EPC) A Naming Scheme for Physical Objects, Auto-ID Center, http://www.autoidlabs.org/uploads/media/MIT-AUTOID-WH-002.pdf,

[2] S. Haller, S. Karnouskos ,Ch. Schroth, "The Internet of Things in an Enterprise Context", in: Future Internet – FIS 2008 LNCS, vol. 5468, J. Domingue, D. Fensel, P. Traverso, Eds. Berlin Heidelberg: Springer-Verlag, 2009, pp. 14-28, doi: 10.1007/978-3-642-00985-3_2

[3] D. Wilusz, J. Rykowski, "The Architecture of Coupon-based, Semi-off-line, Anonymous Micropayment System for Internet of Things", in: Technological Innovation for the Internet of Things IFIP AICT, vol 394, L. M. Camarinha-Matos, S. Tomic, P. Graça, Eds. Berlin Heidelberg: Springer-Verlag, 2013, pp. 125-132, doi: 10.1007/978-3-642-37291-9_14

[4] D. Wilusz, J. Flotyński, M. Sielicka, "Supporting experimentation in a food research laboratory with the Internet of Things", in: PhD Interdisciplinary Journal no. 3/2013, Gdańsk: Gdańsk University of Technology, 2013, pp. 113-119.

[5] D. Waltenegus, Ch. Poellabauer, Fundamentals of wireless sensor networks: theory and practice, John Wiley & Sons, 2010

[6] M. Weiser, "The computer for the 21st century", in: Scientific American vol. 265, 1991, pp. 94-104, doi: 10.1038/scientificamerican0991-94

[7] R. T. Fielding, R. N. Taylor, "Principled design of the modern Web architecture", in: ACM Transactions on Internet Technology vol. 2 issue 2, New York: ACM, 2002, pp. 115-150, doi: 10.1145/514183.514185

[8] J. Rykowski , P. Hanicki, M. Stawniak, "Ontology Scripting Language to Represent and Interpret Conglomerates of IoT Devices Accessed by SOA Services", in: SOA Infrastructure Tools: Concepts and Methods, S. Ambroszkiewicz, J. Brzeziński, W. Cellary, A. Grzech, K. Zieliński, Eds. Poznań: Wydawnictwa Uniwersytetu Ekonomicznego w Poznaniu 2010, pp. 235-262

[9] D. Guinard, I. Ion, S. Mayer, "In Search of an Internet of Things Service Architecture: REST or WS-*? A Developers' Perspective", in: Mobile and Ubiquitous Systems: Computing, Networking, and Services, A. Puiatti, T. Gu, Eds. Berlin Heidelberg: Springer-Verlag, 2012, pp. 326-337, doi: 10.1007/978-3-642-30973-1_32

[10] OSGi Alliance Technology / HomePage http://www.osgi.org/Technology/HomePage

[11] J. Flotyński, K. Krysztofiak, D. Wilusz, "Building Modular Middlewares for the Internet of Things with OSGi", in: The Future Internet LNCS, vol. 7858, A. Galis, A. Gavras, Eds. Berlin Heidelberg: Springer-Verlag, 2013, pp. 200-213, doi: 10.1007/978-3-642-38082-2_17

[12] R. S. Hall, K. Paulus, S. McCulloch, D. Savade, OSGI in Action: Creating Modular Applications in Java, Greenwich: Manning Publications, 2011

[13] OSGi Alliance *OSGi™ Service Platform Release 4 Version 4.2* http://www.osgi.org/javadoc/r4v42/

[14] A. de Castro Alves, OSGi in Depth, Greenwich: Manning, 2011

[15] J. S. Rellermayer, G. Alonso, T. Roscoe, "R-Osgi: Distributed Applications through Software Modularization", in: Middleware 2007 LNCS, vol. 4834, R. Cerqueira, R. H. Campbell, Eds. Berlin Heidelberg: Springer-Verlag, 2007, pp. 1-20, doi: 10.1007/978-3-540-76778-7_1

9th Conference on Information Systems Management

THIS event constitutes a forum for the exchange of ideas for practitioners and theorists working in the broad area of information systems management in organizations. The conference invites papers coming from two complimentary directions: management of information systems in an organization, and uses of information systems to empower managers. The conference is interested in all aspects of planning, organizing, resourcing, coordinating, controlling and leading the management function to ensure a smooth operation of information systems in an organization. Moreover, the papers that discuss the uses of information systems and information technology to automate or otherwise facilitate the management function are specifically welcome.

Topics

The areas and topics of interest include, but are not limited to two groups:

- Management of Information Systems in an Organization:
  - Modern IT project management methods
  - User-oriented project management methods
  - Business Process Management in project management
  - Managing global systems
  - Influence of Enterprise Architecture on management
  - Effectiveness of information systems
  - Efficiency of information systems
  - Security of information systems
  - Privacy consideration of information systems
  - Mobile digital platforms for information systems management
  - Cloud computing for information systems management
- Uses of Information Systems to Empower Managers:
  - Achieving alignment of business and information technology
  - Assessing business value of information systems
  - Risk factors in information systems projects
  - IT governance
  - Sourcing, selecting and delivering information systems
  - Planning and organizing information systems
  - Staffing information systems
  - Coordinating information systems
  - Controlling and monitoring information systems
  - Formation of business policies for information systems
  - Portfolio management,
  - CIO and information systems management roles

Event Chairs

Arogyaswami, Bernard, Le Moyne University
Chmielarz, Witold, University of Warsaw, Poland
Karagiannis, Dimitris, University of Vienna, Austria
Kisielnicki, Jerzy, University of Warsaw, Poland
Ziemba, Ewa, University of Economics in Katowice, Poland

Program Committee

Bialas, Andrzej, Institute of Innovative Technologies EMAG, Poland
Christozov, Dimitar, American University in Bulgaria, Bulgaria
Csikosova, Adriana, The Technical University of Košice, Slovakia
DeLorenzo, Gary, California University of Pennsylvania, United States
Dima, Ioan Constantin
Espinosa, Susana de Juana, University of Alicante, Spain
Gafni, Ruti, The Academic College Tel-Aviv-Yaffo, Israel
Geri, Nitza, The Open University of Israel, Israel
Grabara, Janusz, Czestochowa University of Technology, Poland
Jelonek, Dorota, Czestochowa University of Technology, Poland
Kersten, Grzegorz, Concordia University, Montreal, Poland
Kobyliński, Andrzej, Warsaw School of Economics, Poland
Kohun, Frederick, Robert Morris University, United States
Koohang, Alex, Middle Georgia State College, United States
Lasek, Mirosława, University of Warsaw, Poland
Levy, Yair, Nova Southeastern University - Graduate School of Computer and Information Sciences (GSCIS), United States
Modrak, Vladimir, The Technical University of Košice, Slovakia
Niedźwiedziński, Marian, University of Lodz, Poland
Pańkowska, Małgorzata, University of Economics in Katowice, Poland
Pastuszak, Zbigniew, Maria Curie-SKlodowska University, Poland
Phusavat, Kongkiti, Kasetsart University in Bangkok, Thailand
Rizun, Nina, Alfred Nobel University, Dnipropetrovs'k, Ukraine
Rouibach, Kamel, Kuwait University, Kuwait

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1217–1225
DOI: 10.15439/2014F146
ACSIS, Vol. 2

Towards a Comprehensive Model for E-Government Adoption and Utilisation Analysis: The Case of Saudi Arabia

Saleh Alghamdi
University of Sussex,
Informatics Department
Brighton, UK
Email: sa434@sussex.ac.uk

Natalia Beloff
University of Sussex,
Informatics Department,
Brighton, UK
Email: N.Beloff@sussex.ac.uk

***Abstract*—E-Government increases transparency and improves communication between the government and the users. However, users' adoption and usage is less than satisfactory in many countries, particularly in developing countries. This is a significant factor that can lead to e-Government failure and, therefore, to the waste of budget and effort. Unlike much research in the literature that has utilised common technology acceptance models and theories to analyse the adoption of e-Government, which may not be applicable for e-Government acceptance analysis, this study proposes a more comprehensive and appropriate framework for analysing the significant factors that could influence the adoption and utilisation of e-Government in Saudi Arabia, as this is becoming a necessity.**

I. Introduction

INFORMATION and Communication Technologies (ICTs) are considered to be the backbone of many activities used nowadays. They have tremendous potential to provide solutions and to solve problems in different aspects, which then leads to enhanced quality of life. Given the fact that the accelerated development of Information Technologies (IT) has led to a rapid increase in the number of websites and services provided by governments, nearly all governments of countries around the world have at least a web presence, or so-called e-Government [1]. Currently, the role of ICTs is crucial in governance processes, where they can help to create a structured network for service delivery [2], effectiveness and efficiency [3], interactivity, accountability and transparency [4].

The term e-Government has various definitions in the literature [5] [6], most of which refer only to providing online services (E-Services) to citizens, whereas the definition of e-Government is, in fact, broader and more comprehensive. Although defining e-Government from a single perspective is easy, defining it from general view is relatively difficult [7]. However, we can provide a more comprehensive definition, namely the utilisation of various Information and Communication Technologies (ICTs) for facilitating communication between the government and the stakeholders (citizens, businesses and governmental agencies), providing effective, efficient and integrated e-Services that enhance the relationship between the government and the stakeholders through multiple and flexible channels, leading to a more democratic system and increased engagement.

The field of Electronic Government is growing considerably; therefore, many research areas need to be studied in order to provide scientific insights [8]. One of the most important elements of implementing E-Government systems is the interaction between users and E-Government systems, specifically the adoption and utilisation by users, who are the main target when implementing such systems. This interaction element is considered the main method for measuring the utilisation and success of E-Government systems. If there is no interaction between users and E-Government systems, this means that there is no benefit of implementing such systems. Therefore, this research will investigate and analyse factors that could influence the adoption and utilisation of e-Government in Saudi Arabia from different perspectives, which will then lead to better understanding, ensuring that e-Government has a high level of success.

This paper is divided into five sections. The first introduced the proposed definition of e-Governance. The second section provides a literature review and background information regarding e-Government in Saudi Arabia. The third section explains the research model. The fourth section provides a brief discussion of the research model and the development thereof. The final section provides a conclusion and details our planned future work to utilise, test and validate the proposed model.

II. Background and Context

A. Literature Review

E-Government initiatives are still in the early stages in most developing countries, and face many issues related to adoption, implementation and utilisation. Users' adoption of e-Government systems is less than satisfactory in many countries, particularly in developing countries. Although large amounts of money have been invested in e-Government initiatives in some of these countries, such as the Arab countries, they still have many challenges and shortcomings that slow the adoption and minimise the utilisation of e-Government

978-83-60810-58-3/$25.00

systems, which influence the success of such systems [9]. Most of the studies in the literature have focused on implementing e-Government from technical and structural perspectives [10]. Several studies also focused mainly on analysing barriers to and challenges of implementing e-Government [11]. However, few recent studies have been conducted to discover and analyse factors that can affect the adoption and utilisation of e-Government from the users' perspective [12] [2] [13].

Several models and theories in the literature have been developed to study the acceptance and the diffusion of technologies, including the Theory of Reasoned Action (TRA), the Technology Acceptance Model (TAM), the Diffusion of Innovation Theory (DOI), Perceived Characteristics Innovation (PCI) and the Unified Theory of Acceptance and Use of Technology (UTAUT) [14][15][16][17][18]. Most of the research in e-Government literature has utilised some of these common models and theories to analyse the adoption of e-Government, either by using their original forms, by adding certain constructs to them or by combining them. However, some of the models that have been used for analysing the adoption of e-Government in the literature were critically analysed in this research in order to evaluate their applicability for studying levels of e-Government's adoption and utilisation. This will help us to fill the gaps and to overcome shortcomings, which exist in the conducted studies, while developing this research model.

Some significant constructs involved in the analysed models are important and are supported in the literature; thus, they will be integrated into the current model. However, certain other constructs are not supported in the literature, which means they are not significant and will therefore not be used for analysing the adoption and utilisation of e-Government. Furthermore, a number of significant factors that are likely to have influence on the adoption and the usage of e-Government were not addressed in some of the analysed models. Table (I) shows some of the reasons that certain commonly used models are not applicable to the analysis of the adoption and usage of e-Government.

As stated previously, some recent studies in the literature have analysed e-Government adoption, and most have utilised the common models that are given in Table (1) [2] [22]. Therefore, the outcomes were limited due to the limitations that exist in the frameworks utilised. Several studies have amended the original forms by adding extra constructs, such as risk and trust [23] [12] [24], or by combining some common models [19].

B. E-Government in Saudi Arabia

The Saudi e-Government programme 'Yesser' was officially launched in 2005 [25]. Yesser is an Arabic word that means 'make it easy'. Consequently, the Yesser programme aims to provide services and information to citizens easily [26]. It serves as an enabler and facilitator for transforming the public sector to the information society, whereas each governmental agency is responsible for the actual execution of its own website and is responsible for effective coordination

TABLE I: Inapplicability of common models for the e-Government context

The Model	Inapplicability for the e-Government context
TRA	• The *Subjective Norm* construct is one of least understood aspects of TRA [15]. The *Subjective norm (SN)* construct is likely to have an indirect impact on *Behavioural Intention (BI)* via the *Attitude towards a behaviour (A)* construct. This will make the differentiation between the direct effect of *SN* on *BI* and the indirect effect of *SN* on *BI* via *A* more difficult. • There is a lack of significant constructs that can analyse the adoption and utilisation of large and complex systems such as e-Government.
TAM	• *External Variables* that have been proposed as an influential factor that affect the *Perceived Ease of Use (PEU)* and *Perceived Usefulness (PU)* are not fully explored in TAM [19]. • *Attitude towards using* does not fully mediate *Perceived Ease of Use (PEU)* and *Perceived Usefulness (PU)* [15]. • TAM cannot capture and specify the complete essence of e-Government usage behaviour [20], due to the lack of many important factors and constructs that have direct impact on behaviour in term of intention to use, as well as on the actual usage of technologies, particularly e-Government systems. • TAM excluded some important resources of variance, and did not consider other important factors that could prevent users from using information systems, such as time and money constraints [2].
DOI	• The *Trialability* construct, which refers to how easily a new technology can be used experimentally, cannot be applied to certain technologies and systems, including e-Government, due to the nature of the data and the number of users; thus, such systems are not trialable. This does not, however, mean that testing and validating the system is not essential. • The *Observability* construct, which refers to the extent to which the innovation provides visible and tangible results, can be implicitly integrated into the *Relative Advantage* construct. • DOI Model also lack certain factors and constructs that are crucial when analysing the diffusion and adoption of new technology, particularly e-Government.
UTAUT	• UTAUT did not address some very important constructs and factors, such as perceived awareness and quality of service,although these are highly likely to have a strong impact on behaviours and intentions to use and adopt technologies. It also did not address constructs related to reliability aspects such as security, privacy, trust and perceived regulations. Moreover, UTAUT did not address the influence of culture on the adoption and utilisation of technologies. Although UTAUT considered the influence of certain personal demographic factors, including age and gender, it did not address other important demographic factors, such as the user's location, the user's education level and the user's income, which are likely influence the utilisation and adoption level. • UTAUT ignored the influence of some constructs on the others, although this is very important in such analyses. • Grouping and labelling of items in UTAUT is problematic, since a variety of disparate items is combined to represent a single construct [21].

with the Yesser programme [27][28]. According to the e-Government's first action plan document, the vision of the Saudi e-Government initiative is to provide better government services to users; thus, the vision statement can be summarised as [28]:

"By the end of 2010, everyone in the Kingdom will be able to enjoy from anywhere and at any time —world-class government services offered in a seamless, user-friendly and secure way by utilizing a variety of electronic means."

To achieve this vision, the Saudi e-Government aims to provide 150 top-priority services for citizens and residents by 2010, and to make them available 24/7 with a 75% usage level and an 80% user satisfaction rating [29]. Unfortunately, the Saudi e-Government's vision has not yet been completely realised, despite the fact that it is currently 2014. An evaluation study was conducted by Al-Nuaim in 2011 to evaluate the Saudi ministries' websites, which were considered to be e-Government service providers. This study indicated that nine Saudi ministries, which represent 41% of the evaluated ministries, did not implement a true e-Government website. Moreover, ten ministries, which represent 45.4% of the evaluated ministries, were completely or partially in the 'web presence' stage. Three ministries, which represent 13.6% of the evaluated ministries, were in the 'one-way interaction' stage [26]. Some improvements might have been achieved during the period between Al-Nuaim's study and the present, but we believe that this period is not sufficient to accomplish that which was expected by 2010, based on the vision statement.

The focus of Al-Nuaim's study was the evaluation of e-Government service providers' websites, which measure the achievement of the first important aspect of the vision of Saudi's e-Government, namely providing over 150 top-priority e-Services. However, the focus of our research work will be on the utilisation of e-Government from the user's perspective, which represents the second crucial part of the Saudi e-Government's vision (75% usage level and 80% user satisfaction rating). This is a significant contribution of this research and is one that could lead to a better understanding of the factors that can influence the adoption and usage level of e-Government.

Unlike developed countries, research and studies that investigate and analyse the adoption and utilisation of e- Government in developing countries are limited. Few studies have been conducted on e-Government utilisation, e-Usage and e-Engagement in any country, and this is particularly true of Arab countries, such as Saudi Arabia, [7][4][30] and most have limited findings as a result of the implementation of limited methodologies or the collection of data from limited samples. Therefore, the lack of sufficient and comprehensive studies that explore and analyse factors that influence the adoption and utilisation of e-Government in developing countries is considered an important challenge that needs to be addressed. The lack of utilisation of e-Services that are provided by e-Government is problematic, even though some of these e-Services have been implemented well. Implementing huge national systems, such as e-Government, involves a tremendous amount of effort and cost; however, unless it is utilised as expected, the result is a disaster. Thus, studies and research on investigating and analysing the significant factors that could influence the adoption and utilisation of e-Government in Saudi Arabia are becoming a necessity.

III. The Research Model

In order to analyse factors that influence e-Government adoption and utilisation, this research developed a comprehensive model called the e-Government Adoption and Utilisation Model (EGAUM). This model provides a comprehensive framework to analyse key factors that have crucial influence on the utilisation and spread of e-Government. EGAUM was developed based on a critical analysis of the literature on technology acceptance, in conjunction with insights from several models and theories that are commonly used to analyse the acceptance and usage of technologies. Defects and shortcomings that exist in the models used in e-Government adoption literature are addressed in EGAUM. The main goal of this research model (EGAUM) is to determine factors that could influence the users' beliefs and intentions, as well as the behaviour that influences their adoption and usage levels.

The EGAUM model consists of three dependent variables, namely *Intention to Use E-Government (ITU)*, *E-Readiness of e-Government (ER)* and *Actual Adoption and Use of e-Government (AAU)*. EGAUM also contains four groups of independent variables, which are *Personal Factors (PF)*, *Motivational Factors (MF)*, *Technical Factors (TF)* and *Reliability Factors (RF)*. These independent variables represent the fundamental factors that have a critical influence on the adoption and usage levels of e-Government. The research model, EGAUM, is shown in Fig. 1.

The relationships between the research constructs are represented in two ways, namely by arrows (direct relationship) and small circles (indirect relationship). Direct relationship means that a construct has a direct influence. For example, *Personal Factors (PF)* and *Motivational Factors (MF)* have a direct influence on the *Intention to use e-Government (ITU)*. *Technical Factors (TF)* and *Reliability Factors (RF)* also have a direct influence on the *E-Readiness of e-Government (ER)*. However, an indirect relationship means that one or more factors of a specific construct has/have indirect influence; in other words, one or more factors of a construct can indirectly affect the influence of other constructs. For example, one or more of the *Motivational Factors (MF)* indirectly affect the relationship between *Personal Factors (PF)* and *Intention to use e-Government (ITU)*. Moreover, one or more of the *Personal Factors (PF)* has/have an indirect influence on the relationship between *Technical Factors (TF)* and the *E-Readiness of e-Government (ER)*.

A. Personal Factors (PF)

It is proposed in this research that demographic factors such as the users' age, gender, education and computer literacy levels, as well as users' locations and income are factors

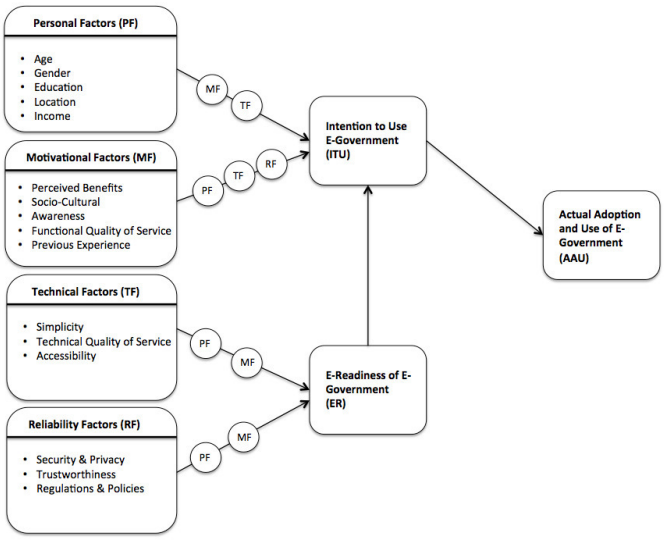


Fig. 1: The research model (EGAUM)

that potentially influence the adoption and utilisation of e-Government. It is further suggested that all these previous factors have a direct influence on the intention to use e-Government.

Age is an important factor that influences e-Government usage levels. Exploring the impact of the user's age on e-Government adoption will lead to better understanding of age-based aspects that need to be improved in order to increase the adoption and utilisation of e-Government. Several studies have found that age is one of the important factors that influence the adoption of Information Technologies (IT), particularly in the Arab countries [31][32]. Moreover, various studies have revealed different results regarding the influence of age on the adoption of e-Government. An early study found that older users tend to correlate positively with the use of e-Government [33]. On the other hand, some recent studies have found that older users are less likely to use e-Services, and that users under 25 years of age are more likely to adopt and use e-Services [34][35].

Gender is another crucial factor that needs to be addressed when analysing the adoption of e-Government systems, particularly in Saudi Arabia. Because of cultural and religious reasons, some Saudi governmental organisations and agencies allocate separate offices to serve women who need to perform governmental services physically. Although this division provides more privacy and comfort for the women and they are not as crowded as they would be in the main offices, these separate offices are not available in all cities, nor at all governmental organisations. This clearly influences the adoption and usage level on the part of women. Moreover, women are not allowed to drive cars in Saudi Arabia and there is insufficient public transport; thus, some women face difficulties when they need to go to government agencies. This also affects women's adoption and usage levels.

Education is also an important factor that needs to be investigated when analysing e-Government adoption and usage. This is because of the strong correlation between computer and information literacy and the education level of a user [36]. Computer and information literacy is defined as everything a person needs to know in order to be able to use computers [37]. Users with higher levels of education are more likely to adopt and use e-Government systems because they know how to use the Internet and computers, even if they are not Information Technology specialists. On the other hand, users with low levels of education are unlikely to adopt and use e-Government.

Location is another important factor that is not covered in any of the previous research into the adoption of e-Government. It is the right of all users to benefit from services that are provided by the government, regardless of whether they live in rural or urban areas; clearly, this poses a challenge for governments. In the context of Saudi Arabia, not all cities have governmental agencies and organisations that can provide governmental services for people who live in these cities. This challenge forces customers, who can be either citizens or business's representatives, to travel to the nearest city with the required governmental agency, a process that is both costly and time consuming. This will most likely have an effect on adoption and usage levels.

Income is another personal factor that will affect the intention to use e-Government systems. The EGAUM model included income as one of the Personal Factors (PF) that have a direct influence on the Intention to Use e-Government systems (ITU) because of two potential scenarios. The first suggested scenario is that income could prevent users from possessing computers and digital devices that could enable them to access e-Government services online, thus, preventing them from adopting and using e-Government systems. The second suggested scenario is that income could prevent users from travelling to the nearest city in the event that they live in rural areas without governmental offices in which to access governmental services physically. In both scenarios, the user's income is suggests a direct influence on the intention to use e-Government.

B. Motivational Factors (MF)

The Motivational Factors (MF) construct is one of the most important aspects to be investigated and analysed with regard to the adoption and utilisation of e-Government. This construct is comprised of four significant factors that are likely to influence the adoption and utilisation of e-Government.

Perceived Benefits in the E-Government Adoption and Utilisation Model (EGAUM) can be defined as 'the degree to which e-Government provides functional and non-functional benefits to its users'. This is related to the belief in gaining benefits and the expected outcomes of conducting e-Services using e-Government systems. The idea is to analyse the influence of perceiving the advantages of using e-Government systems on the adoption and utilisation thereof. The Perceived Benefit factor is classified into two categories:

- Functional benefits, which refer to 'any tangible benefits that a user can obtain from using the e-Government system'. These functional benefits include completing the intended services online from anywhere and at any time, synchronising and updating records among different governmental organisations, saving the history of conducted services, tracking the status of the conducted services, preventing bias from affecting the process of the conducted service and reforming the bureaucracy that exists in governmental procedures.
- Non-functional benefits refer to 'any intangible benefits that the user can obtain from using the e-Government system'. These non-functional benefits include time saving, cost saving, the comfort and convenience of conducting the service, avoiding crowds and long queues, reducing effort and independence, particularly for women in Saudi Arabia, as certain limitations exist.

Much of the research in the literature emphasised the importance of analysing the influence of the perceived benefits of a new technology on the adoption and diffusion of that technology [15] [16] [38] [39].

Socio-Cultural factor is also likely to have a strong impact on the implementation of a new technology, especially when that new technology is related to improving the lifestyle of society. It is a combination of two influential aspects, namely the social aspect and the cultural aspect. There is a strong correlation between them, since one aspect has the power to change the other. Social influence can be defined as "the normative pressure of associated members family or friends influences the intention to use e-government" [24]. Numerous research in the literature found that social influence, which is represented by friends, family members and colleagues, has a strong impact on the users' adoption level of a new technology, such as e-Government [40][41].

Cultural Influence can be defined as "the values, beliefs, norms and behavioural patterns of a group of people in a society for national culture, staff of an organisation for organisational culture, specific professions for professional" [42]. Cultural Influence has been widely investigated in the literature on the adoption of technologies, and it is proposed that cultural norms have a fundamental correlation with the intention to use e-Government systems [43] [44]. Many aspects of social and cultural influences need to be investigated and analysed in order to discover their impact on the adoption and usage of e-Government systems. The most important aspects that are suggested in this research can be explained in the following points:

- *Image* : This refers to the users' perceptions that using and adopting an e-Government system will afford them societal superiority. It is claimed in the literature that adoption of e-Government might reflect the adopter's familiarity with modern technologies, his or her proficiency in using computers and the Internet, a high degree of modernism and a higher level of education. These phenomena add a degree of prestige and social value to the adopter [20]. This research suggests that this phenomenon is even more influential in Saudi society.
- *Influence of others* : It is now widely accepted that "the most powerful influence on human behaviour is other people" [45]. The power of the social influence runs very deeply, since whom we know and how we feel about others affects our behaviour. In the Saudi context, the impact of others is even more influential, since the society in the Kingdom of Saudi Arabia is interrelated and coherent. Most of the relationships between family members, as well as the social relationships between different families and friends, are strong.
- *Resistance* : The adoption and utilisation of any new technology is greatly affected by the behavioural norms of the society, since individuals may tend to resist changes that technologies can make, which is likely to lead to negative consequences with regard to the implementation of such technology [46]. Resistance to change is considered one of the factors that can negatively affect the successful implementation of any system, including e-Government systems. Some users might resist using e-Government due to trust issues, whereas others might resist adopting e-Government because of the face-to-face culture that is still influential in terms of customer confidence [47].
- *The influence of interpersonal social networks (Connections or Wasta)* : Wasta represents one kind of corruption in many government organisations in developing countries [47]. In Arabic, wasta means personal connections with certain employees or top managers who can accelerate processes or break rules in order to complete paperwork and conclude transactions. Since e-Government will limit such corruption because services will be treated electronically and all actions will be controlled, some customers (including citizens and businesses) may resist the use of e-Government and continue to seek help from their personal connections instead. Moreover, some employees may also resist using the e-Government system in order to retain their ability to break the rules when processing their relatives' applications.

Awareness of the functions and services that any interactive system can provide for its users is a very important factor. Perceived awareness is a strong contributor to the adoption of e-Government [20]. Thus, all governments that intend to implement the e-Government system, especially in developing countries, need to be conscious of making users aware of and familiar with e-Government, particularly users in remote areas. Not doing so is likely to create a severe digital divide and fail to achieve the goals of e-Government. The research also assumed that awareness can be improved in a variety of ways, such as interactive advertising, social media and traditional advertising methods. Social media can play a fundamental role in enhancing the adoption and utilisation of e-Government by increasing awareness levels in this regard. Some recent research has begun to analyse the impact of social media on e-Government interaction [48] [49].

Functional Quality of Service refers to the quality level of functional aspects of a service that is provided by e-Government. Functional quality of service is a significant factor that must be studied from different aspects in order to achieve accurate and practical results. Delivery speed of and delivery options for required documents among government agencies and users are considered to be important aspects that can influence adoption and usage levels. Customer care is another important aspect of functional service quality and involves providing on-going technical support to customers, helping customers to perform e-Services, seeking customer feedback and ensuring their satisfaction. Other aspects of the functional quality of the services provided services are also likely to affect e-Government usage, such as providing different payment options for governmental transactions that involve fees, providing up-to-date information regarding e-Services and making e-Services available 24 hours a day, seven days a week.

Previous Experience is another important factor that has been addressed in this research model (EGAUM). It refers to experiences that users have encountered in the past with regard to e-Services. In this research, two aspects concerning the influence of previous experience have been proposed. The first is previous experience of using online applications and services such as e-Business services, e-Commerce services, online banking, online shopping and online payments. The second is related to previous experience of using e-Services provided by the e-Government itself. Positive and negative results of using online services, either e-Government services or any other nongovernmental services, are likely to have a strong impact on future interactions with the e-Government. Furthermore, this factor will also influence the adoption and utilisation of e-Government when first using it, as well as the continued use thereof. The Motivational Factors (MF), which include the Previous Experience factor, have indirect influence on all other factors. This will determine the influence of all the research factors on the first use and on the continued use of e-Government.

C. Technical Factors (TF):

In any interactive information system, there will be technical aspects that need to be addressed and taken into account in order to achieve the desired goals of implementing such systems. Unlike certain research in the literature that focused on the technical infrastructure aspects of the e-Government system itself, such as the network infrastructure and required applications [44] [50], this research will focus on the theoretical technical aspects that are related to the users' adoption and interaction. The technical factors that are addressed in this research model (EGAUM) can be listed as:

Simplicity refers to those factors that can simplify the e-Government system and make it easy to use. It has been established that the easier an information system is to use, a greater number of users will adopt it. This relationship is particularly applicable and important with regard to large systems such as e-Government, as large numbers of users are expected to utilise it, and their skills and abilities may differ significantly. Simplicity includes the following factors:

- *E-Government interface design:*
  E-Government's websites and interfaces are the interaction mediator between the e-Government system and its users. Therefore, they must be designed according to the users' requirements in order to achieve satisfactory levels of utilisation. E-Government interface design includes web page layout, design consistency, colour contrast, being free of spelling errors, having a clear font, clear labels, searchable contents and multiple language options.
- *Reaching e-Services:*
  This refers to how easy it is to reach to the intended e- Service on the e-Government website i.e. the ease of locating and navigating the intended e-Service after accessing the intended e-Government website. Difficulty in locating e-Services on e-Government websites is likely to have a negative effect on the adoption and utilisation of such e-Services, even if access to the e-Government websites is easy. Therefore, it is suggested in this research that if it is easier to reach e-Services, more e-Services will be performed; thus, the high adoption and utilisation thereof will be achieved.
- *Service description and hints:*
  This factor refers to the descriptions of services that are used to explain the e-Service to the users, as well as how to apply for it, the requirements thereof, the way in which the e-Service will be processed and how it will be delivered. Such information provides the users with a better understanding of the methods involved in successfully implementing e-Services. Furthermore, providing tips and examples for conducting e-Services and information regarding the required information type is also important in assisting users to utilise e-Government services.

Technical Quality of Service is another form of quality of service that needs to be addressed when analysing the adoption and usage of e-Government systems. It refers to the technical aspects that are visible to the users and which can affect their willingness and intention to adopt and use the system. These technical aspects reflect the quality of the system, and they include:

- *Suitability for people with special needs:*
  This refers to the e-Government system's ability to support features and technologies that can assist people with special needs to use and interact with the e-Government. Special needs include blindness, low vision, deafness, being hard of hearing and physical disabilities [51] [52].
- *Free of syntax, semantic and linkage errors:*
  This refers to ensuring that e-Government system is free from technical errors such as broken links, payment confirmation errors, and server and network errors. Such errors are likely to lead to user frustration and dissatisfaction, resulting in low levels of adoption and utilisation [26].

Accessibility Accessibility is another technical factor that refers to the users' ability to access the e-Government system, and includes the available methods for reaching and accessing services provided by the e-Government system. Several important aspects that are related to e-Government accessibility will be addressed in this research. These important aspects include:

- *The role of intermediaries in e-Government accessibility:* Intermediaries can be defined as any private or public organisation that facilitates communication and coordination between public services providers and users [53]. Many users will be excluded from benefiting from e-Government services due to certain access barriers, such as limited access to the Internet and computer illiteracy [54]. In this regard, Al-Madinah city, which is a large city in Saudi Arabia, has introduced so-called 'intermediary organisations' or 'e-offices', which operate under the legislation and authority of the Saudi government. They are physical intermediaries (organisations) that help people who have difficulty using e-Government. These difficulties include limited Internet access, computer illiteracy and the inability to pay online. Al-Sobhi et al. have studied the influence of the intermediaries, and the results show that intermediaries are an extremely useful channel for improving the adoption and utilisation of e-Government [24].
- *Mobile e-Government:* With the spread and popularity of mobile communication devices, e-Government has become more accessible and adoptable in many countries [55]. Mobile communication devices include mobile phones, tablets and PDAs (Personal Digital Assistants). E-Government services are often difficult for users in remote areas to access, as well as being problematic for users that are homebound, or who have poor computer skills or chronic illnesses [56]. Thus, m-Government can enhance access for users with these conditions easily and efficiently, which will have a positive effect on the adoption and usage level.
- *Digital divide:* The digital divide refers to the inequality between groups (individuals, households and businesses) in terms of access to and the use of Information and Communication Technologies (ICTs) [57]. Equality in terms of accessibility and the ability to use any information system is an important factor that can significantly affect the adoption and utilisation level. The importance of this factor is emphasised when the information system targets large populations and major users, such as e-Government systems.

D. Reliability Factors (RF)

The Reliability Factor (RF) construct is another important aspect of the EGAUM model that was developed in this research. This construct comprises fundamental factors that are related to the users' perceived risks and trust. These factors are:

Security and Privacy are two significant factors that need to be of a high standard in any interactive system. Their importance is emphasised when the system involves public users and sensitive data. In order to increase the adoption and usage level of information systems and applications, users need to feel safe when interacting with such systems. Implementing high standards of security and privacy is crucial, but it is not sufficient to increase the usage and utilisation level. Users need to be informed about the implementation of high security and privacy standards via the publication thereof, the sending of e-Services receipt confirmations, the requirement for further security criteria when performing financial transactions, and by informing the users when the e-Service requests the sharing of their personal information and requires permission to do so. Such factors are likely to influence perceptions of security and privacy, thereby increasing the users' adoption and utilisation of e-Government.

Trustworthiness plays a vital role in helping users to overcome the perceived risk and uncertainty involved in using online services. Trust issues can strongly affect the users' intention to share their personal information and to perform online transactions when using e-Government systems. Although trustworthiness has been studied and has proved to be an important factor in the literature on technology acceptance, there is still insufficient research that investigates and analyses the influence of trust on e-Government adoption and utilisation [30][58]. In this research, the trustworthiness factor will be investigated according to three aspects, namely trust in the Internet, trust in the e-Government system and trust in the e-Service provider. The importance of this factor is based on the fact that trust can be built from first impressions, and can also be affected by any later shortcomings. Moreover, trust is difficult to regain, particularly in an uncertain and virtual environment.

Regulations and Policies for using any interactive information system must be introduced and set up strictly and clearly in order to reach satisfactory levels of adoption and interaction. The importance of setting up clear, strict regulations is emphasised when the implementation of large public systems, such as the e-Government system, is involved. This factor includes several significant aspects, such as usage terms and conditions, e-Service delivery policies, payment policies, users' and providers' rights, data protection policies, and security and privacy policies. These examples of regulations must be introduced not only for the e-Service provider's records, but must also be published in order for the public users to be informed of their rights and to increase their confidence in the reliability of the system.

IV. Discussion

As stated previously, this research model was proposed based on a critical analysis of the common technology acceptance models that have been utilised in most of the recent e-Government studies. A number of significant factors in these common models were integrated into this research model, albeit with broader insight. For example, the Perceived Benefits

in EGAUM is similar to the concept of Perceived Usefulness in TAM and to Relative Advantages in DOI [15] [16], but has a wider interpretation. Moreover, the Simplicity factor is also more comprehensive than is the Ease of Use in TAM and the Effort Expectancy in UTAUT [15] [18].

On the other hand, several important factors in this research model were not addressed previously in most of the e-Government adoption and usage studies, especially those that were conducted in developing countries. This is because most of the e-Government adoption studies have utilised common technology acceptance models in the e-Government context that originally lacked many fundamental factors. These important factors have been added in this research model as independent factors, and include Cultural Influence, Personal Factors Influence, Awareness, Previous Experience Influence, Functional and Technical Quality of Service, Security and Privacy, Regulations and Policies, and Trustworthiness.

We believe that this research model has proposed and developed a simplified and comprehensive manner for providing a scientific framework for the analysis of e-Government interaction. Moreover, EGAUM is a universal model, which means that it can be utilised in countries other than Saudi Arabia. It can also be adapted and used to analyse the adoption and utilisation of different interactive systems and various service applications because it addresses influential factors that are crucial in such an analysis. EGAUM is currently being tested and validated to explore the influence of its constructs on e-Government adoption and utilisation in Saudi Arabia.

V. Conclusion

In order to investigate and analyse the key factors that can influence users' adoption and utilisation of e-Government services in Saudi Arabia, this research proposed a comprehensive and conceptual model (EGAUM) that would be the basis of our future work. A number of significant factors that have been integrated into this research model were taken from well-studied technology acceptance models in the existing literature. Other crucial factors have been developed and added in this research in order to conclude a model for e-Government adoption and utilisation analysis that is more comprehensive and appropriate.

The EGAUM model is now being validated and tested on various user groups, including public users of e-Government services, office workers that provide e-Government services and business managers who use e-Government services as part of their business activities, in order to determine the direct and indirect influences of the model's constructs on actual e-Government adoption and utilisation.

References

[1] R. Davidrajuh, *Planning e-government start-up: a case study on e-Sri Lanka*. Electronic Government: An International Journal, Volume 1, No 1, pp. 92-106, 2004.

[2] S. Alshafi and V. Weerakkody, *Factors affecting e-government adoption in the state of Qatar*. Proceedings of the European and Mediterranean Conference on Information Systems, Abu Dhabi, UAE, 2010.

[3] S. Archmann, and J. Castillo Iglesias, *eGovernment: A Driving Force for Innovation and Efficiency in Public Administration*. EIPAScope, 2010 (1). pp. 29-36.

[4] T. Gebba, and M.R. Zakaria, *E-Government in Egypt: An Analysis of Practices and Challenges*. International Journal of Technology and Management. Vol. 1 No. 1, pp. 11-25, 2012.

[5] Y. Kitaw, *E-Government in Africa Prospects, Challenges and Practices*. Swiss Federal Institute of Technology, 2006.

[6] U.N., Benchmarking E-government: A Global Perspective. United Nation Division for Public Economics and Public Administration - American Society for Public Administration: USA, 2001.

[7] S. Alateyah, R. Crowder, and G. Wills, *Citizen Adoption of E-government services*. International Conference on Information Society. University of Southampton, United Kingdom, 2012.

[8] A. Grönlund, *INTRODUCING e-GOV: HISTORY, DEFINITIONS, AND ISSUES*. Communications of the Association for Information Systems, Volume 15, 2004.

[9] E. Ziemba, T. Papaj, and R. Zelazny, *A Model of Success Factors for E-Government Adoption - The Case of Poland*. Issues in Information Systems, 14(2), pp. 87-100, 2013.

[10] Z. Ebrahim, and Z. Irani, *E-government adoption: architecture and barriers*. Business Process Management Journal, Vol. 11 Iss: 5, pp.589-611, 2005. DOI: 10.1108/14637150510619902

[11] M. Alshehri, and S. Drew, *Challenges of e-Government Services Adoption in Saudi Arabia from an e-Ready Citizen Perspective*. World Academy of Science, Engineering and Technology, 2005.

[12] Z. Al-adawi, S. Yousafzai, and J. Pallister, *Conceptual model of citizen adoption of E-government*. The Second International Conference on Innovations in Information Technology, 2005.

[13] S. Sahraoui, *E-government in the Arabian Gulf: government transformation VS. government automation*. eGovernment Workshop, Brunel University, West London, 2005.

[14] I. Ajzen, and M. Fishbein, *Belief, attitude, intention, and behaviour: An introduction to theory and research*. Reading, MA: Addison-Wesley, 1975.

[15] D. Davis, P. Bagozzi, and R. Warshaw, *User Acceptance of Computer Technology: A Comparison of Two Theoretical Models*. Management Science, Vol. 35, No. 8, 1989.

[16] E. Rogers, *Diffusion of Innovations*. The Free Press, New York, USA, 1995.

[17] G. Moore, and I. Benbasat, *Development of an instrument to measure the perceptions of adopting an information technology innovation*. Information Systems Research, Vol. 2 Issue 3, pp192, 1991. DOI: 10.1287/isre.2.3.192

[18] V. Venkatesh, M. Morris, G. Davis, and F. Davis, *User Acceptance of Information Technology: Toward a Unified View*. Management Information Systems Quarterly, Vol. 27, Issue.3, 2003.

[19] S. Sang, and J.D. Lee, *A Conceptual Model of e-Government Acceptance in Public Sector*. Third International Conference on Digital Society, Washington, DC, USA, pp. 71-76, 2009.

[20] M. Shareef, V. Kumar, U. Kumar, and Y.K. Dwivedi, *e-Government Adoption Model (GAM): Differing service maturity levels*. Government Information Quarterly. 28, 17 - 35, 2011. DOI: 10.1016/j.giq.2010.05.006

[21] E.M. Van Raaij, and J.J.L. Schepers, *The acceptance and use of a virtual learning environment in China*. Computers & Education, vol. 50, no. 3, pp. 838-852, 2008. DOI: 10.1016/j.compedu.2006.09.001

[22] M. Alshehri, S. Drew, and R. Alghamdi, *Analysis of Citizens' Acceptance for e-Government Services: Applying The UTAUT Model*. IADIS International Conferences, Theory and Practice in Modern Computing and Internet Applications and Research, pp. 69-76, 2012.

[23] A. Taiwo, A.K. Mahmood, and A.G. Downe, *User Acceptance of eGOVERNMENT: Integrating Risk and Trust Dimensions with UTAUT Model*. International Conference on Computer & Information Science (ICCIS), 2012.

[24] V. Weerakkody, R. El-Haddadeh, F. Al-Sobhi, M.M. Shareef, and Y.K. Dwivedi, *Examining the influence of intermediaries in facilitating e-government adoption: An empirical investigation*. International Journal of Information Management, 33, 716-725, 2013. DOI: 10.1016/j.ijinfomgt.2013.05.001

[25] Yesser.gov.sa, E-Government program overview. [Online] Available from: http://www.yesser.gov.sa/en/ProgramDefinition/Pages/Overview.aspx (accessed in 20/01/14).

[26] H. Al-Nuaim, *An Evaluation Framework for Saudi E-Government*. Journal of e- Government Studies and Best Practices, IBIMA. Vol. 2011, 2011.
[27] KH. Al-Sabti, *The Saudi Government In the Information Society*, 2005. [Online] Available from: http://www.yesser.gov.sa/ar/mediacenter/DocLib1/dubaiegov.pdf (accessed in 03/02/14).
[28] Yesser, The National e-Government Strategy and Action Plan, 2006. [Online] Available from: http://www.yesser.gov.sa/en/MechanismsandRegulations/strategy/Pages/implementation_plan_first.aspx (accessed in 25/10/13).
[29] M. Al-Suwail, *E-government Program*. Proceedings of the National e-Transactions Conference, Al-Riyadh, Saudi Arabia, 2007.
[30] H. Alsaghier, M. Ford, A. Nguyen, and R. Hexel, *Conceptualising Citizen's Trust in e-Government: Application of Q Methodology*. Electronic Journal of e-Government, Academic Conferences. Vol. 7, pp. 295-310, 2009.
[31] C.E. Hill, K.D. Loch, D.W. Straub, and K. El-Sheshai, *A qualitative assessment of Arab culture and information technology transfer*. Journal of Global Information Management, Vol. 6 No. 3, pp. 29-38, 1998.
[32] E. Baker, S. Al- Gahtani, and G. Hubona, *The effect of gender and age on new technology implementation in a developing country*. Information Technology and People, 20, 352-375, 2007. DOI: 10.1108/09593840710839798
[33] G, Sciadas, *The Digital Divide in Canada*. Ottawa: Statistics Canada, 2002.
[34] M. Al-Otaibi, and R. Al-Zahrani, *Electronic commerce in the Kingdom of Saudi Arabia*. King Saud University, Riyadh, 2009.
[35] K. Alrawi, and K. Sabry , *E-commerce evolution: a Gulf region review*. International Journal of Business Information Systems, 4, 509-526, 2009. DOI: 10.1504/IJBIS.2009.025204
[36] F. Al-Sobhi, V. Weerakkody, and M.M. Kamal, *An exploratory study on the role of intermediaries in delivering public services in Madinah City: Case of Saudi Arabia*. Transforming Government: People, Process and Policy, vol. 4, pp. 14-36, 2010.
[37] M.K. Alomari, P. Woods, and K. Sandhu, *Predictors for E-government Adoption in Jordan: Deployment of an Empirical Evaluation Based on a Citizen-centric Approach*. Information Technology & People, vol. 25, pp. 4-4, 2012. DOI: 10.1108/09593841211232712
[38] L.L. Tung, and O. Rieck, *Adoption of electronic government services among business organisations in Singapore*. Journal of Strategic Information Systems, 14, 417-440, 2005. DOI: 10.1016/j.jsis.2005.06.001
[39] L. Carter, and F. Belanger, *The utilization of e-Government services: Citizen trust, innovation and acceptance factors*. Information Systems Journal, 15(1), 5-25, 2005. DOI: 10.1111/j.1365-2575.2005.00183.x
[40] Z. Irani, Y.K. Dwivedi, and M.D. Williams, *Understanding consumer adoption of broadband: An extension of the technology acceptance model*. Journal of the Operational Research Society, 60, 1322-1334, 2009. DOI: 10.1057/jors.2008.100
[41] V. Venkatesh, and S. Brown, *A longitudinal investigation of personal computers in homes: Adoption determinants and emerging challenges*. MIS Quarterly,25(1), 71-102, 2001. DOI: 10.2307/3250959
[42] K. Leung, R.S. Bhagat, N.R. Buchan, M. Erez, and C.B. Gibson, *Culture and international business: recent advances and their implications for future research*. Journal of International Business Studies, 36(4), pp. 357-378, 2005
[43] C. Akkaya, P. Wolf, and H. Krcmar, *Factors Influencing Citizen Adoption of E-Government Services: A cross-cultural comparison*. 45th Hawaii International Conference on System Sciences, 2012.
[44] S.A. Alateyah, R.M. Crowder, and G.B. Wills, *Factors Affecting the Citizen's Intention to adopt e-Government in Saudi Arabia*. World Academy of Science, Engineering and Technology Vol:81, 2013.
[45] D. Halper, *Successful public service design must focus on human behaviour*, The Guardian newspaper, 2013. [online] Available from: http://www.theguardian.com/public-leaders-network/2013/jan/17/public-service-design-human-behaviour (accessed in 09/11/13).
[46] R.T. Watson, T.H. Ho, and K.S. Raman, *Culture: A fourth dimension of group support systems*. Communications of the ACM, 37(10): pp. 55, 1994. DOI: 10.1145/194313.194320
[47] S. AlAwadhi, and A. Morris, *Factors Influencing the Adoption of E-government Services*. Journal of Software, vol. 4, no. 6, 2009.
[48] H.M. Abdelsalam, C.G. Reddick, S. Gamal, and A. Al-shaar, *Social media in Egyptian government websites: Presence, usage, and effectiveness*. Government Information Quarterly, 2013.
[49] L. Zheng, *Social media in Chinese government: Drivers, challenges and capabilities*. Government Information Quarterly, 2013.
[50] F.H. Chanchary, and S. Islam, *E-government Based on Cloud Computing with Rational Inference Agent*. High Capacity Optical Networks and Enabling Technologies (HONET), pp.261-266, 2011.
[51] A. Abanumy, A. Al-Badi, and P. Mayhew, *e-Government Website Accessibility: In-Depth Evaluation of Saudi Arabia and Oman*. The Electronic Journal of e-Government Volume 3 Issue 3 pp 99-106, 2005.
[52] Alliance for Technology Access, *Computer Resources for People with Disabilities: A Guide to Exploring Today's Assistive Technology*. 2nd edition, Hunter House, 1996.
[53] M. Janssen, and B. Klievink, *The role of intermediaries in the multi-channel services delivery strategies*. International Journal Of Electronic Government Research,5(3), 36-46, 2009. DOI: 10.4018/jegr.2009070103
[54] H. Margetts, and P. Dunleavy, *Cultural barriers to e-government*. Academic article for the report: 'Better Public Services Through e-government'. London: National Audit Office, 2002.
[55] S.Y. Hung, C.M. Chang, and S.R. Kuo, *User acceptance of mobile e-government services: An empirical study*. Government Information Quarterly, 30, pp. 33-44, 2013. DOI: 10.1016/j.giq.2012.07.008
[56] S.A. Becker, *Bridging Literacy, Language, and Cultural Divides to Promote Universal Usability of E-Government Websites*. Northern Arizona University, 2009.
[57] M.D. Chinn, and R.W. Fairlie, *The Determinants of the Global Digital Divide: A Cross-Country Analysis of Computer and Internet Penetration*. Economic Growth Centre, Yale University, USA, Paper No. 881, 2004.
[58] P.A. Pavlou, and M. Fygenson, *Understanding and predicting electronic commerce adoption: An extension of the theory of planned behaviour*. MIS Quarterly,30(1), 115-143, 2006.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1227–1234
DOI: 10.15439/2014F198
ACSIS, Vol. 2

The Application Of A Conversion Method In A Confrontational Pattern-Based Design Method Used For The Evaluation Of It Systems

Witold Chmielarz
University of Warsaw,
Faculty of Management,
ul. Szturmowa 1/3, 02-678 Warszawa, Poland
Email: witek@wz.uw.edu.pl

Marek Zborowski
University of Warsaw,
Faculty of Management,
ul. Szturmowa 1/3, 02-678 Warszawa, Poland
Email: mzborowski@wz.uw.edu.pl

***Abstract*—The objective of this paper is to examine application possibilities of a conversion method in a new confrontational pattern-based design method used for IT systems evaluation. First, the authors present characteristic features and assumptions of the new methodology. Next, they conducted a study verifying the application of this methodology in the analyses preceding the execution of the project aiming at creating a new comparison engine, based on the model which received the highest scores in the evaluation. The results of the study were analysed, and the authors examined their usefulness with regard to the application of the presented method.**

I. Introduction

IN THIS article the authors present an application of the conversion method for the evaluation of websites. The conversion method may be considered as an extension of the novel methodology to design IT systems. The method is based on the averaging of distances from the average results of the surveys concerning the quality of IT systems [4]. This article shows that the application of the conversion method to the confrontational pattern-design method is possible. It refers both to the activities of project teams and on the other the modern methods of design solutions for design through the service (Service Desig) [10]. The main recommendations of this type are methods [3, 6, 12, 13]: focus on the needs of the user; full cooperation of the parties during the project, to present the full implementation of the service over time; formalization and a peculiar "materialization" of all elements of the service in a way comprehensible to the user and comprehensive account of the trial of the service by a single contractor.

They have one fundamental flaw: they do not provide models, procedures and practices which are the required contribution to the theoretical knowledge. Thus, they are not embedded in a broader economic or praxeological context. Therefore, they cannot be used for the purpose of the optimization of investment in IT (especially services), formulation of new service projects (based on the added value for the customer global and multicultural), and the implementation of projects that allows these project to be immersed in the realm of operation systems as much as possible. This is particularly important for electronic commerce tools.

The integration of these solutions with basic assumptions of modern management methods and some of the best proven traditional solutions can bring, as it seems, very good results with regard to creating design patterns. This conclusion is supported with a number of experiments conducted by the authors during the research into the assessment of e-business [2] and the possibility to use the findings in system designing. In the study the Authors used an iterative approach to identify ex ante needs of the user and confront them with the ex post experiences resulting from a deep analysis of the existing IT solutions of Confrontational Pattern-Based Design Method.

The basic assumptions and recommendations *Confrontational Pattern-Based Design Method* refer to the concept of *Service Design*, on the one hand, and *Agile Design* methods on the other [5]. They are as follows:

- in many cases user's requirements, even in the case of an informed client, are dominated by existing habits connected with the IT systems used in the organization: in the clients' opinion the questions concerning additional functionalities introduce elements of ambiguity, or even contradiction,
- project schedule becomes the result of negotiations between the user, expecting to reduce the time of the realisation of the project, or even be provided with the finished product immediately, and the possibilities of the contractor and his desire to offer a product which is of higher quality than the existing solutions,
- iteration between the initial recognition of the user's needs (even if they tend to be reduced to the experience of the previous installation), and best practices derived from the analysis of the existing solutions on the market; each successive iteration is a compromise bringing us closer to the final solution,
- there are methods of identifying best practices, consistent with the scientific basis of the evaluation of IT systems (and common sense); we should aim at constant

This work was not supported by any organization

978-83-60810-58-3/$25.00

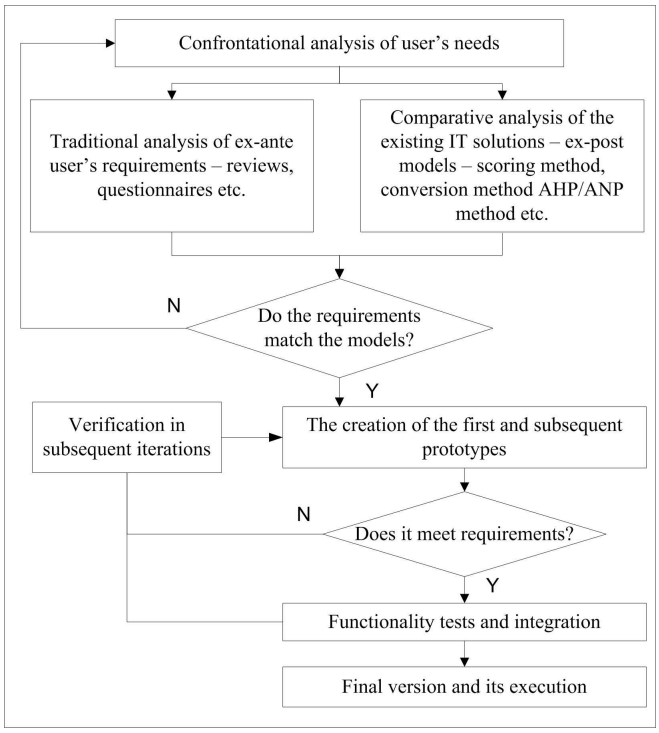


Fig. 1. Basic phases of the life cycle of the project in CPD model Source: on the base of [5]

improvement or creation of an IT system which allows for automatic selection of the method best suited to a particular decision-making situation,

- the user shows a specific tendency to overcomplicate systems, thus, we should aim at their greatest simplicity at the level of design, content, service and methods of solving basic design problems,
- parallel analysis of the user's requirements — on the one hand, the application of traditional analytical methodologies (questionnaires, interviews, conversations, ...), on the other, consideration of the user's evaluation of the existing systems in order to extract certain components which can meet his needs in the best possible way,
- coordination of the language describing design requirements with the assessment criteria of existing solutions,
- ongoing integration of selected components in accordance with previous arrangements in the present version of the project,
- solving, by way of negotiation, confrontational requirements resulting from the projection of the user's requirements, performance and usage analyses as well as expert requirements resulting from the best practices of completed projects in the same area.

The life cycle of the project according to CPD model (see: Figure 1), consists of four basic phases: confrontational analysis of the user's needs — parallel examination of the needs of the final user and of the existing IT solutions in the field. Based on the list of the best solutions, we create a project based on the optimal design patterns used in the existing portals. In the case of differences — negotiations with users which aim at bringing their position closer to the position resulting from the analysis of best practices; on this basis we create a prototype of an IT system — subsequently, it is presented to the client; in the case it does not meet the user's expectations — introduction of changes, creation of another prototype — presentation of the prototype to the user; in the case of fulfilling requirements — testing stage; functionality tests — subsequent versions presented to the client are tested before next modifications, which client may suggest at this stage. In case of doubts, we return to the creation of the next prototype and working on the final version of the project; the last iteration leads to the creation of the last, complete version of the project, which is then executed.

Methodology of the Confrontational Pattern-Based Design Method concerns mainly small and medium-sized e-commerce projects. It assumes full access to the existing software, from the user's perspective, and such a situation is taking place in the Internet. Apart from the classical analysis of the user's requirements, the practical approach consists in the examination of the existing solutions enabling: specification and accurate research into the area in which the software works, creating a ranking of IT solutions existing on the market, identification of the features which make particular solutions better than others.

However, there are concerns regarding: the creation of a coherent methodology which needs to be applied in the examination of the needs of a user of existing IT solutions, interpretations of findings of practical analyses, the necessity of taking into account the high dynamics of technological innovation in this field, the necessity of developing the mechanism of negotiations of the proposed solutions with the end user.

II. Assumptions of the Study

The issue of creating a project of a comparison site may be used as an example of the application of this method. The findings of a research company Tradedoubler [9] shows that Polish clients use comparison sites more and more frequently, anticipating constant growth of e-commerce (Poland and Czech Republic noted over 30% increase of internet sales in 2012, while the average in Europe is 22% [7]). The determinants of the success seem to be: the novelty effect, limited financial resources, which consumers may spend on shopping, promotional activity of comparison sites, the growing awareness of the opportunities posed by internet shopping, and the price, which is a key criterion in the case of buying branded goods. The destimulators are mainly phenomena such as:

- not all shops in Poland, especially new ones, present their offers on comparison sites,
- offers are not continuously updated,
- errors concerning the presented product price,
- there is no active verification of the prices provided by the shops [5].

The most popular software of this class on the market are: Ceneo.pl, Skapiec.pl, Nokaut.pl and Okazje.info.pl. The client verifies the price, brand of a product, reputation of a shop, client's opinions, and subsequently, based on the available data, he or she chooses the most attractive offer. At present, comparison sites are widely known and more commonly used — therefore, they prove to be very useful in conducting the mass research into the needs of the so-called average user of the Internet.

The assumptions of the study were as follows:

- the objective of the study is to design an online comparison service so that it meets the users' needs in a way which is optimal in terms of the qualitative characteristics: functional, technological, etc.,
- we should analyse users' requirements concerning the basic usability parameters of the portal. The study's objective is to indicate the parameters which are most important from the point of view of the websites' clients (considering the class of electronic agents),
- based on the above findings, we create a list of the most important evaluation criteria of the existing comparison websites,
- the list of criteria is used for individual evaluation of comparison websites known to users (one user can evaluate more than one website),
- the group of respondents is not chosen at random, it belongs to the class of convenient samples as the respondents are students of the selected universities in Warsaw (Faculty of Management of the University of Warsaw and Faculty of Information Technology of the Academy of Finance and Business Vistula); they represent all types of studies (B.A., B.Sc., M.A., PhD studies),
- the evaluation will be standardized according to the ten-point scale used to examine other types of IT systems,
- the results of the analysis of user requirements will be confronted with the evaluation of the characteristics of the existing comparison websites in order to identify the best possible implementation of those characteristics which are most important from the user's point of view,
- the analysis will be performed first by a scoring method, and, next by conversion method for minimization of subjectivity and thus their realignment,
- multiplying the preference scale by the table of results of the scoring analysis will allow for the realignment of the findings with the application of relations among particular criteria,
- the realignment will result in finding the design patterns of comparison websites which are considered best from the point of view of management, and which may be used in the creation of the design of the website of this class.

Initially, among the randomly selected five students of the Faculty of Management of the University of Warsaw and Academy of Finance and Business Vistula, declaring the use of comparison websites for making purchases, the authors carried out a pilot study concerning the factors which are most important when using software of this class. After the users' suggestions and as a result of standardization (unification of terminology and concepts of particular categories proposed by them), the following groups of characteristic features have been identified: design (visualization), information presented on the site and the ease of navigation, text search, data operations: selection (filtering) and sorting the results according to selected criteria and additional functionalities used for increasing user-friendliness.

Conclusions from the observation of students' behaviour supplied the initially established sets of features with the detailed criteria and subcriteria, which were thoroughly discussed with the respondents before placing them on the list used in the evaluation. The final list of criteria included:

- visualisation, information available on the website and the ease of navigation: brand awareness (the ease of memorizing data and reasons why we like the website), main page (readability and clarity of the main page, the ease of navigation and finding particular functions), the consistency of graphic elements, clearly marked colours, matching colours of elements, background colouring, icons symbolizing categories of products, colours and clarity of the text, carefully selected pictures of good quality, etc.), photo gallery (large, readable and clear images, which do not obscure navigation and present the actual features of the product etc.), completeness of information (characteristic features, minimum/maximum price, product image), clarity (appropriate font size, distinct colouring, the appropriate distribution of elements on the website), avoiding distracting elements (too much advertising, excessive number of photographs, etc.), opinions about shops (logo, delivery time, minimum price, clients' opinions), product reviews (if they are available, in what form, whether they meet clients' requirements), comparison engines (operating according to features specified by a client — e.g. price, number, delivery time), the ease of using categories (availability: lists, subcategories, characteristic features, etc.), the form of presentation of the product list (lists, different views, presentation according to the product features, icons, photos), rankings of products/shops (position in the ranking of shops/product categories), suggesting products (prompting — the latest model, at a similar price, other customers also bought ..., etc.);
- functionality of the text search (dynamic search of the selected product, without the necessity of looking through many categories): the accuracy of search results (the name of the product, the name of the product + producer, etc. combinations), prompting (lists displayed during the search, text field with autocomplete: prompting the name in the early selection stage, etc.), spellchecker (automatic correction of spelling);
- filtering and sorting results (simple and easy selection of products): filtering functionality (a large number of criteria, speed, limitations on operations which may be

performed by a client; the possibility of creating complex selection rules, etc.), sorting functionality in the product list (criteria of ordering and its combinations), sorting functionality in the list of shops (map, city, list according to prices);
- additional functionalities (functions improving usability) for the user (individually for websites): ordering at the level of a comparison site, memorizing a list of products, creating sets of products, loyalty program, the possibility of using mobile applications, price alert, tracking the changes concerning your favourite products (prices, characteristics).

In November 2013 the authors conducted surveys in the selected universities. Over 110 people filled the survey. Among the survey participants: women constituted 2/3, 1/3 of respondents were men. Most people, 80%, were in the age group of 18-25 — typical for students of full-time studies and 15% from the age group of 26-35 — characteristic of students of part-time studies. The participation of people over 35 years of age was small - 5%. 58% of the respondents were representatives of the cities with over 500,000 inhabitants, 22% of towns with 10-100,000 inhabitants, only 6% were from rural areas. Over 60% declared having secondary education, 28% were students of BA studies and 7% of MA studies. Three-quarters of the group declared the status of a student, 15% are employed in the private sector, 6% in the public sector and 4% are self-employed. Over a third of the sample belongs to the income group with over PLN 4,000 per month, 28% 2,000-2,500, 23% to PLN 1,500, and 12% to the remaining groups.

III. Analysis results by scoring method and their implmentation in the design

The first part of the survey available in the Internet and distributed in its traditional form was used to verify the importance and relevance of the list of criteria established in the interviews and direct discussions (information analysis) with clients of comparison websites (students) in a pilot survey. For all respondents, all the criteria connected with visualization, information contained on the website and the ease of navigation turned out to be the most important factors (34.2%). The second place was taken by the functionalities of the text search (33.8%); the last position (32.0%) was taken by the functionality of filtering and sorting results and additional functions improving usability. The differences in those groups were not very significant (up to 2 percentage points). For particular selected criteria, the differentiation did not appear to exceed (to a great extent) the observed results. The difference between the highest and the lowest scores amounted to 2.1%.

Relatively highest scores were assigned to the accuracy of search results — 5.68% and the completeness of information — 5.58%; the lowest were given to: suggesting products — 3.58% and additional functionalities — 3.83%. In total, this indicates the proper use of this tool (compatible with the objectives behind its creation) and its use in an elementary, rather than extended, range. In the case of website comparison websites, the evaluation of visualization criteria turned out to be very high — the graphic designs of home page obtained the score of 5.56%.

In the evaluation of the usability of criteria, 40% of responses were very high scores and about 31% of responses claimed that the criteria selected for the study were good: this means that almost 3/4 of criteria specified in the pilot survey are regarded as accurate. On average, only 3% of the respondents considered the criteria as poorly matched to the assessment of comparison websites. Small differences in the average scores do not induce the authors to reject any of the criteria. A similar situation occurs when we refer the average values of the obtained scores to the maximum possible score in the evaluation of comparison websites. In this way we obtained a list of functionalities of a comparison website which, from the users' point of view, best suits their expectations with regard to this type of service. Its importance has been verified by means of a survey conducted among the clients of selected comparison websites. All listed elements obtained more than 50% of the maximum possible value, so they may be applied in the ex post analysis of the existing comparison websites.

In the second part of the study, the authors conducted an examination of the comparison websites according to the previously adopted criteria. The examination analysed the four selected comparison websites which are the most popular among clients and the fifth one, which was selected individually by respondents. The first three websites: Ceneo, Skapiec and Nokaut have obtained a total score above the average; okazje.pl and individually chosen websites (different from the most popular ones (Cenuj.pl, Webkupiec.pl etc.)), are positioned below the average. Generally, there emerges one regularity which may be applied when establishing analytical characteristics for the project — it indicates which of the existing websites of this class can be used as a basic pattern to be imitated when creating a new website. Generally, we may conclude that Ceneo.pl will serve this purpose in the best way. However, the detailed analysis of the results is not so clear or univocal. In terms of the consistency of graphics, loyalty scheme and tracking the changes in the favourite products, the leading position is taken by skapiec.pl, in case of price alerts — Okazje.pl. Skapiec.pl, which takes a second position in the ranking, does not offer price alerts which are of considerable importance to clients. Okazje.pl, which takes the lowest place in the ranking, occupies the second position when we take into account the evaluation of the homepage and creating sets and a third place with regard to the photo gallery it presents on its website.

We can also consider individual sites taking into account the features which are evaluated as the best and the worst (see Table 1). In the most popular websites there occurs a puzzling consistency of the best and the worst scores. In other services some of the characteristics are beyond the scope of the list. Hence, we may assume that, basically, the most important elements of the comparison websites are: brand awareness, home page, consistency of graphic elements and the accuracy of search results. However, whenever a particular website feature receives the lowest scores, it points to the fact that

potential customers attach great importance to this particular element. Therefore, avoiding distractions, lack of spellchecker (or inappropriate corrections) and lack or limited number of additional functionalities should be added to the list. Also, we must appreciate the three remaining, important elements: completeness of information, the ease of using categories and filtering functionality. Thus, the design of the optimal comparison website should take into consideration at least the set of criteria presented in Table 1.

We may also analyse the results of the examination considering the number of scores in particular categories. Ceneo.pl was a leader in the ranking because it obtained the greatest number of very good scores. Also, the greatest number of good scores was assigned to Skapiec.pl, which came second in the ranking. The comparison website: Okazje.pl obtained the largest number of satisfactory or poor grades, other comparison websites — unsatisfactory. Nokaut.pl does not distinguish itself in any category. This table shows that it would be best to refer to the design patterns of ceneo.pl and Skapiec.pl when creating a new comparison website.

Apart from the possibility of limiting the number of basic features important for the creation of a prototype — which results from the studies of the existing websites in this category, the presented study shows a high discrepancy between the initial, average results of the clients' opinions concerning comparison websites and their characteristic features and average scores obtained from the ranking of the most popular websites existing in the Internet (see Table 3). Based on the initial evaluations, the first three positions are taken by: accuracy of search results, completeness of information, and the design of the homepage. When we take into account the evaluations of website analyses, the most important factors are: homepage, views of the product list, consistency of graphic elements.

The same situation occurs in the case of other features of comparison websites. However, if we consider not the values of the total of evaluations, but average values from initial analysis and the ranking of websites, then the importance of the features resulting from evaluations from initial analyses, more or less coincides with the importance of the features shown by the ranking of websites. What is the reason of such great variances between the opinions concerning the usefulness and importance of particular criterion for the evaluation of the comparison website, and low scores of these features in the analysis of particular comparison websites? It seems that the situation is caused by the users' awareness of high requirements which it should meet and, on the other hand, dissatisfaction with the implementation shown on the websites of existing comparison engines. It is also a clear indication for the system designer — we should not use the ready-made patterns in the cases where we observe high discrepancy between the user's expectations and the importance of the feature and its fulfilment. We should focus on these elements so that they are presented in the form which would meet the users' expectations.

IV. The Results of the Analyses of Comparison Engines by Means of a Conversion Method and the Application of the Findings in Project Works

The analysis by means of a scoring method (considered only from the point of view of a scoring analysis) faced the objections of far-reaching subjectivity [5]. The application of conversion method is an attempt to eliminate this subjectivity — the author's applied his own conversion method [14].

Here, we adopt the following assumptions: after constructing the experts' table of evaluations of particular criteria for each website, we need to perform the conversion with the established preference vector of the superior level criteria [2]. Next, the authors performs the transformation of the combined scoring table into the preference vector (first converter).

The next steps are:

- constructing a matrix of distances from the maximum value for each criterion in every website, establishing the maximum value; establishing the matrix of the distances from the maximum value, calculating the average distance from the maximum value for each criterion,
- as a result of the above operation, constructing a matrix of differences in the distance from the maximum value and the average distance according to criteria,
- for each bank website: constructing conversion matrices - modules of relative distances of particular criteria to remaining criteria (the distance from the same criterion is 0), the obtained distances below the diagonal are the converse of the values over the diagonal,
- averaging criteria conversion matrices — creating one matrix of average modules of values for all criteria,
- transforming the conversion matrix of criteria into a superior preference matrix (calculating squared matrix, adding up rows, standardization of the obtained preference vector; repeated squaring, adding up rows, standardization of preference vector — repeating this iteration until there are minimum differences in subsequent preference vectors).

As a result of the above operations we establish a criteria conversion matrix.

Subsequently, the author performed a transformation of the scores presented by experts on the level of a matrix specifying expert websites' evaluations for particular criteria (second converter) [14]. The results have been obtained in an analogical way:

- constructing a matrix of distances from the maximum value for each criterion and each website establishing the maximum value; establishing the matrix of distances from the maximum value,
- calculating the average distance from the maximum value for each website,
- constructing a matrix of the differences of deviations from the maximum value and the average distance of the features from the maximum,
- for each criterion: constructing a matrix of transformations (conversions) of the differences of the average distance from the maximum value between the websites,

TABLE I
THE MAP OF THE WEBSITE CRITERIA WITH THE HIGHEST AND THE LOWEST SCORES (WHERE: G - BEST-RATED FEATURES, B - WORST-RATED FEATURES)

Description/Website/Evaluation	Ceneo	Nokaut	Skapiec	Okazje	Other
Brand awareness	G	G	G		B
Homepage	G			G	G
Consistency of graphic elements		G	G	G	
Photo gallery				G	
Completeness of information					G
Avoiding distractions	B	B	B	B	
Product reviews				B	
Product comparison engines				B	
The ease of using categories					G
Accuracy of search results	G	G	G		B
Spellchecker	B	B	B		
Filtering functionality					B
Additional functionalities	B	B	B		

TABLE II
THE PERCENTAGE ASSESSMENT OF PARTICULAR COMPARISON WEBSITES

Evaluation/Comparison website	Ceneo.pl	Nokaut.pl	Skapiec.pl	Okazje.pl	Other
Unsatisfactory	3.50%	4.00%	3.45%	10.65%	**45,68%**
Poor	8.94%	18.55%	15.36%	**23.49%**	19.14%
Satisfactory	27.66%	32.55%	27.12%	**33.66%**	19.14%
Good	40.95%	33.45%	**41.38%**	24.70%	9.26%
Very good	**18.96%**	11.45%	12.70%	7.51%	6.79%

analogically as presented above (the distance for a particular feature in the same website from the same website is 0), values below the diagonal are the converse of the values over the diagonal,
- constructing a module matrix of transformations of the differences of average distance from the maximum value between the websites, for each criterion,
- for each module matrix of transformation of the differences of the average distance from the maximum value between the websites, squaring it, adding up rows, standardization of the obtained ranking vector and repeating this operation until the obtained differences between two ranking vectors for each criterion will be minimal.

As a result of the above presented operations we obtain a conversion matrix of websites' evaluations: using the obtained vectors to construct a combined ranking matrix

- heading names of bank websites by appropriate transfer of the obtained preference vectors for each criterion,
- multiplying the matrix obtained in such a way by the previously calculated preference vector,
- analysing final results and drawing conclusions (Note: the lowest distances in this case are the most favourable, comparability adjustments to other methods can be obtained by subtracting these values from 1 and their repeated standardization).

The basis for the creation of the presented method was the assumption that it should be easy to apply. The objective has been reached, which is visible in the number of the advantages presented below. The only disadvantage of the method is the fact that the transformation of the results of the survey is connected with carrying out many complex operations.

The advantages of this method are:

- the ease of application (similar to the realization of a scoring method) which results from the fact that in the survey form there are questions concerning the subjective evaluation of the element,
- in the case of considering a large number of evaluation criteria or alternatives there is no significant increase in the number of questions in the survey,
- the possibility of the application of the method with the participation of people who are not experts in a given field,
- there are no measures, as in the case of e.g. ELECTRE method — veto threshold, which may not be fully understandable for the respondent [14],
- the result of the calculations which takes the form of the importance of the evaluations of the examined objects.

Similarly to previous cases, the application of a conversion method resulted in flattening of the obtained results. It reduced not only the differences between particular comparison engines but also the differences among particular evaluation criteria. As a result, the obtained findings minimized the subjectivity of the evaluation, and additionally — as it turned out — brought the results closer to the findings of the preliminary analysis of

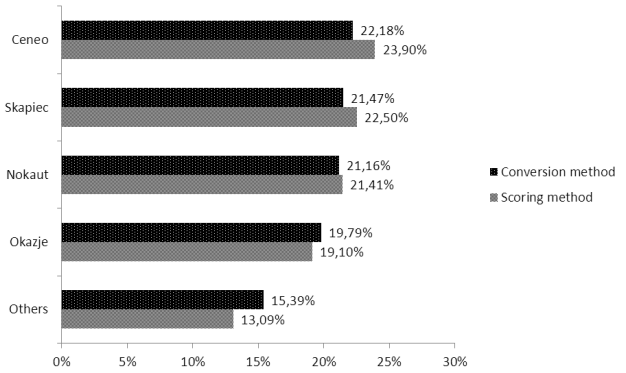


Fig. 2. Ranking of selected comparison engines websites according to a scoring method and a conversion method

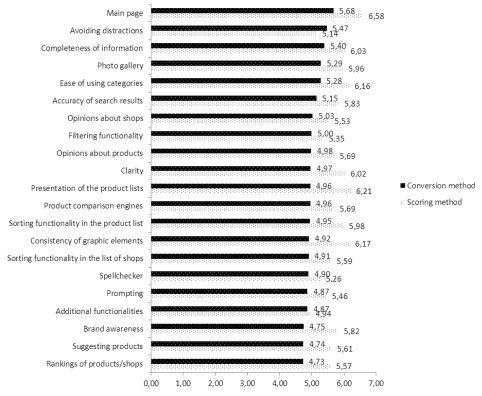


Fig. 3. Ranking of the degree of usefulness of particular criteria in the evaluation of the quality of comparison engines according to a conversion method and a scoring method

the requirements of comparison engines users. Figure 2 shows the results of calculations carried out by means of the scoring method and the conversion method brought to comparability.

In this case the conversion did not bring about any changes in the order of the evaluated comparison engines: it reduced the distance between them in such a way that for the engines which came first in the ranking there occurred differences in minus, and for the last ones in plus. Still, we can conclude that ceneo.pl may be seen as the best service among those selected for the evaluation.

However, there are no regularities with regard to the differentiation of criteria allowing for the detailed evaluation of each website. In this case we noted the reduction of the distances among the criteria. Also, the order of their evaluations changed in a significant way. The first positions were taken by: main page design, avoiding distractions and completeness of information about a product; the last ones were: brand awareness, suggesting products and rankings of products/shops. Only a third of the obtained criteria values exceed the average; thus, we may conclude that only those criteria may be seen as a model to be followed.

V. Conclusions

On the basis of the above results we can create — similarly as in the case of the application of the scoring method — the detailed basic pattern — a prototype which is a compilation of the best features of comparison engines, selected on the basis of the users' evaluations. This prototype should be confronted with the users' expectations established on the basis of their individual preferences.

The obtained findings (see: Table 3) allow for better matching of the results obtained in the rankings by means of conversion method (three times better than in the case of a scoring method). In this case good adaptation does not offer too much room for manoeuvre with regard to the selected evaluation criteria; however, it points to users' awareness of the most important features of the comparison engines.

Particular attention should be paid to those criteria where the discrepancies in the evaluations of the users' expectations and their realisations were the greatest. In subsequent iterations by way of "confrontations" (within the limits established by the obtained pattern), we refine the final prototype, and proceed to the next stage of the project. In the case of the application of a scoring method in the study, the discrepancies were significantly larger and concerned over 50% of the indicators. In the case of the application of a conversion method they do not exceed 10%. Further studies, however, may show whether there is a regularity resulting from the characteristics of a conversion method or it is just a coincidence.

The research is an extension of the previous one and will be continued. On the basis of the study findings the next method supporting website assessment will be applied. This time AHP/ANP method (T. Saaty) will be used, because it allows to remove (even better then conversion method) subjectivity of evaluations from the respondents questionnaires. Then, the selected methods of assessment websites will be compared with respect to their suitability for use inside a confrontational pattern-based design method. So, the theoretical and methodological achievements of this research will be enhanced and the verification will be conducted on the example of next Polish websites and portals.

References

[1] Buchanan J., Sheppard P., Lamsade D. V. Project ranking using ELECTRE III, at http://130.217.168.130/departments/staff/jtb/Electwp.pdf.

[2] Chmielarz W, Szumski O., Zborowski M.: (2011), Kompleksowe metody ewaluacji witryn internetowych *(On Complex Methods of Websites Asessment)*, Warszawa, Wydawnictwo Wydziału Zarządzania UW, Warsaw.

[3] Chmielarz W. (2013), Zarządzanie projektami@rozwój systemów informatycznych zarządzania *(Projrct management@Management Information Systems Development)*, Wydawnictwo Naukowe WZ UW, Warsaw.

[4] Chmielarz W., Zborowski M.: Conversion Method in Comparative Analysis of e-Banking Services in Poland in: rozdz. 4 pt.: Information Systems and Services in: Perspectives in Business Informatics Research eds. A. Kobyliński, A. Sobczak in: Lecture Notes in Business Information Processing nr 158, Springer Verlag, Berlin, Heidelberg, 2013, str. 227-240.

[5] Chmielarz W. (2014): Confrontational Pattern-Based Design Method — realization of service design methods in: proceedings of KM Conference 2014, in printing.

TABLE III
THE POSITIONS OF PARTICULAR CRITERIA IN THE EVALUATION OF COMPARISON ENGINES IN THE PRELIMINARY EVALUATION OF THE USERÂĂŃS REQUIREMENTS AND IN THE EVALUATION RESULTING FROM CALCULATIONS WITH THE APPLICATION OF A SCORING METHOD AND CONVERSION METHOD

Criterion	Preliminary analysis	Ranking according to a scoring method (SM)	Ranking according to a conversion method (CM)	The absolute value of the difference SM	The absolute value of the difference CM
Main page	3	1	1	2	2
Avoiding distractions	4	21	2	17	2
Completeness of information	2	5	3	3	1
Photo gallery	5	8	4	3	1
Ease of using categories	6	4	5	2	1
Accuracy of search results	1	9	6	8	5
Filtering functionality	7	19	7	12	0
Opinions about products	10	11	8	1	2
Clarity	8	6	9	2	1
Product comparison engines	11	12	10	1	1
Consistency of graphic elements	9	3	11	6	2
Opinions about shops	12	17	12	5	0
Presentation of the product list	14	2	13	12	1
Sorting functionality in the product list	13	7	14	6	1
Sorting functionality in the list of shops	16	15	15	1	1
Spellchecker	17	20	16	3	1
Additional functionalities	20	13	16	7	4
Prompting	15	18	18	3	3
Brand awareness	18	10	19	8	1
Suggesting products	21	14	20	7	7
Rankings of products/shops	19	16	21	3	2

[6] Flasiński M. (2006): Zarządzanie projektami informatycznymi *(Informatics Project Management)*, Wydawnictwo Naukowe PWN, Wasaw.

[7] http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/E-commerce_statistics, listopad 2013.

[8] http://wiadomosci.mediarun.pl/artykul/internet-internet,czy-dzieki-porownywarkom-cen-faktycznie-kupujemy-najtaniej,44430,2,1,1.html, listopad 2013.

[9] http://www.tradedoubler.com/pl-pl/informacje-i-zasoby/, listopad 2013.

[10] Meroni A., Sangiorgi D., (red.) (2011), Design for Services, Lancaster University, Farnham, Gower.

[11] Nowoczesne zarządzanie projektami *(Contemporary Project Management)*, (2012) red. M. Trocki, PWE, Warsaw.

[12] Orłowski C., Z. Kowalczuk, E. Szczerbicki (2009), Zastosowanie technologii informatycznych w zarządzaniu wiedzą *(IT Application in Knowledge Managemnt)*, PWNT, Gdańsk.

[13] Sikorski M. (2013), Usługi on-line. Jakość, interakcje, satysfakcja klienta *(On-line Services. Quality, Interaction, Customer satisfaction)*, Wydawnictwo Polsko-Japońskiej Wyższej Szkoły Technik Komputerowych, Warszawa.

[14] Zborowski M. (2013), Modelowanie witryn internetowych o profilu ekonomicznym *(Economic Universites Profile Websites Modelling)*, University of Warsaw, Faculty of Management, Warsaw, 2013, doctoral dissertation under the supervision of W. Chmielarz.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1235–1241
DOI: 10.15439/2014F320
ACSIS, Vol. 2

Semantic Organization of Information Resources for Supporting the Work of Academic Staff

Ilona Pawełoszek
Technical University of
Czestochowa, Al. Armii Krajowej
36B 42-200 Częstochowa, Poland
Email: ipaweloszek@zim.pcz.pl

***Abstract*—The paper presents an on-going project in developing a semantic information portal for academic institution. The concept of a semantic platform called SemLib, grew out of many academic discussions and it reflects the information needs of researchers and educators from Technical University of Czestochowa. The proposed method of needs assessment is especially designed to create the starting point to build semantic structure of the digital library. The author proposes an approach for managing, organizing and populating knowledge by semanticizing existing information resources in the digital and traditional paper form. The prototyped solution is based on the Semantic MediaWiki software, that allows for cooperative resource building, maintaining and offers enhanced querying capabilities. Experimental results demonstrate the potentials of the proposed system as well as some obstacles that are the subject to improvement by further development of the SemLib platform.**

COLLABORATION and exchanging knowledge are vital to the effectiveness of the work of academic staff. In a world permeated with information technology and overwhelming amount of data there is a growing need to improve tools and methods of organizing information resources. The intelligent content management tools are crucial factor for every large organization.

In today's economy the universities play important roles of teaching and research. They are "central generators and repositories of knowledge in our society" [17 p.5].

At present it seems that the most successful research projects and most remarkable publications are characterized by their multidisciplinary environment and ability to tackle research challenges on a broader, wider scale. Effective collaboration and knowledge sharing requires the right information to be published and easily accessible to community members and external stakeholders (like enterprises and public administration seeking for academic partners to cooperate).

Discussions among academic community members reveal their unsatisfied information needs, and necessity for robust knowledge sharing and communication tools [16]. A remarkable improvement can be achieved by harnessing together Semantic Web and social networking tools. Such a combination is often referred to as Web 3.0 [18].

The aim of this paper is to present the concept of information portal based on Web 3.0 paradigm to organize information resources to support the university researchers and educators in their tasks. The main emphasis has been put on identifying and creating the semantic structure of the information resources in a way that would make them most useful and accessible.

I. The need for Semantic Organization of Information

University faculties often have hundreds of employees in teaching and scientific positions, whose tasks are "knowledge intensive" and require the access to the up to date resources including: books, journals, whitepapers, educative multimedia etc. A lot of these resources are created by the academics themselves and published in printed or electronic form. Even if the information is available, it may not be easily accessible, especially if we consider a large collection of information sources spanning diverse domains.

A number of discussions with academic staff of the Management Faculty of Technical University of Czestochowa revealed that the information in many cases is poorly accessible and the information flow between employees is impeded by the lack of appropriate IT solutions.

The activities like teaching, research, organizational tasks, writing publications and self-education, determine the information needs of the academics. Both the electronic and the printed form, appear to be poorly usable in terms of searching for particular information. Having hundreds of volumes such as manuals, conference proceedings, journals close at hand doesn't mean the information they contain is easily accessible. It is always easier to use search engine than to leaf through a book searching for a definition, reference or a person who is recognized as an authority on the given subject. The Author's own experience as well as the multiple discussions with coworkers acknowledge that even having the files in the electronic form (i.e.: pdf, doc or html files) is not very convenient for searching for particular information and it is one of the most time consuming task in the work of academics. Most of the local and global search engines serve a keyword-based method of finding information and are not able to respect the context of the query or to process the detailed query attributes. Finding information is a part of everyday work of every educator or scientist.

978-83-60810-58-3/$25.00

There are many reasons arising from the policies of educational institutions as well as global circumstances that compel the authors to promote their work to be cited by other experts in their domain. On the other hand the authors are supposed to contribute to the domain knowledge, and take into account "the state of the art" of the subject they are working on. Therefore the scientiometric evaluation is important for all of the academics.

In all types of scholarly research it is necessary to attribute the author and the source of information that underpin particular concepts, positions and arguments with citations. This good practice is important due to a number of reasons:

• the citations help readers to identify and retrieve the source work to verify the information, or to learn more about issues and topics addressed by the work,

• citations provide the evidence that the subject is important and grounded in prior research,

• by citing one gives credit to the author of an original concept or theory presented.

For these reasons, it is important to build the systems that facilitate the bibliographic description of information cited in an organized and thorough manner. At the same time there is a growing need for creating effective search engines oriented towards supporting the retrieval of scientific information.

II. The Concept of Semantic Organization of Information Resources

Currently the Semantic Web technology solutions are based on ontologies that describe the structure of domain instances, their classes, attributes, distinctions and relations. The term "ontology" in the context of information sciences and artificial intelligence is most often defined as an "explicit specification of conceptualization" [1]. Such specifications are necessary for knowledge representation and exchange [2]

A conceptualization can be defined as: an intentional semantic structure that encodes knowledge of a piece of some domain. Ontology is a (partial) specification of this structure, i.e., it is usually a logical theory that expresses the conceptualization explicitly in some language. Conceptualization is language independent, while ontology is language dependent. [3]. A shared or common ontology refers to an explicit specification of concepts [4, p.99] which can be used by a group of people or program agents in a multiagent system.

Ontologies can differ in terms of their expressiveness. Constructing ontology-based systems is time-consuming and must be realized by highly skilled staff – knowledge engineers cooperating with domain experts. For some applications it is worth the cost, i.e. investing a lot of time and effort in constructing a 'perfect' ontology, but in many cases this is not feasible [5 p.10]. There is always a need to balance the tradeoff between the complexity and the effort (and cost) of construction and maintaining of the ontology in question.

Thus there is a growing number of methodologies that specifically address the issue of ontology development and maintenance, for example:

• The methodology by Ushold and King [6] – worked out for the construction of Enterprise Ontology. The most interesting part of the methodology is the procedure of informal ontology development which is based on creativity techniques (brainstorming) to develop an informal ontology that can be easily understood and many people and works as a starting point to formalization.

• METHONTOLOGY by [7] is an example of evolutionary prototyping methodology that consists of a set of activities based on ontology life cycle and the prototype refinement;

• AFM:Activity-First Method in Hozo proposed by [8] is a method of building ontologies of tasks and domains by exploiting technical documentation. The task ontology provides the set of roles played by the users in the context of a given task. The domain concepts are organized according to the identified roles.

• 101 Method [9] is an iterative approach to ontology development. The method was worked out for the needs of development of wine and food ontology, using the Protégé-2000 environment. The 101 method proposes four activities for the development of an ontology: (1 definition of the ontology classes; 2) organizing the taxonomy of classes; 3) defining properties of the classes and their permitted values, and 4) filling the attributes values for the instances.

It seems there is no just one best methodology for developing ontologies that would be useful in all cases and could become a standard. The above-mentioned methodologies extensively rely on human creativity and skills. Despite many semi-automatic approaches that have been developed, the ontology construction is still an art rather than craft.

The semi-automatic ontology construction is still a relatively new concept. These approaches are often referred to as ontology learning (OL) techniques. The example of such a semi-automatic approach is presented by [5], and focuses on constructing enterprise ontologies to be used for structuring and retrieving information related to a certain enterprise. The ontologies are quite 'light weight' in terms of logical complexity and expressiveness.

The semi-automatic ontology construction requires creating rather complicated algorithms based on NLP (Natural Language Processing) techniques, therefore these are specific methods for solving partial problems.

III. The Methodology for Identification of Semantic Structure

To the needs of our project (called SemLib – the abbreviation of Semantic Library) we decided to adapt and particularize the aforementioned general frameworks of manual ontology building based on human experience. The semantic organization of information resources can be defined as a process consisting of the following steps:

1. Gathering electronic documents and metadata about printed documents and objects.

2. Identification of the semantic structure - a global structure of information in a document base.

3. Choosing a tool and a formal language for designing ontology and describing information resources.

4. Designing an ontology - network of semantic knowledge.

5. Applying the formal language to annotate the documents according to the structure identified in step 2 with the chosen tool.

Our methodology of organizing information resources is especially dedicated to the situations where there exists the set of documents that must be semanticized according to user's needs in a multidisciplinary users community.

Semantization of a document base means adding data and structure to the documents to make them understandable for intelligent search engines [see also: 19 p.70].

The first task is to decide about the scope of documents that are to be organized in the semantic structure. In our case the resources include:

- The users profiles containing the information about: the scientific degree, professional affiliation, interests, position. The users profiles already exist on the website of the university, but they are useless when it comes to search for the persons with particular attributes, professional background, interests etc.
- Publications: books, journals, whitepapers, reports, educative multimedia etc.
- Descriptions of research projects and grants – it is an important information while searching for partners with professional experience in particular field.
- Descriptions of external institutions cooperating in research project or education.

A lot of these resources are created by the academics themselves and published in printed or electronic form. Surprisingly both electronic and printed form, appear to be poorly usable in terms of searching for particular information. Having hundreds of volumes such as manuals, conference proceedings, journals close at hand doesn't mean the information they contain is easily accessible. Therefore the semantic organization of information resources is the prerequisite of semantic processing and searching.

Some of the aforementioned resources usually cannot be directly semanticized because their document format does not allow for adding annotations. Therefore in this case only the metadata can be the representation of the document on the semantic platform. The semantic platform plays the role of the catalog consisting of metadata of things that exist in a real world (employees, organization units, electronic and hardcopy documents). Some of the documents can be directly semanticized – the precondition is the possibility to transform the document to HTML format and having the copyright that allows to use, publish and edit the content.

The identification of the document base semantic structure can be performed in different ways. For example [10] propose systematic and automatic approach to ontology construction through the automatic identification of keywords from a corpus of randomly collected unstructured texts.

Different approach is presented in the study of [11]. The authors present an approach that starts with a list of relevant domain ontologies created by human experts, and techniques for identifying the most appropriate ontology to be extended with information from a given text.

Although the automatic approaches are very accurate they can be applied only to the texts in a specific domain. In our case study the documents are heterogeneous considering their domains and their structure. For example the descriptions of the academics, their scientific and professional attainment may differ significantly. On the other hand the descriptions of literature (journal articles, monographs and other publications) have quite similar bibliographic structure and can be easily cataloged using common metadata format like Dublin Core. The descriptions of projects realized by the academics can also be multidisciplinary and differ in terminology.

Therefore the automatic approach to ontology extraction would require to distinguish many categories and to run the procedure many times to extract separate ontologies. The next step would be mapping the ontologies because the documents from different categories are interrelated.

Taking into account the multidisciplinary character of the designed platform and the requirements of usability and maintainability we decided not to use automatic methods of ontology extraction and concentrate on recognizing the users' information needs as a base of ontology construction.

The semantic structure of the document base should be first of all meaningful and relevant for the users. The functions of the platform should be adjusted to the context of the work tasks of the users.

The next section describes the method for identifying the semantic structure in the context of the users work.

The aim of the designed solution is to fulfill the information needs of the users by exploiting the easily accessible document base with a semantic query interface. An information need is a gap in one's knowledge[12] that can be bridged by providing an answer to a question. Thus asking the users to specify the questions they ask while searching for information seems to be the most appropriate research method in this case. Table I summarizes the questions specified by 20 potential users of the system, the questions were assigned to the groups that represent categories of objects described in the document base. The categories reflect the areas of professional activities.

The areas of context are general categories (classes) of objects. Questions represent attributes and/or relations between objects.

For example finding "experts in particular domain" (let us call the domain: X) can possibly involve finding the persons who participated in conferences and projects in the domain X or teaching subject X.

The next important thing to specify was the form and the particularity of the answer to our questions – what additional information is to be displayed. This additional information presents the context of the query (Table II).

In our example of finding experts in the given domain the query result could include: The names of the persons inter-

TABLE I.
THE QUESTIONS SPECIFIED BY POTENTIAL USERS OF THE SYSTEM

Question about:	Categories of objects								
	publication	Person	Organization structure	External institutions	Projects	Conferences	Lectures	Domains and subjects	Places
Definitions	x							x	
Research results	x							x	
Research methods	x							x	
Citations	x	x							
Scientific degree		x							
Workplace		x	x						x
Experts in particular domains	x	x			x		x	x	
Universities and academic centers				x					x
Articles about a given subject	x						x	x	
Citations in articles	x								
Taking part in projects as a member of the team or leader		x			x			x	
Conducting individual research,		x	x		x			x	
Participating at conferences,	x	x				x			x
Authoring and co-authoring of research papers, books, chapters, etc.	x	x							
Editorialship of scientific publications	x	x							
Educational Multimedia resources							x	x	
Being in charge of organizational units		x	x						x
Cooperation with industry				x	x		x	x	x

ested in domain X, the scientific degrees, the affiliations, number of citations, number of research projects.

IV. THE SEMANTIC SOFTWARE PLATFORM

In order to benefit fully from the networked information it is vital for organizations to have a single point knowledge shop for quality information with certainty, authority and consistency. [4, p.103]

Today's software platforms are designed with the architecture of participation in mind, this means they are meant for active user contribution in creation of content, functions and services. Web 2.0 platforms continuously grow in popularity due to the phenomenon of "network effect" which is driving force behind a number of users joining online communities [13].

Information portals allow easy access to heterogeneous data resources, applications and services in a consistent way.

Portals may also improve corporate communication by exploiting various tools based on Web 2.0 paradigm. The semantically enhanced information portal SemLib for the community of academic staff was meant to be a user-centered web application exploiting selected Semantic Web technologies and adhering to the following requirements:

- Scalable and extensible architecture.
- The portal content should be created, reviewed, published, and managed by web Content Management System (CMS) using well-defined content authoring and publication process. web 2.o architecture seems to be the best choice considering the large number of internal stakeholders and the dynamics of information (new publications will be added frequently).
- CMS should provide multiple authorization levels.
- Users, involved in editing, reviewing and publishing content.
- Easy semantic query interface – many Semantic Web technologies like RDF and OWL only offer a SPARQL query interface, which is difficult for users not having SPARQL expertise and no background in knowledge engineering.
- The portal should serve as a tool for promotion of scientific research and collaboration.

After considering the above requirements, we chose Semantic MediaWiki (SMW) as the platform to present the structured knowledge and manage the knowledge reasonably.

The aim of wiki is to organize and share information resources in a collaborative environment. Collaborative authoring is the effective way of creating knowledge proved by many examples, of which the most spectacular is Wikipedia. Apart from many advantages and flexibility, the MediaWiki platform is not free from shortcomings. The main disadvantage is the lack of knowledge structuring functionality. The answer to this shortcoming is Semantic Mediawiki package which is a set of extensions that provide semantic annotations and reasoning features.

The SMW achieves most of the requirements we preset while few of the requirements cannot be fulfilled and some of them cannot be directly realized. SMW is a free and open source extension of MediaWiki, released under the GNU Public License. SMW collects semantic data by letting users add annotations to the wiki source text of pages via a special markup [14].

The additional features of MediaWiki, such as the possibility of setting the authorization for the groups of users, and searching for historical changes extensively facilitates management tasks.

One of the most important advantages offered by Semantic MediaWiki is its semantic query language that lacks the complexity of SPARQL (e.g. querying within a particular namespace) [15].

Semantic MediaWiki has become the most popular semantic wiki engine by far, achieving several hundred installations worldwide and engaging an active open source developer community.

V. The Semantic Structure of the SemLib Platform

The semantic structure defines the categories, subcategories and attributes of the entities described in the document base. The attributes can also be categories. The seman-

TABLE II.
EXAMPLE INFORMATION NEEDS AND CONTEXT PARAMETERS

Needed information	Query context parameters
definitions of concepts	author, year of publication, Author's affiliation
research reports and results	author, institution, year of publication, geographical region, branch
citations of one's own publications	years of the referencing work
figures, tables, charts	subject, year of publication, relevant definition
multimedia	creator, subject, year of publication, file type
bibliographic references	year and place of publication, isbn or issn, author/authors
potential partners in project realization or coauthoring.	people, interests, authorship, participation in projects
conferences	subject, place, publication type, price

tic structure was designed according to the users' information needs unveiled by the survey. The structure of the information needs were analyzed from the point of view of categories and attributes. Figure 1 presents the simplified class diagram with relations between categories defined and applied in the SemLib prototype platform. Some of the relations were omitted on a diagram to make it more legible. For example an article can be understood as a part of the journal or as the book chapter. We decided not to define the separate class for the entities like book chapter and article because they would have exactly the same attributes. Although it can be inferred that if an article is a part of the book it can be called a "chapter", so it would be possible to display the separate lists of articles and chapters if needed. The relation "is written by" can contain multiple values, each of the values points to one entity of the class "Author".

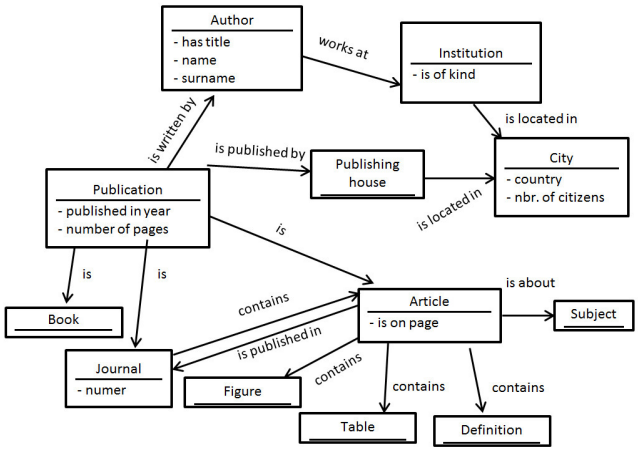


Fig. 1 Classes and relations in a prototype semantic platform

Source: Own study

The presented data model allows to ask compound queries respecting many context parameters.

The "subject" class that is used to categorize the content of publications, has many instances of the hierarchical structure. The example fragment of marketing subject class is:

- Marketing
  - o Customer relationship management
  - o Marketing mix
    - Promotion
    - Advertising
  - o Web marketing
  - o Customer service

The subjects hierarchy facilitates annotation tasks. For example if the described article is annotated with the tag "advertising", the search engine "knows" that the article is from the marketing domain although it may not be explicitly written in the page describing the given article.

Descriptions of example queries and their syntax are presented in table II.

The list displayed by the query 1 will also contain articles about promotion and advertising although they are not annotated with the word "marketing".

While designing the semantic portal based on SMW there is no need to define the whole categories structures in advance. For example the categories of subjects are created or developed when there is a need to describe the particular publication from a given domain. This ongoing iterative development approach will probably bring a better result of more accurate tagging if we assume that the authors will add their own publications in the domain of their expertise.

Analogously there is a possibility of searching for relations between projects, conferences and persons, so it is possible to list for example "all the publications from conference Y authored by person X" or "all the conferences in which the people from University C took part".

There are a lot of combinations of attributes and categories. Such lists can be settled according to date, place, institution, person and practically every attribute defined in the system. This functionality is very useful while preparing reports about activities of the University departments or individual employees.

Although the semantic wiki query syntax seems easy and intuitive thanks to its similarity to natural language the practice shows that after some time of disuse it is hard to remember the names of attributes used to pose the queries.

For now we are developing a help system in the form of example queries which users can modify according to their current needs. The example queries that were included in help system were chosen from the most often input queries. Additionally the users can use special wiki pages for listing the categories and attributes existing in the system. The namespace for special pages listing categories is: Special:Categories

VI. Conclusion and Future Work

The SemLib platform is very flexible and the directions of its evolution should be determined by analyzing the users' needs. The needs assessment method consisting in formulating example queries posed by users proved to be an appropriate and agile way to transform the users' needs to semantic categories and attributes.

The platform has been created and developed at Technical University of Czestochowa to complement the missing semantic functionalities of the existing information portals.

Probably the most time-consuming part of the discussed undertaking is the developing, maintaining and annotating the document base. Semanticizing the publications from conferences, journal articles and monographs is rather tedious task in spite of forms and templates that are accessible in Semantic MediaWiki interface. But the effort may turn out to be worthwhile if the portal brings results in facilitating collaboration and information access. It is hard to evaluate the success of the semantically enhanced portal in a quantitative way. Some measures can be applied, like: precision, recall, time spent to get answers to particular questions comparing to traditional keyword search or literature browsing, the number of emerging research teams or publications co-authored by groups using the SemLib platform.

At present the SemLib platform is continuously developed and it is perceived very promising and innovative.

Having all the metadata at hand makes the work faster when it comes to find particular information in the library. It is also possible to prepare multiple reports from the scientific activities much faster. The future research directions can encompass the proposition of a detailed methodology of assessing the performance of the system.

Apart from all the advantages there are also some obstacles that are subject to further considerations and development. One of potential problems is the possibility of inaccurate annotations. An idea to overcome that problem is to engage the community of end-users. If we assume the users are experts in their subject it is undeniable that after a short training, they can do the tagging far more accurately than the system administrators who will only perform the controlling functions.

As the practice shows the wiki passive users (the persons who are not editors) encounter some difficulties in using SemLib. The most often reported one was the lack of knowledge about wiki functions in general. Not all of the users seem to understand the nature of the wiki and the need for respecting some organizational guidelines, e.g. using the categories and attributes defined earlier by other users. Moreover it is in the interest of the users to describe their own publications in a way that ensures they will be frequently found, retrieved, and consequently more often cited.

The SemLib would be far more useful if all the described resources (books, journals, articles) were accessible in electronic form and linked directly to the semantic wiki site where they are described. Transformation of all the library resources to electronic form is inevitable in the future, but it requires a lot of work and changes in the copyright law. The old fashioned fossil library catalogs of publications do not have a potential to reveal the full value of information resources they describe because their data models are not flexible and cannot be easily adapted to dynamic users' needs. The semantic technology along with cooperative editing tools like MediaWiki support a community of participants, who interact with the content to varying degrees based on their permissions. All participants can be creators, editors, and curators [20].

The SemLib platform is quite new project that is still in the development phase, but we continuously add new publications to the semantic base and at the same time we are working to expand the ontology of research topics.

We are working on the concept of the more intelligent query interface that is based on the idea of contextual hints displayed to the user after analyzing the part of the query that has been already input into the search box.

REFERENCES

[1] T. Gruber, A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition. 5. 1993

[2] M. Obitko, Ontologies - Description and Applications. Report No. GL 126/01. Gerstner Laboratory for Intelligent Decision Making and Control Series of Research Reports. 2001. http://cyber.felk.cvut.cz/gerstner/reports/GL126.pdf

[3] Specification of Conceptualization Retrieved April 20, 2014, from http://www.obitko.com/tutorials/ontologies-semantic-web/specification-of-conceptualization.html

[4] Z Cui, V A M Tamma and F Bellifemine "Ontology management in enterprises", BT Technology Journal Vol 17 No 4 October 1999 pp. 98-107

[5] E. Blomqvist, "Semi-automatic Ontology Construction based on Patterns", Dissertation No. 1244, Linköping Studies in Science and Technology, Linköping 2009

[6] M. Uschold, M. King, "Towards a Methodology for Building Ontologies.Workshop on Basic Ontological Issues in Knowledge Sharing". 1995.

[7] A. Gómez-Pérez, M. Fernández López and Corcho O., "Ontological Engineering with Examples from the Areas of Knowledge Management, E-commerce and the Semantic Web". London: Springer 2004

[8] R. Mizoguchi, and J. Bourdeau, "Using Ontological Engineering to Overcome AI-ED Problems", International Journal of Artificial Intelligence in Education, Vol.11, No.2, pp.107-121, 2000.

[9] F. N. Noy, and D. L. Guinness, "Ontology development 101: A guide to create your first ontology". Stanford University, 2001. Retrieved April 20, 2014, from http://protege.stanford.edu/publications/ontology_development/ontology101.pdf

[10] A. Khurshid and G. Lee "Automatic Ontology Extraction from Unstructured Texts" Retrieved April 20, 2014, from https://www.scss.tcd.ie/Khurshid.Ahmad/Research/OntoTerminology/ODBASE2005.final.pdf

[11] Raghu Anantharangachar, Srinivasan Ramani, S Rajagopalan, "Ontology Guided Information Extraction from Unstructured Text", International Journal of Web and Semantic Technology (IJWesT) Vol.4, No.1, January 2013 Retrieved April 20, 2014, from http://arxiv.org/ftp/arxiv/papers/1302/1302.1335.pdf

[12] D. Nicholas: "Assessing Information Needs, tools, techniques and concepts for the Internet age", Taylor and Francis e-Library, 2005

[13] I. Pawełoszek, "Semantically Enhanced Information Portal for Community of University Researchers", Prace Naukowe UE we Wrocławiu nr 187, Informatyka Ekonomiczna nr 20. Wybrane zagadnienia. Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław 2011.

[14] D. Vrandecic, „Ontology Evaluation" Retrieved April 20, 2014, from http://www.aifb.kit.edu/images/b/b5/OntologyEvaluation.pdf

[15] A. Dumitrache and C. Lange, "BauDenkMalNetz – Creating a Semantically Annotated Web Resource of Historical Buildings" Retrieved April 20, 2014, from http://ceur-ws.org/Vol-721/paper-04.pdf

[16] D. Viehland, "ISExpertNet: Facilitating Knowledge Sharing in the Information Systems Academic Community", in: Issues in Informing Science and Information Technology pp 441-450, Retrieved June 17, 2014, from: http://proceedings.informingscience.org/InSITE2005/I36f30Vieh.pdf

[17] M. Abreu, V. Grinevich**, A. Hughes, M. Kitson and P. Ternouth, "Universities, Business and Knowledge Exchange" Council for Industry and Higher Education, and Centre for Business Research 2008, retrieved June 17, 2014, from: http://www.cbr.cam.ac.uk/pdf/University%20Business%20Knowledge%20Exchange%20v7.pdf

[18] P. Mika, "Social Networks and the Semantic Web", Springer 2007.

[19] R. M. Fay, M. P. Sauers, "Semantic Web Technologies and Social Searching for Librarians", Amercian Library Association 2012

[20] S. McNally, "Empowering the Distributed Editorial Workforce", In: M. Wolfgang, T. Kowatsch, (Eds.), "Semantic Technologies in Content Management Systems", Springer 2012.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1243–1250
DOI: 10.15439/2014F264
ACSIS, Vol. 2

Barriers in Creating Regional Business Spatial Community

Dorota Jelonek
Czestochowa University of Technology
ul. Dabrowskiego 69,
42-201 Czestochowa, Poland
Email: jelonek@zim.pcz.pl

Cezary Stepniak
Czestochowa University of Technology
ul. Dabrowskiego 69,
42-201 Czestochowa, Poland
Email: cstep@zim.pcz.pl

Tomasz Turek
Czestochowa University of Technology
ul. Dabrowskiego 69,
42-201 Czestochowa, Poland
Email: turek@zim.pcz.pl

***Abstract*—The article describes the obstacles that may arise during the planning phase of the project – Regional Business Spatial Community. The authors are working on a project to prepare RBSC . It involves creating a community around a selected GIS (Geographic Information System) software available in Cloud Technology. In the design of the project, the participants should be all the entities responsible for creating the infrastructure of the region. After conducting preliminary discussions with potential participants, the authors paid attention to the barriers and problems that threaten the planned project. Therefore, in the course of the study they drew attention to the mentioned aspect. The study identifies four groups of barriers: organizational, psychological, technological and financial.**

The authors also conducted empirical research to identify the most significant barriers to the creation of RBSC . The study involved five institutions providing utilities (gas, water, internet, TV) to the residents of the city . The results showed that the most significant barriers are organizational and psychological aspects. Slightly less importance is attributed to technological and financial barriers.

I. Introduction

REGIONAL infrastructure is one of the basic elements defining life standards and security of inhabitants in a given region. The aforementioned infrastructure consists of different elements, such as: roads, telecommunication, energy and gas networks, plumbing and different ones. Creating and maintenance of the mentioned elements of infrastructure are the obligations of many different entities, for instance enterprises functioning on commercial principles, community partnerships, government or local government departments or public services. All mentioned entities are interdependant.

Admittedly, decisions about the development of self-activity, at least a part of them, can be self taken, however, huge investments often demand intercommunication among the representatives of the selected entities. The necessity of the data exchange occurs, among other things, during planning and execution of municipal investments, crisis management [1] or different types of modernising infrastructure of media situated close to one another.

The authors of the study are working on the project called Regional Business Spatial Community (RBSC) - [2]. It seems that all the projecting entities, designing, owning or maintaining media should be interested in creating a proper system which enables data exchange and undertaking the cooperation within the confines of the suggested undertaking.

For the sake of the realization of RBSC, the GIS software functioning in Cloud technology available on the market was analysed, the group of potential entieties which should be interested in the creation of the suggested undertaking was selected and the condition of their IT infrastructure was examined.

The present study was dedicated to the analysis of the barriers which were singled out during the research and can make difficult or even impossible the realization of the mentioned undertaking.

II. The Idea of Regional Business Spatial Community Based on GIS in Cloud Technology

The idea of RBSC assumes the use of the Internet as the communication and spatious data exchange platform between the partners. After the analysis of GIS tools it was assumed that for the needs of the suggested project, the possible solution would be the use of Cloud Technology [3]. The mentioned technology enables a direct access to the worked out system (within the confines of the possessed entitlements) and the data content. It also enables electronic communication between the partners.

The worked out RBSC conception is based on the following assumptions:

- the GIS tool available in Cloud Technology is necessary,
- within the confines of the tool, a closed room, available for the entitled users only, is created,
- a communicator available for all the members of community functions within the confines of the room,

This work was not supported by any organization

- all the entities responsible for the infrastructure of a given region are invited to the community,
- the owner of the GIS tool is provides with the basic spatial data (geographic data),
- the other entities provide spatial data concerning their own infrastructure,
- data provided in the GIS tool should be accurate with the dictionary system and the codes used in the internal IT systems of the users (internal integration of different types of IT systems is desirable),
- participants of the community have access to the data on the basis of their entitlements,
- within the confines of RBSC, processes and code of conduction can exist.

The aim of creating RBSC is to build a platform enabling the access to the currently updating data concerning the infrastructure of the region [2]. By means of the communication tools, it will be possible to make arrangements between the entities concerned, involving different current or being planned investments, concerning the crisis management as well [4]. Within the confines of the tool, it will be possible to create a set of procedures facilitating the selected procedures of cooperation (on the level of investment planning or crisis management for example) (see also [11]). It can also be assumed that the provided maps will be intelligent. Due to it, it will be possible to capture all the problems arising on the project level (the representatives of the concerned entities will be informed currently about it) or they will facilitate the valuation of the arranged activity, for example. The conceptual model of RBSC is shown on figure 1.

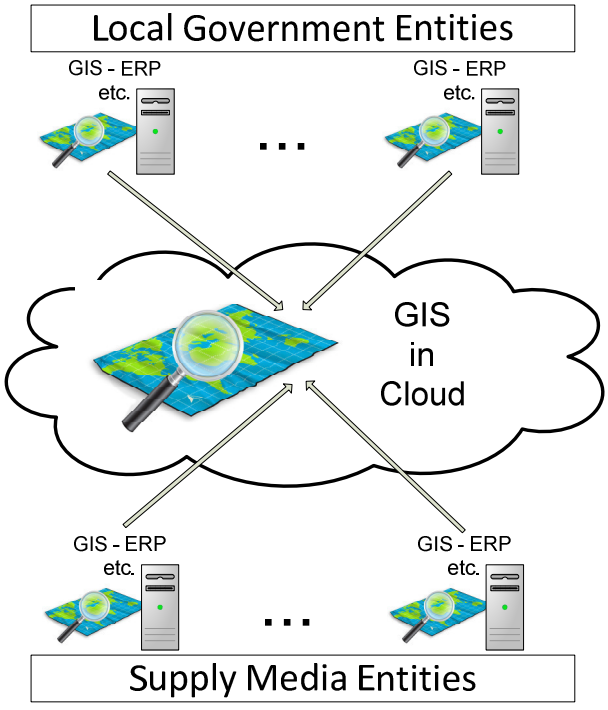


Figure 1. The conceptual model of RBSC
Source: [2]

Technological basis for the suggested project should be GIS tool available in Cloud Technology. Within the confines of the tool, the GIS software for visualisation, as well as actualisation of maps and project purposes should be enclosed. The supplier of the device should also provide the layers of maps which describe geographical object of the region. The demand of map intelligence means that the necessity of the description of link between individual layers of maps may be necessary. Intelligent maps would facilitate the process of infrastructure designing (pointing at all the potential conflicts in the infrastructure and enabling the preliminary valuation) – [5].

The use of Cloud technology system solves the problem of localization of the constructed system. The provider of GIS system in Cloud technology indicates the rules of access for the entitled users (the example is [6]).

RBSC are assumed to have limited territorial access of interaction. Usually they can apply to singular agglomeration, second level administrative units, however, only in the case of big systems they can include whole regions. As a result, one elaborated system can be duplicated in future for users from a different region. Generally, individual regions should be closed rooms. However, it may occur that some entities such as telecommunication enterprises can work in many regions. Then, the software supplier should decide the rules of using many rooms.

While creating electronic communities [7], the possibility of using GIS software only is not enough. GIS tools must be complemented with communicators which enable electronic consultation between the partners. It is vague especially in the case of these regions where the management or units responsible for designing and maintenance are outside the region. Due to this fact their share in the works of the agreement team (organized, among other things, during undertaking municipal infrastructural projects) is made more difficult. Another element which can be joined to the system is the tools used for designing processes (based on, for example; BPMN [8], BPML, UML, OWL and different ones). The mentioned tools enable to describe the procedures which would be conducted within the designed system. Introducing them can become the development of RBSC project.

The planned project originally should include all the entities responsible for the regional infrastructure. They can be, among the others, public administration offices, state or local government agencies, companies that provide utilities (municipal and commercial), and, to a limited extent, other social organizations and companies performing service functions with respect to infrastructure managing entities in the region.

Practically, in the initial stage of the project, a limited number of entities can join. Only with the development of the project other partners can join.

The supplier of the tool can be both a commercial company and the relevant departments of the State Geodetic. This second type of entities can provide the basic layers of maps. The suppliers of communicators and business process design tools can be commercial enterprises or groups of users working on the project. Using spatial data is done on the basis of reciprocity. Operators shall provide each other their spatial data and/or lexical systems and codes for the managing items of infrastructure. Most of the companies that form the infrastructure use in their business different IT systems, such as: MRP/ERP/BI, CAD/CAM, CRM, GIS. The used systems can be at different levels of internal integration. Taking part in RBSC should favor making work on the integration of internal integration and then support the integration of data available in GIS systems. The mentioned integration should not only allow the exchange of data between different systems, but also the creation of a coherent dictionaries systems and codes used in scale of enterprises, as well as in and RBSC. Each participant of RBSC receives appropriate permissions to use the data depending on several factors. These factors include, among the others: the existing law on the rights of access to data, the scope of services of its own data, the scope of the purchased rights (access to selected functions of the created system can be determined on a commercial basis). RBSC is meant a social enterprise rather. Admissibility of commercial solutions stems more from the need to maintain the system and the need to finance its development.

Preparing the project, it was taken into consideration that it is complex, both organizationally and technologically. Therefore, its conduction would be a multistep process. Its launch and eventual success will require overcoming many barriers.

III. TYPOLOGY OF BARRIERS

RBSC is a complicated undertaking mainly due to a large number of potential users coming from different organizations. Another problem that hampers the realization of the suggested project is of the environment system. It consists of several independent entities, none of which has decision-making powers to others. Environmental data binding components are only specific state legislation created by local law (enacted by local governments - concerning, inter alia spatial region), and the sense of community of interest (ie, projects that require collaboration infrastructure entities forming regions).

RBSC Organization should be treated as a project on two levels: organizational and technological. Its implementation should be carried out according to the methodology used in IT Projects.

Generally, IT Projects are carried out in several stages: the planned projects, analysis, modeling, verification and implementation. For each phase other measures are taken. Also at the various stages, various difficulties and obstacles, which should be overcome, are recorded.

In the case of RBSC the project is currently in the planning stage. Hence the occurring barriers must be overcome. Otherwise, the proposed project may never move to the implementation phase. The barriers existing in the planning stage can be divided into four basic groups:

- organizational,
- psychological,
- technological,
- financial.

A. Organizational barriers

Organizational barriers are the major obstacle to undertake the RBSC project. If they are not overcome, it will not be possible to undertake further work.

Analysing organizational issues, the following questions should be undertaken:

- defining the RBSC environment,
- identifying the processes and their targets in the RBSC environment,
- examining the sense of common interests of entities operating in the RBSC
- environment,
- identifying the relationship between the entities,
- suggesting the potential leaders of the project,
- identifying mechanisms to study the needs of users.

Defining the RBSC environment RBSC consists in determining the geographical impact of the system and determining the entities, which in the first instance must be acquired for the project. It is necessary to indicate whether there are entities that should definitely join RBSC.

The geographical scope may be a single or a set of contiguous administrative units. The problem is that the jurisdictions of individual organizations or their selected organizational units will not always perfectly overlap. It seems that the key factors for the success of RBSC will be adequate local government units. They have a real impact on the planning of spatial order, and may also initiate investment in infrastructure of an individual administrative unit. Moreover, they are responsible for crisis management. Their important asset are also the laws requiring them to collect specific type of data. These data can be collected on paper, but more and more frequently GIS are used. In addition, local government units may raise relatively greatest confidence.

The second group of entities represents companies supplying the media. Their importance lies in the fact that they design, build, and operate certain types of media in the region. They can affect the development of the region (still developing its infrastructure), as well as collect data on the course of networks providing a specific type of media. Very often their representatives are invited to the consultation, which are organized in order to arrange routes of particular media. On a smaller scale, access to these systems may also include sub-contractors involved in the construction,

maintenance or improvement of existing infrastructure networks.

When determining the environment of a project, as independent entities should also be considered IT companies providing IT tools, such as GIS systems, MRP / ERP and other Web and Cloud Services, and Business Procesess Modeling Tools. The closer the integration of information systems is assumed to be, the role of IT vendors will be greater.

The layout of the entities in the environment of a project enterprise should undertake the roles of particular entities. This will enable determining the objectives of the created community and to the basic processes carried out within its framework. In the course of the future formalization of processes, it will be possible to define RBSC functionality. Examining the sense of community of interest among potential participants RBSC is to determine the levels of cooperation that can be implemented using the proposed Community. In the future, it may be a basis for the negotiations on extending the scope of activity of the system.

Determining the relationship between the entities has to indicate that the mutual obligations of individual entities. What is results from the state law, local regulations, and what should be considered good business practice that should be encouraged, or in extreme cases, force the participants in their mutual relations.

The problem of every project is to create a suitable leader (integrator) of the project. In an environment where most companies are independent of one another it is a difficult task, taking into account the interests of the various participants. Generally for potential leaders, originators may be created, local government units or agencies designated by them, possibly IT companies providing IT Tools.

For the proper functioning of the created system it is necessary to create a mechanism to study the needs of users. It is a condition of the development of the system and may be a mechanism winding up the need for cooperation in this environment.

Taking into account the foregoing, the most serious potential organizational barriers include:

- lack of developed cooperation mechanisms or dislike between the identified entities in the researched environment,
- lack of widely respected leader,
- lack of knowledge of the proposed project,
- unclear legal provisions controlling the cooperation between these entities,
- marginality of the region for which RBSC is developed in the activity of suppliers of the individual infrastructure elements.

These elements can effectively prevent the adoption of this project.

B. Psychological barriers

Psychological barriers are a kind of supplement of organizational barriers. In the planning phase of the project, the key issues to be considered can include:

- a sense of common interests of potential participants in the project,
- self-independence,
- interests of individual entities operating in the environment.

The main problem is to develop a sense of common interests. At the stage of individual regions, it can be difficult especially when the provider of individual media are corporations, for which the region is of marginal importance. Investments for a given region are not of primary importance for their business. In contrast, the supervision over an infrastructure functioning in the region is held by units at a relatively low level in the hierarchy of the organization. Even more difficult situation may occur when supervisin over infrastructure can be taken over by an outsourcing company, whose only task is to maintain the infrastructure tools without permission for their development. Then, a tendency to participate in the proposed project may be relatively small.

Another element is the question of self-independence. Creating the infrastructure of individual media in the region could have signs of some kind of internal competition. The course of individual types of media is defined by the rules. Therefore, individual entities may be willing to act on the principle of the first in the game. Who is the first to take certain design and construction actions, can be given priority in the selection of the course of their own network. Other entities will be forced to adapt to a fait accompli. This type of phenomenon will occur, especially in those regions where there is no tradition of co-operation or where the local government has failed to practice cooperation with providers of media.

The phenomenon can still be affected by the internal systems of interest groups. Not all organizational units within individual companies may be interested in such a high level of IT system integration, fearing an increase in the level of internal control (external as well). Single IT service providers, for which in the event of realization of the proposed project may run out of space, can also be threatened.

To sum up the basic psychological barriers include, among others:

- too close corporate systems and closeness to cooperation among providers of media in the region,
- lack of tradition and willingness to cooperate,
- mutual dislike or distrust to cooperation between potential participants RBSC,
- negative attitudes of managers and selected organizational units of individual entities,
- disregard of project proposals,

- lack of knowledge about the range of costs and potential effects of the project,
- diversionary activities of external entities interested in maintaining the current status quo.

These barriers can have a particularly strong role in the initial stages of starting up the project, where there is still relatively large uncertainty about the potential effects of the project, whereas the current rules of functioning have been violated.

C. Technological barriers

Technological barriers are associated with the preparation of appropriate solutions for IT technology [9]. On the one hand, the proposed solutions should provide an appropriate level of quality of services provided. The indicators should include, among others: the appropriate scope of the data collected, clear and understandable rules of visualization, high level of intelligence of the implemented tools and system security.

Undertaking the project such as RBSC, it is necessary to assume that there are solutions available on the market which enable technical development of the system and achieving its full assumed functionality while maintaining economic rationality of the project.

In this case, among the others, the following issues should be analysed:

- analysis of commercially available IT tools (GIS application, instant messaging, BPM Tools),
- analysis of the IT systems used by potential participants in the project,
- possible integration of the systems used by potential RBSC users,
- analysis of potential and planned developments of previously used IT systems,
- providing an adequate level of security of IT systems,
- possibility of data flow between (system rooms) for entities operating in many regions.

On the IT market there are many different types of tools: GIS, instant messaging, BPM Tools. Requirements for GIS is a tool available in Cloud Technology for creating maps of the region, together with the use of technology of intelligent design. Intelligent solutions can be based on semantic technologies for defining the relationship between the various leyers (classes of objects in spatial databases). They can be an independent software which, however, must be compatible with the used GIS system.

Communication within the community may be also provided by the commercially available enterprise instant messaging (eg Microsoft Lync). Unlike traditional communicators they enable the smooth functioning of corporate they network, they have a high level of security and allow the integration of systems (eg ERP, GIS).

BPM Tools the expected part in the future. A prerequisite for its use is to create a community and define the basic processes that will be implemented in its framework. Its use will depend, inter alia, of information systems that will be used among community participants. Moreover, the question arises whether the participants would be interested in such a high level of integration. However, already at the planning stage, it is necessary to take into account this type of opportunity.

Another technological indicator of the discussed project are internal IT systems used by potential participants in the project. The point is the systems of MRP / ERP and BI, CRM, CAD / CAM types, various types of business systems and others. They are systems that will provide data on real events, and therefore create a description of the current state of the slice of reality. In addition, they will create the basis for the code, lexical and semantic of the planned project. Hence, the question of mutual integration of these systems will be crucial. IT Systems used by potential participants in the project are continuously being improved. Therefore, the suggested solution should not constitute any restrictions in this regard.

An important issue is the problem of security of data available in the system. On one hand, it is a technological challenge for manufacturers of IT solutions, on the other hand, an adequate level of safety should encourage increased willingness to cooperate in the planned community.

To sum up, the basic technological barriers include:

- The scope of the internal computerization of potential participants in the proposed project.
- The level of internal integration of IT systems.
- Ability of integrating GIS tools with the available instant messaging and BPM Tools.
- Ability to integrate all systems envisaged in the planned community,
- Providing an adequate level of safety.

From a technological point of view, planning, designing and implementation of the proposed system is feasible. The thing is, it is to be more economically rational, and therefore it is important to use as much as possible systems which can be duplicated.

D. Financial barriers

RBSC is a project in which it is difficult to determine the primary beneficiary. Moreover, the assumption is not to create a sense sricto economically effective project, but first of all, it has to be useful for the participants of it. Despite this, at its creation, and then the exploitation, economic rationality should be kept. In many cases of planned projects, financial considerations were the main reason for the failure of the project. RBSC is an extensive project and requires the cooperation of many entities. Therefore, it seems to be a cost demanding project.

In practice, the drawn up solutions should include, in the widest possible range, the ones which can be duplicated (thus, relatively cheap). On the other hand, the implementation of all the planned functionality of the system

can be decomposed into a number of stages and thus, extend it in time.

It can be assumed that analyzing the financial issues of the planned project the following issues should be considered:

- sources of project financing,
- the range of payments for the use of the created system,
- entities responsible for the functioning of the system.

Drawing up the proposed project requires work and commitment of many people from different organizational units. The primary objective of the project is to create a useful system. Therefore, at least the initial phase of the project can be financed from different types of funds. Only after activation of the project, it is possible to specify the rules for the financing of the system and to determine the extent of the payment for the selected system functions It is also important to define the principles of financing the development of the system in the future. Considering the fact that the system should be largely made up of duplicated tools, the need for payments to certain suppliers of IT services should be taken into account. Taking into consideration the issues presented above, the main financial barriers for the implementation of the project include:

- the need to pay the installments for the use of the system,
- reluctance to pay multiple suppliers for the use of similar IT services where some operators are already using systems with similar functionality,
- legal problems that might arise for certain entities (referring especially to the units of public administration), related with the possibility of the use of commercial systems,
- height and multiplicity of the charges for the use of the system,
- lack of decision-making powers to bear the costs for the use of IT tools selected at the regional level (referring mainly to corporations with head offices outside the region).

The presented barriers are just some of the problems to be dicussed before starting the project. The results of empirical research on the barriers to the implementation of RBSC in Częstochowa provide further points of the article.

IV. THE EMPIRICAL RESEARCH

In order to verify practically the issues brought up in the article, the authors conducted a study determining the most significant barriers connected with the formation of RBSC. The theoretical bases for the creation of RBSC and pilot studies in the area of information resources of such projects have been presented in previous publications written by the authors [2] [10].

Five entities with their own technological infrastructure, where resources are distributed and therefore can be managed through GIS systems, functioning in the region of Częstochowa were researched. Obtaining a larger study sample was relatively difficult because the entities are often a part of larger entities. People working in local offices are not authorized to provide information, and reaching the management requires complex formal procedures. Among the entities of Częstochowa region who have agreed to participate in the study are:

- Silesia Board of Amelioration and Water Division of Czestochowa,
- Polish Oil and Gas Company SA Upper Silesian Sales Department. Upper-Silesian Region. Customer Service Czestochowa,
- one of the nationwide providers of cable television, broadband internet and telephone services (did not give the permission for the use of the name),
- Częstochowa City Hall,
- Water and Sewage District Czestochowa SA in Częstochowa.

The research was carried out by notifying the appropriate application to the management of the enterprise. A representative of the board usually passed conducting the research on relevant departments. Additionally, a conversation with a representative of the company providing solutions in GIS Cloud was conducted. Representatives of the company have to prepare their own proposed solution within the confines of this concept. These studies are a pilot, expected to be extended in the future.

Research questionnaire, which was used in the study contained 16 questions divided into 6 main areas:

- Area 1 - (Questions 1-2) on the business nature of the entity,
- Area 2 - (Questions 3-5) on the level of computerization of the entity,
- Area 3 - (Questions 6-8) on willingness to cooperate with other entities,
- Area 4 - (questions 9-11) concerning the use of GIS,
- Area 5 - (Questions 12-15) for the local business communities - such as RBSC,
- Area 6 - (Question 16) concerning barriers to the creation of RBSC.

The study was conducted in April 2014. They will be a part of a larger project. For the purposes of this study only that part of the results that directly corresponds to the subject of the article (technology, level of computerization, barriers of accessment to RBSC) were taken into account.

Most subjects (4 out of 5 respondents) declared being a unit independent in decision making. This means that in spite of belonging to the structure of larger organizations or networks, they have autonomy and can take specific initiatives (e.g. accession to RBSC). Only the provider of cable television declared that they could not make decisions for themselves, and the management is done through the control panel.

The condition for the establishment a local business community (or joining such a project) is to have an appropriate level of computerization. The use of ICT solutions (information and communication technologies) is

the basis of RBSC. All the researched entities declared that they meet the basic requirements in this area, namely: they have broadband internet access, they use ERP and GIS. Unfortunately, no company has made a full integration of these systems. In the majority of cases (3 to 5) between ERP and GIS systems, only a simple (using files), or semi-automatic data exchange is possible. The other two entities (Water and Sewage and the City Hall) has integrated only some of trade systems.

The third group of questions contained in the form of research related to willingness of the entities to cooperate on the platform RBSC. Three entities (water, drainage, gas works) are interested in permanent cooperation, particularly in the area of investment planning and in crisis situations. A representative of the City Hall sees such a need only temporarily, in the implementation of the relevant investment. Internet, television and telephony supplier sees no need for such cooperation.

Entities interested in cooperation agree on the scope of the data that should be on the RBSC platform. The most important task in this area is to place the cadastral maps of infrastructure objects. This will help to streamline and accelerate processes of investment and planning the course of the subsequent parts of the network (gas, water, land reclamation). The most important part of the form of the research were questions that directly relate to the barriers to RBSC creation. These correspond directly with the theoretical considerations set out in the previous paragraph of the article.

The most important barrier to be considered is the organizational factor. This is indicated in the responses in all five forms. Respondents fear of who would organize a platform for cooperation and how they would define the powers and competences of the entities involved in the project.

Another important barrier is the psychological aspects (3 replies to 5). Respondents are concerned about the security of resources and the way of their share. Enterprises treat data and information as a resource of great strategic importance. The mechanisms to protect data are rooted in the mentality and psyche. RBSC in some way, destroys these behaviors, which may be unacceptable, even taking into account the common interests of the users of the platform of cooperation. Indication of technical barriers occurred in two cases.

Main barriers of creating Regional Business Spatial Community are shown on figure 2.

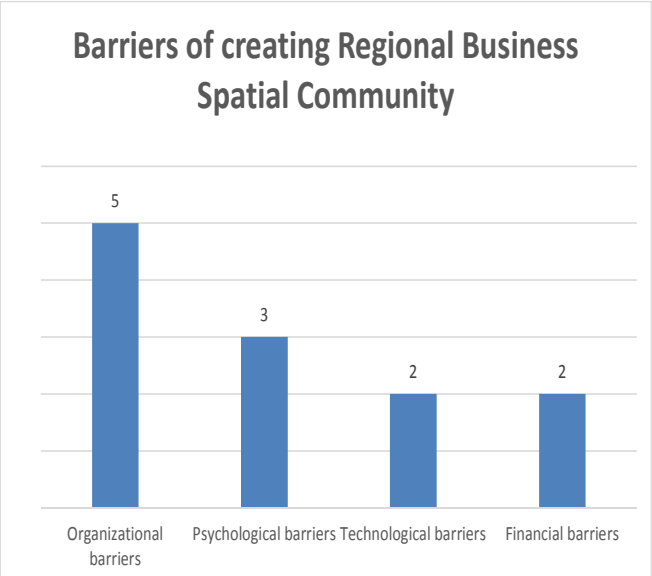


Figure 1. Main barriers of creation RBSC

Respondents believe that the integration of data on a common RBSC platform can be difficult due to the different formats of the data contained in the systems and incompatible software. Number of problems in this area will accumulate together with joining of other entities to the project. In two cases, respondents also related to the financial barriers. In the assumption, RBSC would exist as a separated entity.

Debatable would be platform budgeting, maintaining the necessary IT infrastructure, software or service.

V.SUMMARY

Inherent feature of the information age is the formation of the various communities that communicate using information and communication technologies. Factors affecting the formation of such communities are common interests, science, origin, occupation, etc.

The idea of RBSC is a new concept which indicates the possibility of the formation of business communities, based on ICT. Such communities can be formed only on the condition of cooperation and common benefit. A factor influencing the creation of RBSC may be having and infrastructure development of companies on a common geographical area. The article focuses on companies operating within the city, which provide utilities for the residents (water, gas, the Internet, etc.). The infrastructure possessed by these entities usually takes the form of a network. Networks of different companies often intersect, or are a barrier for itself. Creating a business community, which in cooperation would make specified information resources available, could help to improve the planning, development, and infrastructure repairs. As a survey among companies supplying the media revealed, there is a potential willingness to cooperate in this area.

The rise of RBSC, however, involves the occurrence of certain barriers. In the theoretical part of the article, four main areas of their occurrence were pointed out: organizational, psychological, technological, and financial. The study among potential participants and beneficiaries of RBSC confirmed these considerations. Surveyed companies have declared that the biggest problem may be psychological and organizational barriers. However, technical and financial aspects were considered less important.

References

[1] A. Kavanaugh,, E. A. Fox, S. Sheetz, S. Yang, L. T. Li, D. Whalen, D. Shoemaker, P. Natsev and L. Xie, *Social media use by government: from the routine to the critical.* In: Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times, dg.o '11, pages 121–130, New York, NY, USA. ACM 2011.

[2] D. Jelonek, C. Stępniak, T. Turek, *The Concept of Building Regional Business Spatial Community.* In: ICETE 2013. 10th International Joint Conference on e-Business and Telecommunications. Proceedings. 29-31 July 2013, Reyklavik, Iceland 2013.

[3] M. A. Bhat, R. M. Shah and B. Ahmad, Cloud Computing: *A solution to Geographical Information Systems (GIS),* International Journal on Computer Science and Engineering", vol. 3, no. 2, pp. 594–600, 2011.

[4] T. Howard, *Design to Thrive: Creating Social Networks and Online Communities that Last,* Morgan Kaufmann Publishers, Burlington 2010.

[5] C. Yang, M. Goodchild, Q. Huang, D. Nebert, R. Raskin, Y. Xu, M. Bambacus, and D. Fay, *Spatial cloud computing: how can the geospatial sciences use and help shape cloud computing?,* International Journal of Digital Earth, vol. 4, no. 4, pp. 305–329, 2011.

[6] V. Kouyoumjian, *The new age of cloud computing and GIS,* ESRI white paper, 2010.

[7] H. Fulford, *A Local Community Web Portal and Small Businesses In: Encyclopedia of Portal technologies and Applications.* IGI Global Harshey PA, p. 559 – 563, 2007.

[8] J. Mendling, M. Weidlich (Ed.), *Business Process Model and Notation: Proceedings of the 4th International Workshop,* BPMN Vienna, Austria, September 12-13, 2012, Springer London 2012.

[9] J. Wieczorkowski, P. Polak, *The Specificity of Software for Distributed Organizations – The Proposal of an Enterprise Model. In: Proceedings of the IADIS International Conference Information Systems, edited by Miguel Baptista Nunes,* Pedro Isaías and Philip Powell, International Association for Development of the Information Society, IADIS Press, Lisbon 2013, s.134-141, 2013.

[10] C. Stepniak, T. Turek, *Integration of Spatial Infromation Resources on the Example of Utility Companies in Czestochowa Region, Online Journal of Applied Knowledge Management,* A Publication of the International Institute for Applied Knowledge Management, Volume 2, Issue 2, 2014, http://www.iiakm.org

[11] N. Geri, *Overcoming the challenge of cooperating with competitors: Critical success factors of interorganizational systems implementation. Informing Science,* 12, 123-146, 2009, Available at http://inform.nu/Articles/Vol12/ISJv12p123-146Geri532.pdf

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1251–1258
DOI: 10.15439/2014F314
ACSIS, Vol. 2

Identification of mental barriers in the implementation of cloud computing in the SMEs in Poland

Dorota Jelonek, Cezary Stępniak, Tomasz Turek, Leszek Ziora
Faculty of Management, Czestochowa University of Technology,
Al. Armii Krajowej 19B, 42-200 Czestochowa, Poland
Email: jelonek@zim.pcz.pl; cstep@zim.pcz.pl; turek@zim.pcz.pl; ziora@gazeta.pl

***Abstract*—Cloud computing is an emerging new computing paradigm designed to deliver numerous computing services through networked media such as the Internet. This solution offers many possibilities that did not exist before for all enterprises but especially for small and medium sized companies, which very often cannot afford for huge investments in contemporary IT solutions. The aim of the paper is an attempt of barriers identification especially mental barriers of managers which hinder making decision concerning implementation of cloud computing in small and medium sized companies in Poland. In the paper the notion of cloud computing was presented as well as benefits in the aspects of: technical, financial, organizational ones and barriers of cloud computing adoption. Next the results of research conducted in Polish SMEs companies were presented. The research showed that the biggest mental barriers perceived by managers concern lack of trust, incomplete knowledge about cloud computing and not perceiving necessity of changes in currently used IT model.**

I. Introduction

CLOUD computing is currently the fastest growing IT service. This approach offers several advantages to potential users such as "metered" use (i.e., pay-as-you-go) which offers scalability, online delivery of software and virtual hardware services (e.g., collaboration programs, virtual servers, virtual storage devices) which would enable organizations to obviate the need to own, maintain and update their software and hardware infrastructures. The flexibility of this emerging computing service has opened many possibilities for organizations that did not exist before. Among those organizations are those engaged in healthcare provision.

Cloud Computing for many years has been mentioned among the most important trends of IT development. According to Forrester report the global market of Cloud Computing services in 2011 was 40,7 billion USD [1] and prognoses tell that in 2020 this market will grow to assessment value of 241 billion USD, it is six times.

The literature review reveals that many studies were and currently are being conducted on the use of cloud computing by large scale enterprises primarily on their perceptions about cost reduction, ease of use and convenience.

Cloud computing is adopted by enterprises [2],[3], public sector [4], eGovernment [5], regional business community [6], for education [7], for healthcare provision [8] and many other organizations.

The literature review reveals that many studies were and currently are being conducted on the use of cloud computing by large scale enterprises primarily on their perceptions about cost reduction, ease of use and convenience, reliability, sharing and collaboration and lastly but not the least, security and privacy [2], [9]. Large enterprises have quickly adopted this model of IT resources management [2] but among SMEs companies it is more and more often used solution [10], [3], [11]. The importance of cloud computing parameters on SMEs adoption in a large degree convergent with perspective of big companies e.g. King [3] indicated that the cost reduction, avoiding natural disaster mishaps, better security but lack of reliability in using cloud computing are the most important parameters. On the other hand Gupta et al [10] indicated that for SMEs companies which want to implement cloud computing model the most important are such factors as: ease of use and convenience which is the biggest favorable factor followed by security and privacy and then comes the cost reduction. The fourth factor reliability is ignored as SMEs do not consider cloud as reliable. Last but not least, SMEs do not want to use cloud for sharing and collaboration and prefer their old conventional methods for sharing and collaborating with their stakeholders [10].

The aim of this paper is to identify mental barriers in implementation of cloud computing in SMEs in Poland. There were proposed considerations of mental barriers from the perspective of: lack of trust, incomplete knowledge on cloud computing and not perceiving necessity of changes in the currently used model of IT resources management. The research was conducted in 134 SMEs companies, taking into consideration perception of mental barriers from the perspective of strategic level managers (CEO), tactical level managers and IT managers (CIO).

The rest of this paper is organized as follows: presentation of the notion and models of cloud computing, discussion concerning benefits of implementation of this solution and most often identified barriers of cloud computing implementation. Next the development of cloud computing in Poland was presented. As an introduction to presentation of research results and identification of mental barriers of cloud computing implementation the components of trust in cloud computing were presented.

978-83-60810-58-3/$25.00

II. Cloud Computing

There are many definitions of cloud computing, in which different aspects are underlined of this still new for some enterprises model of IT resources management. The technology is most often the key element of definition [12], some authors focus on the business model e.g. collaboration and pay-as-you-go [13] and the reduction in capital expenditure [14]]. Different authors have different opinions about the core conceptualizations of cloud computing. The reasons for these different opinions are as follows [15]:

- As an emerging IT service model, cloud computing has had only a relatively short lifetime in which to develop into a fully formed paradigm.
- Contributors to the development of cloud computing theory are extremely varied, both in terms of industry and academic background.

Grossman and Gu [16], described cloud computing as an infrastructure that provides resources or services over a network, often the Internet, usually at the scale and with the reliability of a data center.

Buyya et al. [17] defined cloud computing as a type of parallel and distributed system consisting of a collection of inter-connected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and consumers.

One of the most complex definition is provided by National Institute of Standards and Technology: cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [18]. This cloud model is composed of five essential characteristics, three service models, and four deployment models.

The most important characteristics include: on-demand self-service, broad network access, resource pooling, rapid elasticity, measured service [18].

Cloud services delivery models can be broadly categorized into three types: Infrastructure-as-a-service (IaaS), Platform-as-a-service (PaaS) and Software-as-a-service (SaaS). Each of the service types serve different purposes and target different customers however they share a common business model. The fundamental of the model is an offer of making accessible of their computing resources including services, applications, infrastructures, and platform to customers.

Infrastructure as a Service (IaaS) means using computer hardware via the Internet. IaaS is divided into Compute Clouds and Resource Clouds [19]. Compute Clouds provide access to computational resources such as CPUs, hypervisors and utilities. Resource Clouds contain managed and scalable resources as services to users.

Infrastructure as a Service model refers to the tangible physical devices (raw computing) like virtual computers, servers, storage devices, network transfer, which are physically located in one central place (data center) but they can be accessed and used over the Internet using the login authentication systems and passwords from any dumb terminal or device [10].

Platform as a Service (PaaS) is a more advanced level of cloud computing service than IaaS. PaaS provides a full or partial application development environment that enables developers to access resources for application development and to collaborate with others online.

Clouds and Resource Clouds [19]. Compute Clouds provide access to computational resources such as CPUs, hypervisors and utilities. Resource Clouds contain managed and scalable resources as services to users.

Infrastructure as a Service model refers to the tangible physical devices (raw computing) like virtual computers, servers, storage devices, network transfer, which are physically located in one central place (data center) but they can be accessed and used over the internet using the login authentication systems and passwords from any dumb terminal or device [10].

Platform as a Service (PaaS) is a more advanced level of cloud computing service than IaaS. PaaS provides a full or partial application development environment that enables developers to access resources for application development and to collaborate with others online.

By choosing a specific model of cloud computing the service user defines the division of control over the IT resources employed between himself and the service provider [20]. Division of control in cloud computing models was shown in Fig. 1.

In the traditional model the user has almost total control over the infrastructure and software he owns. However, in many cases his self-sufficiency is somewhat limited by the necessity to use Internet Service Providers.

In the IaaS model almost the entire substantial part of the IT infrastructure is outsourced. The user retains control over his data and software.

In the PaaS model the service provider also provides the service customer with the operating environments, where the user may operate the applications he installs.

In the SaaS model the entire infrastructure along with the software is under the control of the service provider. The user retains control over his data.

There are four different cloud deployment models within organizations namely [21],[10]:

1. Public cloud: It is available from a third party service provider via Internet and is very cost effective for SMEs to deploy IT solutions.
2. Private cloud: It is managed within an organization and is suitable for large enterprises (managed within the walls of the enterprises). (…) Private clouds provide the advantages of public clouds but still incur capital expenditures.
3. Community cloud: It is used and controlled by a group of enterprises, which have shared interests. For example, the US federal government using community cloud (built on Terremark's Enterprise cloud platform) for forms.gov, flu.gov, cars.gov, USA.gov, Apps.gov.

Traditional model On-premise	Infrastruc-ture as a Service IaaS	Platform as a Service PaaS	Software as a Service SaaS
Data	Data	Data	Data
Application	Application	Application	Application
Runtime environment	Runtime environment	Runtime environment	Runtime environment
Virtual machine	Virtual machine	Virtual machine	Virtual machine
Server	Server	Server	Server
Data store	Data store	Data store	Data store
Network	Network	Network	Network

Under control of the user
Under shared control
Under service provider control

Fig. 1. Division of control in cloud computing models [20]

4. Hybrid cloud: It is a combination of public and private cloud.

For many users handing over almost total control over their own IT resources is hard to accept and that is why the users very often decide to apply private or hybrid cloud model.

III. CLOUD COMPUTING BENEFITS

Cloud Computing will have a profound impact on the cost structure of all the industries using hardware and software, and therefore it will have significant impact on [22]:

- Business creation
- Macroeconomic performance
- Job creation in all industries
- Job reallocation in the ICT sector
- Public finance accounts, through direct impact on public sector spending and the indirect one on the tax revenues.

Data processing in the cloud gives new possibilities thanks to which organization uses its resources in a more costly effective and flexible way. Benefits may be considered in a three aspects:

- technical benefits,
- financial benefits,
- organizational benefits.

Technical benefits are perceived by the prism of flexible usage of computational power. In case of typical computing architecture with dedicated servers for every application the computational power of big and expensive computers is not used by most of the time. In a cloud computing shared resources are made accessible to those applications which require it at a given period of time. Thanks to this solution [23]:

- an access to computational resources for every application is growing,
- investments into hardware are limited with simultaneous decrease of time needed for computing purpose,
- computing power is allocated for those applications which currently need it,
- the constant level for access to services is maintained regardless of the number of users (scalability),
- there exist the possibility of fast acquisition of external computing power.

In case of cloud computing usage enterprises pay only for resources which were used.

Providers of cloud computing ensure for their customers access to the latest version of software (automatic upgrade) and SaaS.

Financial benefits contain most of all cost reduction (installation and maintenance) [14]. Cloud computing allows for decrease of costs connected with purchasing software licenses. Shorter time–to-market, means the possibility of faster introduction of product onto a market and achievement of financial benefits.

Organizational benefits embrace increase of flexibility and enterprise's openness on introduction of new solutions and increasing efficiency of its functioning. Its sources may be find in:

- limitation of backstage resources in a form of resources (external companies, IT).
- faster implementation of new business applications and IT solutions,
- transfer of the risk connected with IT infrastructure on the service provider,
- easier access to applications for mobile employees.

Summing up considerations concerning benefits of cloud computing application it is worth mentioning IDC research results [24]: demonstrable, tangible economic benefits are available from the adoption of cloud in the EU amongst enterprises. Based on the survey conducted for this study, that included 479 enterprises already using cloud for their businesses, 81% reported lower IT costs with a 10 to 20% reduction being typical, but 12% reported savings of 30% or more. Business benefits included more effective mobile working (46%), higher productivity (41%), more use of standard processes (35%), better ability to enter new business areas (33%) and the ability to open up in new locations (32%) [24].

The main perceived benefits of the cloud computing model in SMEs were also the subject of The European Network and Information Security Agency research [25]. Fig. 2 shows the results of the study.

The costs savings are thought to be the most important benefits of cloud computing. 68% of SMEs specified that this solution would help to eliminate or minimize investments in IT infrastructure. Companies indicated that adjustment and flexibility of IT resources is also very

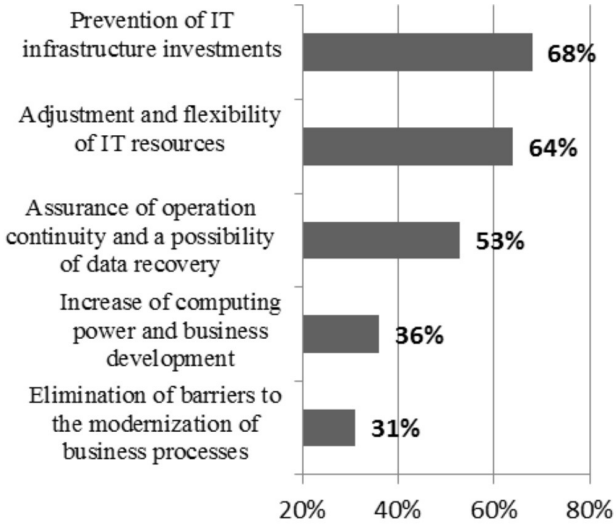


Fig. 2. Benefits of cloud computing model in SMEs [25]

important for them (64%), mostly due to the fact that the company do not need to handle software updates, new versions of products or any other necessary actions that should be undertaken to adapt systems to the changes in regulations (e.g. changes in tax rates). 53 % of respondents supported opinion that cloud computing offers business operation continuity and permits data recovery. The remaining part of the respondents (47%) is concerned with the efficiency of Internet connection, data transfer security and storage. The increase of computing power and business development (36%) and the elimination of barriers to modernization of business processes (31%) are not of a great importance to SMEs.

IV. Barriers in Cloud Computing Adoption

The managers who consider the change of IT resources model management from traditional one to cloud computing model have numerous less or more justified fears. Identification of these barriers is a subject of many research and scientific discussions [20], [26],[10].

In research conducted by IDC analysts [24.] the impact of 12 identified barriers was showed. In the assessment of managers the barriers which have the strongest impact are: legal jurisdiction, security and data protection, trust, data access and portability, data location, local support, change control, ownership and customization, evaluation of usefulness, slow internet connection, local language and tax incentives. As it was showed in the cited report none of the barriers cannot be accepted as the only most important one which was mentioned by most of respondents. There can be distinguished six strongly correlated barriers which are acknowledged as the most significant ones by 62% of respondents. They are: legal jurisdiction, security and data protection, trust, data access and portability, data location, local support [24.].

The relevance of the barriers to the cloud computing implementation was presented in Figure 3.

There are also interesting research results [26], where cloud computing providers were the respondents. Providers

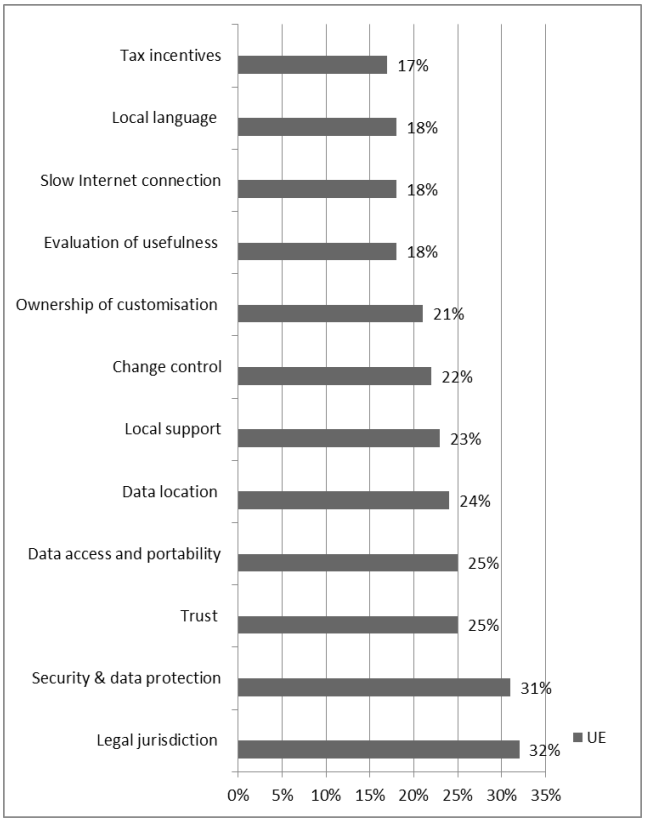


Fig. 3. The relevance of the barriers to the cloud computing implementation [24]

say that customers' main concern over switching to cloud is losing control – an issue voiced by almost half of all respondents (48%). As well, 39% say that data loss and privacy risks are a major worry. Worries of 41% customers concerned integration with existing architecture, a 28% were not sure the promise of cloud environment can be realized, 28% claimed that implementation of integration costs is too high, 27% worried of risk of intellectual property theft and lack of legal and regulatory compliance 22%.

Companies are afraid that their IT performance is controlled not by their own staff but by off-premises cloud providers; and that they may not be able to make necessary changes in application features easily and when needed [27].

Security concerning in particular privacy and data confidentiality, is one of the most cited objections to cloud computing [28].

In the light of above research in order to ensure cloud computing development technical, legal, organizational and mental barriers should be removed in the highest possible degree.

V. Development Of Cloud Computing In Poland

According to Business Software Alliance [29.] Poland takes 12th place among 24 countries in the ranking of government policies having influence on development of cloud computing. This ranking evaluates status in 24 countries, which are together responsible for 80 percent of world IT and communication resources and the policy of

these countries in 7 basic areas: data protection, security in cyberspace, cyber-crime, intellectual property, technological interoperatibility and law harmonization, free trade and IT infrastructure.

5 top places in the ranking were achieved by: Japan, Australia, Germany, USA and France.

Macroeconomics factors which positively influence on such high position of Poland are most of all [29]:

- good legal regulations in the scope of privacy protection, electronic signature, e-commerce, and cyber-crime which constitute a basis for building trust for cloud computing technology.
- Poland also possesses one of the most comprehensible systems of intellectual property protection. In 2008 legal regulations were amended in the scope of Internet provider responsibilities.
- Polish government promotes innovations and interoperatibility and apply non discriminating policy in the scope of governmental orders.

The factor influencing negatively on the position of Poland is still limited access to broadband Internet comparing to such countries as Germany or Japan.

Among microeconomics factors technological preferences of IT (CIO) managers have a big influence on implementation of cloud computing in Polish enterprises. In the table 1 were presented research results on the sample of 3000 (CIO) managers [30].

TABLE I.
TECHNOLOGICAL PREFERENCES OF IT CIO MANAGERS IN THE WORLD AND IN POLAND

Category	Global 2009	Global 2011	Poland 2011
Cloud Computing	33%	60%	49%
Mobile solutions	68%	74%	76%
Virtualization	75%	68%	81%

Presented in the table 1 research results indicate that technological preferences of CIO IT managers in Poland are strongly directed at virtualization (81%) and the application of mobile solutions (76%) and exceed percentage values in a global scale. Preferences in implementation of cloud computing model (49%) are lower than values in a global scale (60%), but comparatively to other countries are high.

Identification, analysis and trial of assessment of the rest factors having influence on development of cloud computing in Polish SMEs, with special consideration of factors having negative influence (barriers) will be presented in the next points.

VI. MENTAL BARRIERS AND COMPONENTS OF TRUST IN CLOUD COMPUTING

Many researchers claim that among barriers of cloud computing implementation the central place should belong to mentality, which can be in favor of open attitudes towards changes or it can constitute crucial barrier.

Swida-Zieba [31] defines mentality as a structure resistant to changes, deeply rooted, non verbalized and subliminal. It constitutes the result of the whole of socialization: system of situational incentives, own experiences, perception of reality and influence of education processes. It is formal and generalized structure, conditioning involuntary and spontaneous reactions on incentives and situations [31].

Mentality of managers is shaped in the process of interaction of many factors from social environment and their own experiences and in the future it has influence on their attitudes, leadership styles or undertaken decisions.

Mental barriers of managers manifest in natural perception reactions e.g. „cloud computing will not bring any benefits for the company", emotional e.g. „we do not trust cloud computing model" and behavioral e.g.. „cloud computing is not needed in the company and we are not going to implement it".

To best mitigate barriers to confidence, we need to understand the main components affecting cloud trust [32]:

- Security,
- Privacy,
- Accountability,
- Auditability

Security mechanisms (e.g. encryption) which make it extremely difficult or uneconomical for an unauthorized person to access some information [32].

Privacy - protection against the exposure or leakage of personal or confidential data (e.g. personally identifiable information (PII)) [33].

Accountability - the obligation and/ or willingness to demonstrate and take responsibility for performance in light of agreed-upon expectations, accountability goes beyond responsibility by obligating an organization to be answerable for its actions [32], [33].

Auditability – the relative ease of auditing a system or an environment. Poor auditability means that the system has poorly-maintained (or non-existent) records and systems that enable efficient auditing of processes within the cloud [32]. Auditability enables review of undertaken activities and stating responsibilities of persons (employees) or organization for undertaken activities, especially when they were not compliant with stated security policy.

Mental barriers are constituted also by insufficient knowledge of managers concerning notion of cloud computing, its advantages and disadvantages and risk connected with implementation of this solution. The current state of managers' knowledge on cloud computing is hardly to categorize as satisfactory one [20]. It causes ungrounded worries about: data location and connected with it data access and portability.

Managers very often do not perceive the needs of changes of current model of IT resources management for cloud computing model, because they assess its usability very low and the risk of introduction of changes is assessed very high.

VII. Mental Barriers To Adoption Cloud Computing. Research Results

In Poland exist 1.9 mln of enterprises, and over 1,8 mln is constituted by microenterprises (employing up to 9 employees), 55 thousand small enterprises (employing from 10 to 49 employees), 15 thousand medium sized enterprises (employing from 50 to 249 employees and 3,2 thousand of big enterprises (above 250 employees).

As far as research methodology is concerned it is worth mentioning that the study in the light of the above was carried out on the sample of 134 companies which are not representative to the entire Polish SME sector. Nevertheless in the authors' opinion, the results are an interesting attempt of building the image of Polish SMEs, which face the challenge of cloud computing implementation.

The aim of conducted research was to verify the following research hypotheses:

H1: Among mental barriers the biggest worries are connected with lack of trust for cloud computing model.

H2: Mental barriers connected with insufficient knowledge are not perceived by managers as „high barrier".

H3: Between "level of management" and perception of highlighted mental barriers occurs stochastically essential dependence.

The study involved: 80 micro-enterprises, 42 small-enterprises and 10 medium-enterprises from the following sectors: manufacturing - 39, trade - 65 and services – 30.

The study was conducted in January - February 2014 with the use of electronic questionnaire. Link to the survey was provided via email sent to the companies. Questionnaire return rate was 27%. In the research participated: 54 managers of strategic level, 48 managers of tactical level and 32 IT CIO's. The results of a few chosen survey questions are presented below.

In order to examine the relevance of each barrier to the cloud computing implementation decision, twelve questions from IDC report [24] research carried out on the sample of European enterprises were used.

- Security & data protection: "We are worried about the security and data protection guaranteed by cloud services"
- Trust: "It is difficult to judge which cloud services are"
- Data location: "We do not know and/or cannot control the location of our corporate data"
- Local support: "There is no local support for the services"
- Change control: "We cannot control software changes and upgrades made by the vendor"
- Ownership of customization: "We do not know who owns the customizations/changes we make to the cloud services"
- Evaluation of usefulness: "We do not know how to evaluate the usefulness of cloud service for our organization"
- Slow Internet connection: "Our Internet connection(s) is/are not reliable or fast enough"
- Local language: "There is no local language version of the services"
- Tax incentives: "Tax and other incentives make buying with capital more attractive than paying for what we use on subscription"
- Legal Jurisdiction: "If we have a dispute with the cloud service provider, I may have to go to court in another country inside the EU"
- Data Access and Portability ""Concern about our ability to move data from one vendor to another or onto our own IT"

Respondents rated each response on a scale: low barrier, average barrier, high barrier. Next all the barriers were grouped into three groups: mental, technical and legal barriers. In the opinion of respondents mental barriers are more frequent than technological or legal barriers perceived as a strong obstacle in the decision on whether or not to implement cloud computing solutions. Among all responses "high barrier" was indicated by as much as 38% of responses related to the mental barriers, 33% to the technical barriers and 29% to the legal barriers.

The next question involved more specific identification of mental barriers.

For the assessment of lack of trust barrier were used questions concerning: security and data protection, privacy, accountability and auditability. In the assessment of

Table II.
THE ASSESSMENT OF MENTAL BARRIERS INFLUENCE ON CLOUD COMPUTING ADAPTION

	low barrier	medium barrier	high barrier
	Trust		
Security and data protection	32	54	47
Privacy	33	50	51
Accountability	56	40	38
Auditability	51	40	43
	Lack of knowledge		
Data location	29	64	40
Data access and portability	19	70	46
	Lack of the need		
Evaluation of usefulness	50	57	27

insufficient knowledge barrier cloud computing model the respondents' answers concerning: data location and data access and portability were evaluated. In the assessment of lack of the need of change of IT resources model evaluation of usefulness was used.

In the table II was presented evaluation of mental barriers' influence on cloud computing adaptation.

In the table II the greatest number of „high barrier" indication obtained: privacy (51) and „security and data protection" (47). It means that respondents as the most important barrier perceive „lack of trust" for cloud computing.

Hypothesis H1 was positively verified.

The barriers connected with the lack of knowledge were most often assessed as „medium barrier", what confirmed rightness of H2 hypotheses.

In the further considerations „high barrier" indications of respondents were only taken into account. For „trust" barrier the number of „high barrier" indications was summed up for four analyzed of such components as: security and data protection, privacy, accountability and auditability. For „lack of knowledge" „high barrier" indications were summed up for data location and data access and portability.

In the correlation table presented in table III the responses of strategic, tactical and IT CIOs level managers were taken into account.

TABLE III.
CORRELATION TABLE

	Trust	Lack of know-ledge	Lack of the need	Sum-mary
Strategic managers	108 (93,79)	36 (45,06)	9 (14,15)	153
Tactical Managers	52 (66,82)	44 (32,10)	13 (10,08)	109
IT (CIO)	19 (18,39)	6 (8,84)	5 (2,77)	30
	179	86	27	292

In the correlation table quantities of indications were placed as well as numerical strengths in parentheses. Using the chi-square independence test (χ^2) we test the H0 hypothesis: concerning independence of investigated features in face of alternative H1 hypothesis: features are not independent.

This statistics has χ^2 out of (r-1)(k-1) layout of the degrees of freedom, where r - the number of columns, k - the linage. The calculated value of the statistics χ^2=17,11, the value from tables for χ^2=0,05 and 4 degrees of freedom χ^2_α = 9,45. Therefore the following condition is fulfilled:

χ^2=17,11 > χ^2_α = 9,45 and the hypothesis H0 can be rejected. This means that between the type of manager and perception of mental barriers connected with trust, lack of knowledge or evaluation of usefulness the stochastic dependence occurs.

For the measurement of the power of correlations of examined variable the coefficient of Czuprow convergence (T) was used. This coefficient accepts values from the [0,1] range. The nearer to the zero the coefficient of convergence is the dependence between variables is weaker. The calculated value of coefficient of convergence T=0,17 means that the dependence is weak, but statistically essential.

The obtained results allowed for positive verification of H3 hypothesis.

VIII. CONCLUSION

Presented research results among SMEs allowed to formulate the following conclusions:

1. Among mental barriers the biggest worries are connected with the lack of trust for cloud computing model.
2. Mental barriers connected with insufficient knowledge are not perceived by managers as „high barrier".
3. Between "level of management" and perception of distinguished mental barriers stochastically essential dependence occurs.
4. The trust as a high barrier was most often indicated by strategic level managers.
5. Lack of knowledge concerning cloud computing was most often assessed as „high barrier" by managers of tactical level.
6. IT CIOs recognize advantages of cloud computing model and very rarely assess "lack of implementation need" as „high barrier".

The key for liquidation of mental barriers rebounding negatively on attitudes of managers in the face of decision concerning implementation of cloud computing model can be consciously shaped program of general education, and the assurance of instruments by means of which it will be implemented. Operations should be directed at the wide formation of features and social competences, favoring open to changes attitudes. Not flexible enough attitudes of managers and management of specific levels, anxiety of undertaking risk and lack of interest in new solutions are consequences of low level of the needs understanding of modern information society.

REFERENCES

[1] „Sizing The Cloud", http://www.forrester.com/Sizing+The+Cloud /fulltext/-/E-RES58161 (accessed on: 1 April 2014).

[2] Q. Li, C. Wang, J. Wu, J. Li, Z.-Y. Wang, "Towards the business-information technology alignment in cloud computing environment: An approach based on collaboration points and agents". *International Journal of Computer Integrated Manufacturing*, vol. 24, no. 11, pp. 1038–1057, November 2011.

[3] R. King, "Cloud computing: Small companies take flight". *BusinessWeek Online*,4, 2008. http://www.businessweek.com/ technology/content/aug2008/tc2008083 619516.htm (accessed on: 1 April 2014)

[4] C. Russell, F. Jeff, J. Norm, M. Seanan, P. Carolyn, S. Patrick, J. Stanley, "Cloud Computing in the Public Sector: Public Manager"s

Guide to Evaluating and Adopting Cloud Computing", Cisco Internet Business Solutions Group 2009.
[5] C. Tsaravas, M. Themistocleous, "Cloud Computing and eGovernment: a Literature Review", European, Mediterranean &Middle Eastern Conference on Information Systems, Greece, pp. 154-164, 2011.
[6] D. Jelonek, C. Stępniak, T. Turek, "The Concept of Building Regional Business Spatial Community". *In: ICETE 2013. 10th International Joint Conference on e-Business and Telecommunications*. *Proceedings*. 29-31 July 2013, Reyklavik, Iceland 2013.
[7] N. Sultan, "Cloud computing for education: A new dawn?" *International Journal of Information Management*, vol. 30, pp. 109–116, 2010
[8] N. Sultan," Making use of cloud computing for healthcare provision: Opportunities and challenges*", International Journal of Information Management,* vol. 34, pp. 177–184, 2014
[9] E. Savitz, W. Vogels, "How the cloud changes businesses big and small". Forbes.com, vol. 14, 2012.
[10] P. Gupta, A. Seetharaman, J. R. Raj, "The usage and adoption of cloud computing by small and medium businesses", *International Journal of Information Management*, vol. 33, pp. 861– 874, 2013.
[11] S. Mahesh, B. J. L. Landry, T. Sridhar, K. R. Walsh, "A decision table for the cloud computing decision in small business". *Information Resources Management Journal*, vol. 24, no. 3, pp. 9–25, July–September 2011.
[12] M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, "Above the Clouds: A Berkeley View of Cloud Computing", Technical report No. UCB/EECS-2009-28 University of California at Berkley, USA, 2009.
[13] W. Kim, "Cloud computing: Today and tomorrow". *Journal of Object Technology*, vol. 8, no. 1, pp. 65–72, 2009.
[14] S. Bhardwaj, L. Jain, and S. Jain, "Cloud computing: A study of infrastructure as a service*", International Journal of engineering and information Technology,* vol.2, no. 1, pp.60-63, 2010.
[15] C. Madhavaiah, I. Bashir, S. I. Shafi, "Defining Cloud Computing in Business Perspective: A Review of Research", MDI SAGE Publications, Vision, vol. 16, no. 3, pp.163–173, 2012.
[16] R. L. Grossman, Y. Gu, "On the varieties of clouds for data intensive computing". *IEEE Computer Society Bulletin of the Technical Committee on Data Engineering*, vol.32, no. 1, pp. 44–51, 2009.
[17] R. Buyya, C. S. Yeo, S. Venugopal, "Market-oriented cloud computing: Vision, hype and reality for delivering IT services as computing utilities". *Proceedings of the 10th IEEE International Conference on High Performance Computing and Communications*, Dalian, China, 25–27 September 2008.
[18] P. Mell, T. Grance, "The NIST Definition of Cloud Computing. Recommendations of the National Institute of Standards and Technology", NIST Special Publication 800-145, September 2011.
[19] V. Chang, R. J. Walters, G. Wills," The development that leads to the Cloud Computing Business Framework", *International Journal of Information Management*, vol. 33, pp. 524-538, 2013.
[20] "Cloud computing: Flexibility, Efficiency, Security". Report ThinkTank, Microsoft 2011.
[21] S. Marston, Z. Li, S. Bandyopadhyay, J. Zhang, and A. Ghalsasi,. "Cloud computing—The business perspective". *Decision Support Systems*, vol. 51, no. 1, pp. 176–189, April 2011.
[22] F. Etro, "The economics of cloud computing". *IUP Journal of Managerial Economics,* vol. 9, no. 2, pp. 7–22, 2011.
[23] "Cloud Computing in financial sector", Report Banking Technologies Forum at Polish Bank Union 2013.
[24] *Quantitative Estimates of the Demand for Cloud Computing in Europe and the Likely Barriers to Up-take*, SMART 2011/0045, D4- Final Report, IDC, 2012
[25] "An SME perspective on Cloud Computing. Survey", The European Network and Information Security Agency (ENISA), November 2009. http://www.enisa.europa.eu/ (accessed on: 10 March 2014).
[26] KPMG International's 2012 Global Cloud Providers Survey,
[27] N. Leavitt, "Is cloud computing really ready for prime time?" Computer, vol. 42, no. 1, pp.15–20, 2009.
[28] Q. Zhang, L. Cheng, R. Boutaba, "Cloud computing: State-of-theart and research challenges", *Journal of Internet Service Application*, vol. 1, no. 1, pp. 7–18, 2010
[29] "2013 BSA Global Cloud Computing Scorecard. A Clear Path to Progress", BSA The Software Alliance: IBM CIO Study 2011
[30] . "The Essential CIO. Insights from the Global Chief Information Officer Study", IBM Institute for Business Value, 2011.
[31] H. Swida-Zieba, Mechanisms of social enslavement Reflections at the decline of formations., The University of Warsaw, Warsaw 1990.
[32] R. K L Ko, P. Jagadpramana, M. Mowbray, S. Pearson, M. Kirchberg, Q. Liang, B. S. Lee, "TrustCloud: A Framework for Accountability and Trust in Cloud Computing", *The 2nd IEEE Cloud Forum for Practitioners (IEEE ICFP 2011),* Washington DC, USA, July 7-8, 2011.
[33] . S. Pearson and A. Charlesworth, "Accountability as a way forward for privacy protection in the cloud," *Cloud Computing*, 2009, pp. 131-144.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1259–1267
DOI: 10.15439/2014F121
ACSIS, Vol. 2

Assessing the quality of e-government portals – the Polish experience

Ewa Ziemba
University of Economics
ul. 1 Maja 50, 40-287 Katowice, Poland
Email: ewa.ziemba@ue.katowice.pl

Tomasz Papaj
University of Economics
ul. 1 Maja 50, 40-287 Katowice, Poland
Email: tomasz.papaj@ue.katowice.pl

Danuta Descours
Silesia Centre of Information Society
ul. Powstańców 34, 40-954 Katowice, Poland
Email: ddescours@e-slask.pl

***Abstract*—The transition from a government to e-government poses continuous challenges in employing increasingly sophisticated web portals as the gateway to government units, their information and services. The high quality of those portals is needed for the successful adoption of e-government. Therefore, the aim of this study is to assess quality of selected Polish e-government portals based on a proposed framework. This framework employs the International Organisation for Standardisation (ISO) standard. Firstly, the paper presents various definitions of quality and different theories/models for assessing the quality of e-government portals. Secondly, the framework for assessing the quality of e-gov portals is presented. Thirdly, the assessment of Polish e-government portals based on the proposed framework is shown. The paper concludes with a discussion of research findings and recommendations for studies on e-gov portals' evaluation. Finally, the future works are submitted. The obtained research findings will prove to be helpful for researchers in developing studies on e-government, especially in the research issue of e-gov portals. Moreover, they can be useful while undertaking empirical activities aimed at the e-government adoption.**

I. Introduction

THE research on e-government has a relatively short history. Analyzing the three most known bibliographic databases, that are EBSCO, ISI Web of Knowledge and Scopus, reveals that scientists have started exploring e-government research since the mid 1990s. By contrast, practical e-government initiatives have been launched since the late 1990s [1]. The European Union countries, including Poland, have written into their strategic planning the building of e-government since 1999 [2], [3], [4], [5], [6], [7], [8], [9]. Studies and empirical activities aimed at e-government have strongly been developing from 2000.

Generally, e-government means using information and communication technologies (ICTs) for [10], [11], [12], [13], [14]:

- supporting processes in government units;
- delivering e-government services at different levels of maturity to government stakeholders (i.e. citizens, enterprises and government units);
- improving government transparency, citizen's participation, and democratic decision making; and
- cooperation, networking, and maintaining partnership relations between government stakeholders.

Studies on e-government focus on the variety of issues relating to the above picture of e-government. They look at e-government from different angles. An important research issue relates to e-government maturity [15], [16], [17]. The maturity levels of e-government reveal the degree of technological sophistication and the degree of organizational transformation in a government. They reflect how enterprises and citizens can interact with government units and how government units can cooperate and communicate. The European model of e-government maturity is comprised of the following levels: information, one-way interaction (downloadable application forms), two-way interaction (electronic application forms, e-forms), transaction (full electronic) and personalization (targetisation/automation) [18], [19], [20]. E-government services (e-gov services) are another important issue of e-government research. Those services are investigated from different perspectives, e.g. their implementation [21], [22], [23] and acceptance by all government stakeholders [15], [24]. Furthermore, researchers are investigating factors that play key roles in successful adoption of e-government [25], [26], [27], [28], [29], [30], [31], [32], [33], [17]. Additionally, studies focused on implementation and acceptance of e-government portals (e-gov portals, e-government web sites) are carried out [34], [35], [36]. E-gov portals provide a single point of access to government services via the Web-enabled interface. Such portals deliver convenient online access to e-gov services for government stakeholders. Thanks to them, enterprises and citizens can interact with government units as well as government units can cooperate and communicate. Therefore, e-gov portals strongly influence successful adoption of e-government [36].

A review of the literature on e-gov portals reveals that research on evaluating them and assessing their quality is very limited. To address this gap, this paper aims to assess the quality of selected Polish e-gov portals based on a comprehensive framework. This framework employs the International Organisation for Standardisation (ISO) standard.

This paper is structured as follows. Firstly, the paper presents various definitions of quality and different theories/models for assessing the quality of e-gov portals. Secondly, the framework for assessing the quality of e-gov portals is presented. Thirdly, the evaluation of Polish e-gov portals based on the proposed framework is shown. The paper concludes with a discussion of research findings and recommendations for studies on e-gov portals' evaluation. Finally, the future works are submitted.

978-83-60810-58-3/$25.00

II. THEORETICAL BACKGROUND

A. Quality of e-government portals

The term "quality" means characteristic, feature, trait. It rarely relates to one thing or a phenomenon. Often, it refers to a sum of these characteristics against requirements. There is not a single definition of quality. The reason is that quality is basically only "the perception of quality" [37]. Many known experts have contributed to comprehensive quality definitions.

Garvin [38] defined and Kitchenham and Pfleeger [39] expanded five different perspectives on quality:

- The transcendental perspective tackles the metaphysical aspect of quality. From this standpoint, it is "something toward which we strive as an ideal, but may never implement completely;"
- The user perspective pertains to the suitability of the product for a stated context of use. In comparison, the transcendental view is volatile, the user view is more specific, based on the product attributes that satisfy user's needs;
- The manufacturing perspective indicates quality as conformable with requirements. This view of quality is stressed by ISO 9000, which defines quality as "the degree to which a set of inherent characteristics fulfills requirements" [40];
- The product perspective indicates that quality can be acknowledged by measuring the intrinsic attributes of the product; and
- The value-based perspective assumes that the different perspectives of quality may be approached differently by various stakeholders, especially with regard to its importance or value.

Quality is relevant for both products and services. This just refers to issues of the e-gov portals quality. The American Society for Quality perceives quality as "the totality of features and characteristics of a product or service that bear its ability to satisfy stated or implied needs" [41, p. 615]. Griffin referred to basic dimensions explored by Garvin, which determine the quality of a particular product or services. These dimensions are: performance, features, reliability, conformance, durability, serviceability, aesthetic, and perceived. For Juran quality is "fitness for use", and Deming believed that "quality should be aimed at the needs of the customer, present and future." Crosby said that quality is "conformance to requirement" [37]. Feigenbaum defined quality as "a customer determination based upon a customer's actual experience with a product or service, measured against his or her requirements – stated or unstated, conscious or merely sensed, technically operational or entirely subjective – and always representing a moving target in a competitive market" [42, p. 1].

In this research the quality of e-gov portals is considered as software product quality. However, a straightforward definition for software product quality is difficult to formulate. Depending on who is evaluating the quality of software product it can be either good or bad. For instance, to an end user, if a software product offers efficient and necessary functionalities to perform the task it was developed for, then software quality is good. A different perspective may have a software developer for whom maintainability or testability, how easy it is to maintain and eliminate bugs, are the main attributes of a good quality software product. And finally, a software architect can perceive good software quality from, for example, the angle of the reusability of the used software components or the scope and quality of the descriptive literature of the software product.

B. Models for assessing the quality of e-government portals

Approaches to assessing the quality of e-gov portals have been explored for a few years [43], [44], [45], [46], [47]. In the work on them, researchers indeed mainly examined, improved and adopted models, such as D&M [48], [49], [50], [51], TAM [52] and Wang's [53].

More generally, a framework for assessing the quality of e-gov portals covers three aspects of quality: system, information and service quality. System quality can be viewed as a measure of e-gov portals' functionalities. It is comprised of four constructs: usability, responsiveness, ease of access and privacy. Information quality is defined as the measure of the value which the information provides to a user. More specifically information quality can be described by four constructs: accuracy, dependability, coverage and ease of use. Service quality focuses on four others: empathy, interactivity, playfulness and aesthetic [36].

Wang and Liao [46] conducted an empirical test of D&M model adoption in the context of G2C e-government in Taiwan. Except for the link from System Quality to System Use, the hypothesized relationships between the six success variables were significantly or marginally supported by the data. The authors emphasized that "researchers can also use the validated model as the foundation for developing comprehensive e-government systems success measures and theories, exploring relationships between the proposed constructs, and comparing e-government success empirical studies" [46].

Almalki, Duan and Frommholz [43] suggested a conceptual framework for assessing e-gov portals' success which integrates the updated D&M model, TAM, self-efficacy theory and perceived risk. In this model 13 constructs are indentified for measuring success of e-gov. portals. Those are: system quality, information quality, service quality, computer self-efficacy, perceived risk, personal values, perceived usefulness, perceived ease of use, attitude towards using, behavior intention to use, user satisfaction, net benefits.

The next model is guided by both TAM, D&M model and the policies framed by the government of India. It includes dimensions for assessing e-service quality of government portals, such as quality dimensions: citizen centricity, usability, technical adequacy, privacy and security, usefulness of information, transaction transparency, comprehensive information, interaction [44].

Papadomichelaki and Mentzas conceptualized an e-government service quality model, eGovQual, where they indicate six constructs: ease of use, trust, functionality of the interaction environment, reliability, content and appearance of information, interactivity interaction [44].

Some researchers showed that navigation facility and accessibility are important in determining citizens' perceived system quality. Whereas information preciseness, timeliness, and sufficiency were found to be key measures in information quality in government e-services [54].

C. Basis of proposed framework for assessing the quality of e-government portals

The proposed framework for assessing the quality of e-gov portals is primarily built on the ISO/IEC 25010 [55]. This standard was chosen because of its breadth and completeness, and because of the prestige of the organization.

According to ISO/IEC 25010, the quality of a software product is "the degree to which the system satisfies the stated and implied needs of its various stakeholders, and thus provides value" [55]. The SQuaRE series of International Standards by quality models is based on those declared and indicated needs: the quality in use model and the product quality model in this International Standard, and the data quality model in ISO/IEC 25012. In this study the product quality model was employed. It defines eight main dimensions and related sub-dimensions for assessing software products (Table 1). Factors (items, constructs, metrics) in each sub-dimension should be defined individually according to the nature of assessed software product. The stated quality factors can build a checklist for ensuring a thorough handling of quality requirements, thus founding a basis for activities that will be necessary during software product development.

The ISO/IEC 25010 software product quality model relates to four layers of quality assessment explored by Halaris et al [45]:

- back-office process performance layer, addressing factors mainly found in quality of back-office operations (this layer relates mainly to the ISO/IEC *Compatibility* dimension);
- portal technical performance layer, addressing the factors of the technical performance of the portal (this layer relates mainly to the ISO/IEC dimensions: *Reliability*, *Security*, *Maintainability*, *Portability*, *Performance efficiency*);
- portal quality layer, addressing the factors of the portal usability (this layer relates mainly to the ISO/IEC dimensions: *Usability*, *Functional suitability*); and
- customer's overall satisfaction layer, addressing the overall level of quality perceived by the user against user's expectations (this layer relates mainly to the ISO/IEC dimensions: *Usability*, *Functional suitability*, *Security*, *Maintainability*, *Portability*, *Performance efficiency*).

TABLE I.
ISO/IEC 25010 SOFTWARE PRODUCT QUALITY MODEL

Quality dimensions	Quality sub-dimensions	Description
Functional suitability	Functional completeness	The software product's ability to provide functions and operations which are required to fulfill stated and implied users' needs
	Functional correctness	
	Functional appropriateness	
Performance efficiency	Time behavior	The software product's ability to offer sufficient efficiency and using reasonable amount of resources
	Resource utilization	
	Capacity	
Compatibility	Co-existence	The software product's ability to be interoperable with other software products
	Interoperability	
Usability	Appropriateness recognisability	The software product's ability to be easy to use, learnable and understandable
	Learnability	
	Operability	
	User error protection	
	User interface aesthetics	
	Accessibility	
Reliability	Maturity	The software product's ability to perform specified function under specified conditions for a specified period of time with the minimum crashes possible
	Availability	
	Fault tolerance	
	Recoverability	
Security	Confidentiality	The software product's ability to secure its internal information so that no unauthorized usage is possible
	Integrity	
	Non-repudiation	
	Accountability	
	Authenticity	
Maintainability	Modularity	The software product's ability to be changeable, maintainable and updatable
	Reusability	
	Analysability	
	Modifiability	
	Testability	
Portability	Adaptability	The software product's ability to be portable from one software (hardware) environment to another
	Installability	
	Replaceability	

Source: based on (ISO/IEC, 2011)

The third model employed in the proposed framework for assessing the quality of e-gov portals was constituted and utilized by IT Project Centre in Poland [56]. This model focuses on usability of e-gov portals and includes 96 factors grouped into the following five dimensions: interface cohesion and user friendliness, pages design, navigation, content, e-gov services form, search engine, main page. Some factors of the IT Project Centre have been used in the *Usability* dimension.

III. RESEARCH METHODOLOGY

This study is a part of research on the holistic and systems approach to the e-government adoption in the context of sustainable information society [57]. Various methods and techniques have been applied in this research at three steps:

- The first step – a review of literature was conducted to explore various definitions of quality and different models for assessment of the e-gov portals;
- The second step – the framework for assessing quality of e-gov portals was established on the basis of literature findings, empirical observations, brainstorming and methods of creative thinking and logical deduction; and
- The third step – the pilot study (feasibility study, experimental trial) was carried out to verify and test the proposed conceptual framework. It was a small-scale, short-term experiment that helped pre-test the reliability and importance of indicated factors in assessing e-gov portals. The study was conducted in December, 2013. Three Polish e-gov portals that can be examples of "good practices" were evaluated. It was a self-assessment; each portal was evaluated by the qualified employees of the appropriate government unit. To classify each factor, the respondents had to carefully read and analyze their portal. E-gov portals' quality has been assessed on a two-point scale: 1 – the capability has been implemented correctly or the capability has been implemented partially; and 0 – the capability has not been implemented. All factors of the framework are equally important and none of them has been prioritized. The final evaluation of each portal is defined as the total sum of ratings of subsequent factors. Microsoft Excel was used to conduct the entire analysis. In the final step of our study, on the basis of creative thinking and logical deduction, recommendations on the evaluation of e-gov portals were formulated.

IV. Research findings

A. The proposed framework for assessing the quality e-government portals

In the proposed framework, a quality of e-gov portals is understood as their capability to satisfy stated and implied needs for the e-gov portal to be used under specified conditions. The dimensions and sub-dimensions of this framework were adopted from ISO standard. For each sub-dimension, the capability of e-gov portals is determined by a set of factors that can be measured. 89 factors were indicated that include 14 factors within *Functional suitability*, 6 factors – *Performance efficiency*, 6 factors – *Compatibility*, 27 factors – *Usability*, 11 factors – *Reliability*, 6 factors – *Security*, 13 factors – *Maintainability*, 6 factors – *Portability*. Detailed enumeration of these factors is presented in [58]. The dimensions, sub-dimensions and factors are structured in the framework as shown in Table 2.

The quality of e-gov portals can be evaluated by all government stakeholders, i.e. government units, citizens and enterprises. Maintainability concentrates on government units' self-assessment, while other dimensions put emphasis on citizens' and enterprises' as well as government employees' evaluation.

Table II.
THE CONCEPTUAL FRAMEWORK FOR ASSESSING THE QUALITY OF E-GOV PORTALS

Dimensions/ Sub-dimensions	Factors descriptions (capabilities)
1. Functional suitability (14 factors)	
1.1. Completeness	Delivering e-gov services at the different levels of maturity
1.2. Correctness	Correct operation of e-gov services
1.3. Appropriateness	Matching e-gov services to the current and future needs of stakeholders
2. Performance efficiency (6 factors)	
2.1. Time behaviour	Shortened time required to the settlement of the matter
2.2. Resource utilization	Reducing consumption of resources (e.g. paper, toner), decrease in employment in government units; reduction in various transaction costs
2.3. Capacity	Lack of restrictions for users while using portal (e.g. time, number of documents); swift loading and running
3. Compatibility (6 factors)	
3.1. Co-existence	Integration between portal, back-office systems and other e-gov portals; updated specification sheet comprising necessary information to integrate portal with other systems and portals
3.2. Interoperability	Unified and consistent layout and content of mandatory information about government units and e-gov services; standardized electronic forms for all e-gov services
4. Usability (27 factors)	
4.1. Appropriateness recognisability	Different assistance methods for users, e.g. help module, assistance available at every level of portal use, electronic courses, wizard for filling out electronic forms, advanced search engine accessible from every level of portal use, virtual adviser
4.2. Learnability	Intuitive operating of portal
4.3. Operability	Site map; at any time users know where they are and may return to any portal page at any time; not more than nine submenu level created for a menu
4.4. User error protection	Portal is protected against admitting wrong (incorrect) data; users are kept informed on the errors and the need to correct them; suggestions for users how to improve errors
4.5. User interface aesthetics	Defining the look of portal by users; portal pages are divided into clear blocks (sections); consistent layout and navigation on pages; highlighted important information (e.g. names of sections, names of e-gov services, names of government units); clearly formatted contents of pages; graphic signs (icons) facilitating portal use; clear and understandable descriptions of e-gov services; comprehensible electronic forms for e-gov services
4.6. Accessibility	Adjusting font sizes to users requirements; special version of portal for the visually impaired and the blind; foreign language versions of portal
5. Reliability (11 factors)	
5.1. Maturity	Portal is protected against non-standard behavior of users; certified information security

Dimensions/ Sub-dimensions	Factors descriptions (capabilities)
	management system according to PN-ISO/IEC 27001:2007 standard has been implemented
5.2. Availability	Portal accessible to the users 7 days a week and 24 hours a day; failure-free of portal working; delivering information on planned breaks in portal operations
5.3. Fault tolerance	In the event of non-critical errors on the portal, its other functions can still be used; in case of errors when sending the completed application form – the application is saved and it is available to the user after the next login
5.4. Recoverability	There is procedures in cases of portal failure to recover portal after failures and to create the archive data entered through the portal; there is a redundant portal (replacement) in the event of failure
6. Security (6 factors)	
6.1. Confidentiality	Users password is masked; required cyclical change of the password by users; encrypted the connection with the portal
6.2. Integrity	There is a procedure of data protection
6.3. Non-repudiation	For the study was combined with Accountability
6.4. Accountability	All of the actions on the portal are identifiable by the user (e.g. first name, surname) and the time when an action was completed (making history of actions undertaken)
6.5. Authenticity	Checking the complexity of the users' passwords from the viewpoint of safety standards
7. Maintainability (13 factors)	
7.1. Modularity	Modular construction of portal; updated specification sheet of relationships between portal modules
7.2. Reusability	Portal modules implemented in its different parts or other portals
7.3. Analysability	Easy analysis for change of modules and its impact on other modules or other integrated portals; informing about changes of module that affect other modules or integrated portals
7.4. Modifiability	Portal's functionalities can be modified and extended according to users' new needs and in accordance with the new letter of the law; users are informed about the availability of new portal functionalities
7.5. Testability	There are: a portal test environment, procedures for portal testing, procedures for changes of the production version of portal, procedures for returning to the previous version of portal; subsequent versions of portal are identifiable
8. Portability (6 factors)	
8.1. Adaptability	There is a mobile version of portal; portal can be used by any web browser and by any operating system
8.2. Installability	Use of portal does not require the purchase of specialist software and the installation of special software
8.3. Replaceability	Users are informed about the need to install the free software and its new version necessary for the proper functioning of the portal

B. Assessing the quality of e-government portals using the proposed framework

The cases of Polish e-government portals

Three Polish e-gov portals that can be examples of "good practices" were evaluated by using this framework. Those were: SEKAP (www.sekap.pl), Digital Malopolska (http://www.wrotamalopolski.pl) and the Gate of Podlasie (http://cu.wrotapodlasia.pl). Currently, all these portals are being improved and this research can help the local government authorities to manage improvement activities.

SEKAP is a result of strategic, innovative projects, that were carried out by the municipal and district authorities of the Silesian Voivodeship in 2005-2008 and 2009-2012 [13], [20]. It enables the provision of e-gov services, including five forms of the relations between government units and their stakeholders: C2G/G2C, B2G/G2B, G2G. The e-government services are provided at different levels of maturity, from the information level to the transaction level. Currently, 122 government units provide e-gov services through SEKAP to citizens, enterprises and other government units. SEKAP includes different kinds of 650 various e-gov services, while the number e-gov services delivered by an individual government unit is 27,264.

Digital Malopolska is a result of a strategic, innovative project that was carried out by the municipal and district authorities of the Lesser Poland (Malopolska) Voivodeship in 2004-2006. The e-gov services for government stakeholders are provided at different levels of maturity, from the information level to the transaction level. Currently, 187 government units provide e-gov services through Digital Malopolska to citizens, enterprises and other government units. Digital Malopolska includes different kinds of 128 various e-gov services, while the number e-gov services delivered by an individual government unit is 23,936.

The Gate of Podlasie is a result of a strategic, innovative project that was carried out by the municipal and district authorities of the Podlaskie Voivodeship and ended in 2006. The e-gov services are provided at different levels of maturity, from the information level to the transaction level. Currently, 97 government units provide e-gov services through the Gate of Podlasie to citizens, enterprises and other government units. The Gate of Podlasie includes different kinds of 1,060 various e-gov services, while the number e-gov services delivered by an individual government unit is 3,000. Currently, this e-gov portal is under reconstruction.

Assessing Functional suitability

All examined e-gov portals received very low scores in *Functional suitability* dimension (Figure 1). This mainly results from incomplete catalogue of e-gov services at the appropriate levels of maturity. In both cases of Digital Malopolska and the Gate of Podlasie the users of the portals pointed out at some malfunctions of the available e-gov services and the need to deliver new ones. Correct operation of the available e-gov services and the need to create new e-

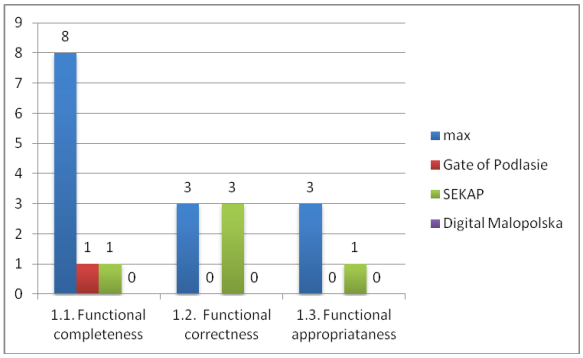


Fig. 1. Assessing *Functional suitability* of the Polish e-gov portals

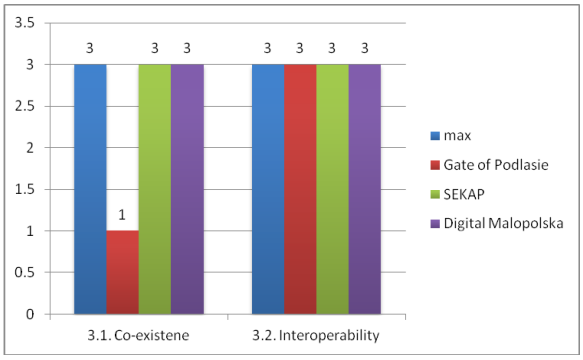


Fig. 3. Assessing *Compatibility* of the Polish e-gov portals

gov services were reported by e-gov employees to the SEKAP portal.

Assessing Performance efficiency

In *Performance efficiency* dimension, Digital Malopolska, received – 83% of all points, SEKAP – 67%, the Gate of Podlasie – 33% (Figure 2). All respondents indicated that the time of settling the matter initiated electronically is not shorter than of the matter initiated in a traditional way. Moreover, in both cases of Digital Malopolska and SEKAP there was smaller consumption of resources (paper, toner, etc.) and lower costs when dealing electronically than traditionally. In addition, in the case of these two portals there are no restrictions on their use, and government stakeholders did not report their observations on their operation being too slow.

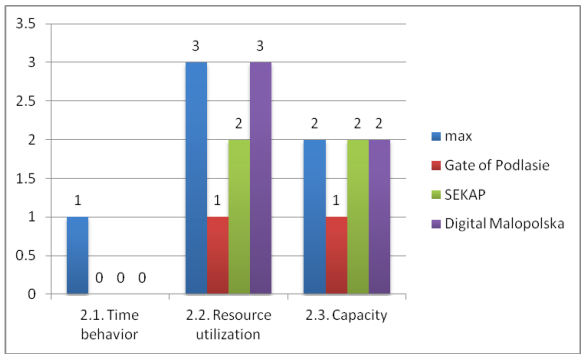


Fig. 2. Assessing *Performance efficiency* of the Polish e-gov portals

Assessing Compatibility

Digital Malopolska and SEKAP received 100% points in *Compatibility* dimensions, the Gate of Podlasie – 67% (Figure 3). All portals are integrated with different e-gov portals and back-office information systems (including different document management systems). Moreover, the documentation required to complete the integration is available. On all examined e-gov portals, information on e-gov services, information about e-government units and application forms is standardized.

Assessing Usability

Digital Malopolska received 85% of all points in *Usability* dimension, SEKAP – 74%, the Gate of Podlasie – 56% (Figure 4). All examined e-gov portals have help modules, providing assistance when filling out forms, and informing users about the errors made by them. They are protected against admitting wrong (incorrect) data. All examined e-gov portals have intuitive interface, easy navigation systems, not more than nine menu (submenu) levels. Only Digital Malopolska makes a portal map and e-learning on using the portal available. The portals do not have virtual advisers, and do not allow users to configure the layout of any portal (customization). Also, none of the portals is accessible for the blind, they do not offer voice messages, and there are also no foreign language versions of the portals.

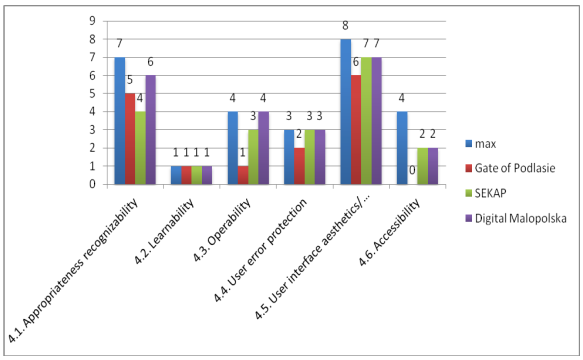


Fig. 4. Assessing *Usability* of the Polish e-gov portals

Assessing Reliability

SEKAP and Digital Malopolska received 91% of all points in *Reliability* dimension, the Gate of Podlasie – 73% (Figure 5). All portals are protected against non-standard behavior of users. It is worth emphasizing that the government units providing e-gov services on SEAKP and Digital Malopolska implemented certified information security management systems according to PN-ISO/IEC 27001:2007 [59]. All portals are accessible to users 7 days a week and 24 hours a day. Moreover, information on planned breaks in portals operations is delivered. In the event of non-critical errors on the portals, their other functions can still be used. There are procedures in cases of portals failures; to recover portals after failures and to create the archive data entered through the portals. SEKAP does not have a redundant portal (replacement) in the event of failure, while Digital Malopolska does not provide the re-send request after the failure of the portal.

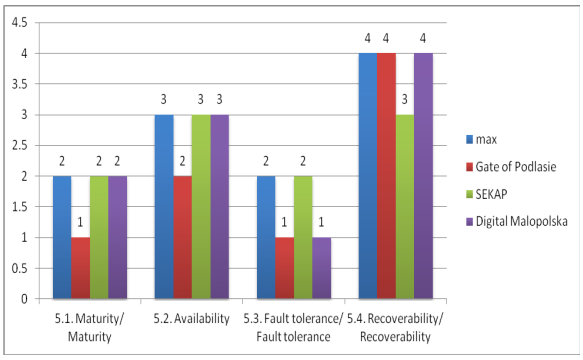


Fig. 5. Assessing *Reliability* of the Polish e-gov portals

Assessing Security

In *Security* dimension, Digital Malopolska received 100% of all points, whereas the Gate of Podlasie and SEKAP – 83% (Figure 6). Users' passwords are masked and the connection with the portals is encrypted. SEKAP and the Gate of Podlasie do not enforce the cyclical change of the password to the user's accounts. For all portals, procedures of data protection have been drawn. All of the actions on the portals are identifiable by the user (e.g. first name, surname) and the time when actions were completed. The complexity of the users' passwords from the viewpoint of safety standards is checked on the portals.

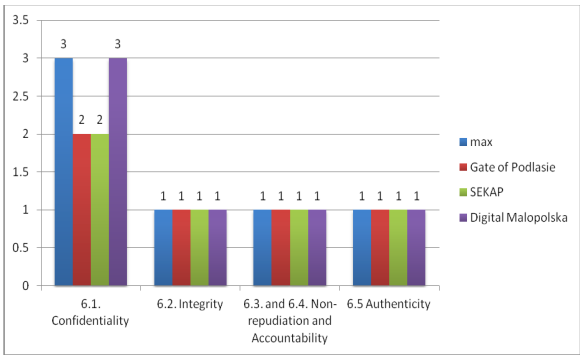


Fig. 6. Assessing *Security* of the Polish e-gov portals

Assessing Maintainability

In *Maintainability* dimension, Digital Malopolska received 100% of all points, SEKAP – 92% and the Gate of Podlasie – 31% points (Figure 6). All portals are made of modules and there are updated specification sheet of relationships between portal modules. Furthermore, the functionalities of the portals can be modified and extended according to users' new needs and in accordance with the new letter of the law.

Assessing Portability

In *Portability* dimension, Digital Malopolska received 83% of all points, whereas the Gate of Podlasie and SEKAP – 67% (Figure 7). All portals can be used by any web browser and by any operating system. There are not mobile versions of SEKAP and the Gate of Podlasie. The use of portals does not require the purchase of specialist software and the installation of special software.

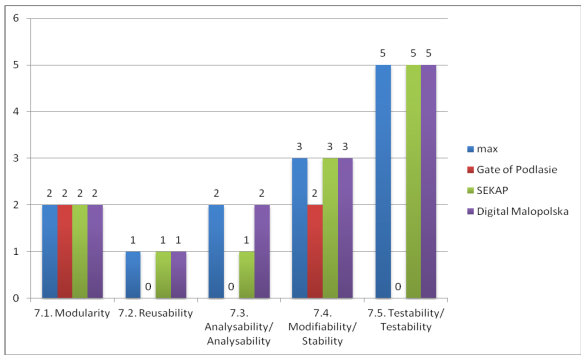


Fig. 7. Assessing *Maintainability* of the Polish e-gov portals

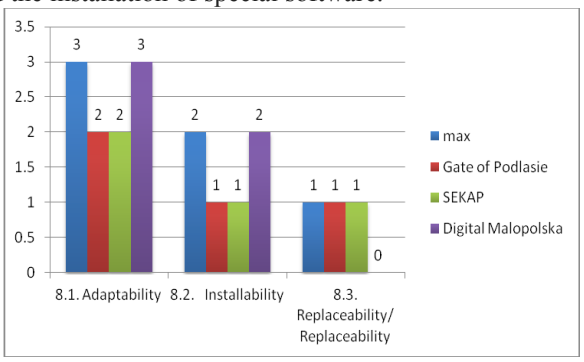


Fig. 8. Assessing *Portability* of the Polish e-gov portals

Overall assessing the quality of the Polish e-government portals

The study showed that Digital Malopolska portal meets the factors identified in the proposed framework (76%) to the greatest extent, the next one is SEKAP (74%), and the last – the Gate of Podlasie (48%). All in all, out of 89 points, the examined e-gov portals received sequentially – 68, 66 and 43 points (Figure 8). The lowest position of the Gate of Podlasie portal may be the result of its reconstruction. The portals have received the highest notes in *Compatibility*, *Reliability*, *Security* and *Portability* dimensions. They have been assessed worst in *Functional* suitability.

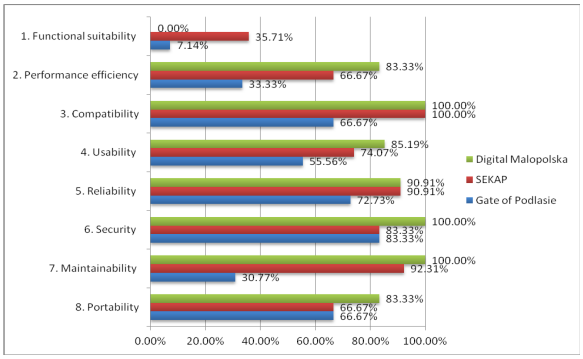


Fig. 9. The total assessment of Polish e-gov portals

V. Conclusion

So, in this paper a framework for the quality evaluation of e-gov portals based on International Organisation for Standardisation (ISO) standard is proposed. The aims of this research were to build a framework for a quality evaluation of e-gov portals and employ it in comparing three popular Polish e-gov portals. This framework can become a platform for developing a quality assessment of e-gov portals. Additionally, this work contributes to extant research by suggesting a framework based on ISO standard. The ISO was chosen because of its breadth and completeness, and because of its prestige.

Evaluation of three Polish e-gov portals confirmed each of the eight dimensions i.e. *Functional suitability*, *Performance efficiency*, *Compatibility*, *Usability*, *Reliability*, *Security*, *Maintainability* and *Portability* can be a good measure for e-gov portals quality. The identified factors measure the quality of e-gov portals from different points of views. At the same time, they show how e-gov portals should be improved.

This study provides several implications for research and empirical activities. It indicates an instrument which can identify the e-gov portals' features that are important for their quality. Researchers and scholars who develop studies on e-gov portals evaluation could find significant guidelines in this paper. They can use the instrument more confidently in their own research on e-gov portals and the successful e-gov adoption. Moreover, for practitioners, the results of this study can be used to undertake empirical activities aimed at evaluation and improvement of e-gov portals. Moreover, the framework can constitute important support for the procurement processes. The specified factors can be used to create the terms of reference for e-gov portal projects.

One limitation of this study was the small sample size, which is a concern because of reliability and validity issues. While this did not constrain the data analysis, a larger and more representative sample may yield more useful results. Another limitation was the verified and tested framework by e-gov employees only, while some factors of *Functional suitability*, *Performance efficiency* and *Usability* should be evaluated by citizens and enterprises. The third limitation was the two-point scale of evaluation, while generally the level of capabilities implementation should be measured. However, this study helps provide some insights that can lead to improvement of e-gov portals, not only in Poland but also in other countries.

The research will be continued. On the basis of the pilot study findings the factors will be modified and extended in each dimension, and Likert scale will be applied before launching a larger study. The theoretical and methodological paradigm of this research will be in-depth explored and the verification will be conducted on the example of Polish and other countries' e-gov portals.

Acknowledgements

This research has been supported by a grant entitled "Designing a system approach to sustainable development of the information society – on the example of Poland" from the National Science Centre in Poland, 2011/01/B/HS4/00974, 2011-2014.

References

[1] O. Almalki, Y. Duan and I. Frommholz, "Developing a conceptual framework to evaluate e-government portals' success," in *Proceedings of the 13 European Conference on e-Government*, University of Insubria, E. Ferrari, W. Castelnovo, Eds. Varese Italy, 13-14 June, 2013, 1, pp. 19-26.

[2] COM (1999) 687 final, *eEurope - An Information Society For All*, Communication of 8 December 1999 on a Commission initiative for the special European Council of Lisbon, 23 and 24 March 2000, retrieve from: http://europa.eu/legislation_summaries/information_society/strategies/index_en.htm, 2012.

[3] COM (2001) 140 final, *eEurope 2002: Impact and Priorities*, Commission Communication of 13 March 2001 on a Communication to the Spring European Council in Stockholm, 23 and 24 March, retrieve from: http://europa.eu/legislation_summaries/information_society/strategies/index_en.htm, 2012.

[4] COM (2003) 567, *The Role of eGovernment for Europe's Future*, Communication of 26 September 2003 from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions, retrieve from: http://europa.eu/legislation_summaries/information_society/strategies/index_en.htm, 2012.

[5] COM (2006) 173 final, *i2010 eGovernment Action Plan - Accelerating eGovernment in Europe for the Benefit of All*, retrieve from: http://europa.eu/legislation_summaries/information_society/strategies/index_en.htm, 2012.

[6] COM (2010) 245, *A Digital Agenda for Europe*, Communication from the Commission of 19 May 2010 to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, retrieve from: http://europa.eu/legislation_summaries/information_society/strategies/index_en.htm, 2012.

[7] COM (2010) 2020 final, *EUROPE 2020 A Strategy for Smart, Sustainable and Inclusive Growth*, European Commission, retrieve from:http://europa.eu/legislation_summaries/information_society/strategies/index_en.htm, 2012.

[8] MSWiA (2008), *Strategia Rozwoju Społeczeństwa Informacyjnego w Polsce do roku 2013*, Warszawa: Ministerstwo Spraw Wewnętrznych i Administracji, 2008, retrieve from: http://www.mswia.gov.pl/strategia/, 2010.

[9] Sejmik (2009), *Strategia Rozwoju Społeczeństwa Informacyjnego Województwa Śląskiego do roku 2015*, Katowice: Uchwała nr III/37/2/2009 Sejmiku Województwa Śląskiego, 2009.

[10] A. V. Anttiroiko, "A brief introduction to the field of e-government," in *Electronic Government: Concepts, Methodologies, Tools, and Applications*, A. V. Anttiroiko, Ed. New York: Hershey, 2008, pp. xli-lxxv.

[11] H. Michel, "E-administration, e-government, e-governance and the learning city: A typology of citizenship management using ICTs," *The Electronic Journal of e-Government*, vol. 3, no. 4, pp. 213-218, 2005.

[12] J. R. Gil-García and N. Helbig, "Exploring e-government benefits and success factors," in *Encyclopedia of Digital Government,* A. V. Anttiroiko and M. Mälkiä, Eds., Hershey: Idea Group Reference, vol. 2, pp. 803-811, 2007.

[13] E. Ziemba and T. Papaj, "E-government Application at the Regional Level in Poland – the Case of SEKAP," *in Proceedings of the Federated Conference on Computer Science and Information Systems,* Wrocław, 9-12 September 2012, pp. 1075-1082.

[14] E. Ziemba, T. Papaj, and R. Żelazny, "A model of success factors for e-government adoption – the case of Poland," *Issues in Information Systems*, vol. 14, no. 2, pp. 87-100, 2013.

[15] M. Kachwamba and A. Hussein, "Determinants of e-government maturity: Do organizational specific factors matter?," *Journal of US-China Public Administration,* vol 6, no.7 (Serial No.50), pp.1-8, 2009.

[16] T. Almarabeh and A. AbuAli, "A general framework for e-government: definition maturity challenges, opportunities, and success," *European Journal of Scientific Research*, vol.39, no.1, pp. 29-42, 2010.

[17] P. Ifinedo and M. Singh, "Determinants of eGovernment maturity in the transition economies of Central and Eastern Europe," *Electronic Journal of e-Government*, vol. 9, issue 2, pp. 166-182, 2011.

[18] *The User Challenge Benchmarking the Supply of Public Services - 7th Measurement*, September 2007, European Commission, Directorate General for Information Society and Media, Diegem, 2007.
[19] *Digitizing Public Services in Europe: Putting Ambition into Action, 9th Benchmark Measurement*, European Commission, CapGemini, Rand Europe, 2010.
[20] E. Ziemba and T. Papaj, "A Pragmatic approach to the e-government maturity in Poland – implementation and usage of SEKAP," in *Proceedings of European Conference on eGovernment,* University of Insubria, Como, 2013, pp. 560-570.
[21] S. al Shafi and V. Weerakkody, "Understanding citizens' behavioural intension in the adoption of e-government services in the state of Qatar," in *Proceedings 17th European Conference on Information Systems*, 2009, p.1-13.
[22] S. Angelopoulos, F. Kitsios and T. Papadopoulos, "New service development in e-government: identifying critical success factors," *Transforming Government: People, Process and Policy*, vol. 4, issue 1, pp. 95 – 118, 2010.
[23] S. C. Srivastava, "Is e-government providing the promised returns?: A value framework for assessing e-government impact," *Transforming Government: People, Process and Policy*, vol. 5, issue 2, pp.107-113, 2011.
[24] S. C. Srivastava, "Is e-government providing the promised returns? A value framework for assessing e-government impact," *Transforming Government: People, Process and Policy*, vol. 5, issue 2, pp. 107 – 113, 2011.
[25] S. Marche and J. D. McNiven, "E-government and e-governance: the future isn't what it used to be," *Canadian Journal of Administrative Science*, vol. 20, no. 1, pp. 74-86, 2003.
[26] R. M. Davison, C. Wagner and L. C. K. Ma, "From government to e-government: a transition model," *Information Technology & People*, vol. 18, no. 3, pp. 280-299, 2005.
[27] J. Choudrie, V. Weerakkody and S. Jones, "Realising e-government in the UK: rural and urban challenges," *Journal of Enterprise Information Management*, vol. 18, no. 5, pp. 568-585, 2005.
[28] Z. Kovacic, "The impact of national culture on worldwide egovernment readiness," *Informing Science*, 8, pp. 143-158, 2005.
[29] M. Asgarkhani, "Digital government and its effectiveness in public management reform: A local government perspective," *Public Management Review*, vol. 7, no. 3, pp. 465-487, 2005.
[30] A. Sultan, K. A. AlArfaj, and G. A. AlKutbi, "Analytic hierarchy process for the success of e-government," *Business Strategy Series*, vol. 13, no. 6, pp. 295-306, 2007.
[31] N. Letch and J. Carroll, "Excluded again: implications of integrated e-government systems for those at the margins," *Information Technology & People*, vol. 21, no. 3, pp. 283-299, 2008.
[32] E. N. Nfuka and L. Rusu, "The effect of critical success factors on IT governance performance," *Industrial Management & Data Systems*, vol. 111, no. 9, pp. 1418-1448, 2011.
[33] F. Zhao, "Impact of national culture on e-government development: a global study," *Internet Research*, vol. 21, no. 3, pp. 362-380, 2011.
[34] E. Daniel and J. Ward, "Integrated service delivery: exploratory case studies of enterprise portal adoption in UK local government," *Business Process Management Journal,* vol. 12, no. 1, pp. 113-123, 2006.
[35] D. J. Calista and J. Melitski, "Digitized government best practices in country web sites from 2003 to 2008: the results are bifurcated," *Business Process Management Journal*, vol. 18, no. 1, pp. 138-162, 2012.
[36] A. Y. K. Chua, D. H. Goh and R. P. Ang, "Web 2.0 applications in government web sites: prevalence, use and correlations with perceived web site quality," *Online Information Review*, vol. 36, no. 2, pp. 175-195, 2012.
[37] R. Luburić, "Total quality management as a paradigm of business success," *Journal of Central Banking Theory and Practice*, vol. 3, no.1, pp. 59-80, 2014.
[38] D. A. Garvin, *Managing Quality - the strategic and competitive edge*. New York, New York: Free Press, 1988.
[39] B. Kitchenham and SL. Pfleeger, "Software quality: the elusive target," *IEEE Software*, vol. 13, no. 1, pp. 12-21, 1996.
[40] ISO (2005), *Quality management systems – Fundamentals and vocabulary* – ISO 9000:2005.
[41] R. W. Griffin, *Management*, Mason, South-Western Cengage Learning, 2013.
[42] L. Wankhade and B. Dabade, Quality uncertainty and perception: information asymmetry and management of quality uncertainty and quality perception, Berlin: Springer-Verlag, 2010.
[43] O. Almalki, Y. Duan, and I. Frommholz, "Developing a conceptual framework to evaluate e-government portals' success," in *Proceedings of the 13 European Conference on e-Government*, E. Ferrari, W. Castelnovo, Eds. University of Insubria, 13-14 June 2013, 1, 2013, pp. 19-26.
[44] D. Bhattacharya, U. Gulla and M. P. Gupta, "E-service quality model for Indian government portals: citizens' perspective," [online], http://www.emeraldinsight.com/journals.htm?issn=1741-0398, vol. 25, no. 3, pp. 246-271, 2012.
[45] Ch. Halaris, B. Magoutas, X. Papadomichelaki and G. Mentzas, "Classification and synthesis of quality approaches in e-government services," *Internet Research,* vol. 17, no. 4, pp. 378-401, 2007.
[46] Y. S. Wang and Y. W. Liao, "Assessing e-government systems success: a validation of the DeLone and McLean model of information systems success," *Government Information Quarterly*, vol. 25, no. 4, pp. 717-733, 2008.
[47] S. K. Sharma, H. Al-Shihi and S. M. Govindaluri, "Exploriong quality of e-governmnet services in Oman," *Education, Business and Society: Contemporary Middle Eastern Issues,* vol. 6, no. 2, pp. 87-100, 2013.
[48] W. H. DeLone and E. R. McLean, "Information systems success: the quest for the dependent variable," *Information Systems Research*, vol. 3, no. 1, pp. 60-95, 1992.
[49] W. H. DeLone and E. R. McLean, "Information systems success revisited," in *Proceedings of the 35th Hawaii International Conference on System Sciences IEEE Computer Society,* Hawaii, US, 2002, pp. 1-11.
[50] W. H. DeLone and E. R. McLean, "The DeLone and McLean model of information systems success: a ten-year update," *Journal of Management Information Systems*, vol. 19, no. 4, pp. 9-30, 2003.
[51] S. Petter, W. H. DeLone and E. McLean, "Measuring Information Systems Success: Models, Dimensions, Measures, and Interrelationships," *European Journal of Information Systems,* 17, pp. 236-263, 2008.
[52] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319-339, 1989.
[53] Y. S. Wang, "Assessing e-commerce systems success: a respecification and validation of the DeLone and McLean model of IS success," *Information Systems Journal*, vol. 18, no. 5, pp. 529-557, 2008.
[54] P. Saha, A. K. Nath and E. Salehi-Sangari, "Evaluation of government e-tax websites: an information quality and system quality approach," *Transforming Government: People, Process and Policy*, vol. 6, no. 3, pp. 300-321, 2012.
[55] ISO/IEC (2011), *Systems and software engineering. Systems and Software quality*. *Requirements and Evaluation* (SQuaRE). System and software quality models. ISO/IEC 25010:2011(E), International Organisation for Standardisation, Geneva, 2011.
[56] Analysis o*f Good Practice in the Area of E-government* (Analiza Dobrych Praktyk w Obszarze E-administracji), IT Project Center (Centrum Projektów Informatycznych), Warsaw, 2013.
[57] E. Ziemba, "The holistic and systems approach to the sustainable information society," *Journal of Computer Information Systems*, vol. 54 no. 1, pp. 106-116, 2013.
[58] E. Ziemba, T. Papaj and D. Descours, "Factors affecting success of e-government portals – a perspective of software quality model," in *Proceedings of European Conference on eGovernment,* Brasov, Spiru Haret University, 2014, pp. 252-262.
[59] PN-ISO/IEC 27001:2007, System zarządzania bezpieczeństwem informacji, PKN, Warszawa, 2007.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1269–1275
DOI: 10.15439/2014F178
ACSIS, Vol. 2

Investigation of the Cobit Framework's Input\Output Relationships by Using Graph Metrics

Mesut Ateşer
Department of Information Security and Management, ÖSYM, Ankara, Turkey
Email: mesut.ateser@osym.gov.tr

Özgür Tanrıöver
Department of Computer Engineering, Ankara University, Turkey
Email:ozgurtanriover@yahoo.com

***Abstract*—The information technology (IT) governance initiatives are complex, time consuming and resource intensive. COBIT, (Control Objectives for Information Related Technology), provides an IT governance framework and supporting toolset to help an organization ensure alignment between use of information technology and its business goals. This paper presents an investigation of COBIT processes' and inputs/outputs relationships with graph analysis. Examining the relationships provides a deep understanding of COBIT structure and may guide for IT governance implementation and audit plans and initiatives. Graph metrics are used to identify the most influential/sensitive processes and relative importance for a given context. Hence, the analysis presented provide guidance to decision makers while developing improvement programs, audits and possibly maturity assessments based on COBIT framework.**

I. Introduction

INFORMATION technology has become a vital and integral part of many business activities and also in the support, sustainability, and growth of enterprises. Business and IT departments, must understand each other and make the strategic / tactical plans together for achieving goals of the organization. IT should provide the necessary services to business, plan, manage existing services, be ready for agile developments, store and protect the data, consider operational jobs and so on. Managing that kind of complex organizations is very hard and to achieve well established management, a set of policies and processes are needed on corporate level [1]. IT governance is the structure of relationship and processes that ensure the effective and efficient use of IT to achieve organizational goals.

IT governance includes decision making structures, alignment processes and communication tools [2]. Demands of business departments, by force of competitive market, must be aligned with the plans of IT [3] [4]. IT needs to monitor all services, their life-cycles, and resources by considering the business expectations. To achieve this, enterprises seek for practical knowledge and well defined guidelines. The best known and generally accepted IT governance framework is COBIT. COBIT, now in its fifth edition, describes a set of good practices for the board and senior operational and IT management [5]. According to the ISACA COBIT 4.1 has been downloaded more than 100.000 times over 160 countries. Although COBIT version 5 is published, COBIT 4.1 is still in use in most organizations and widespread so that we use COBIT 4.1 as source in this research.

COBIT provides a governance framework, supporting toolset and maturity model to help an organization ensure alignment between use of information technology and its business goals in the areas of risk management, resource management, performance measurement, value delivery and regulatory compliance. It is based on best practice in IT management and control. COBIT framework defines 34 processes under four domains and also 318 detailed control objectives and associated audit guidelines. The framework identifies seven information criteria such as effectiveness, efficiency, confidentiality, integrity, availability, compliance and reliability as well IT resources as people, applications, information and infrastructure [6] [7] [8].

COBIT version 4.1 management guidelines provides a section, describe inputs and outputs for each process. These input and output tables represent a brief description for the processes' relationships. Examining the relationships provides a deep understanding of COBIT structure and may guide for IT governance implementation and audit plans and initiatives. In this paper, an investigation of COBIT processes' and inputs/outputs relationships with graph analysis is presented. We aim to analyze the relative importance of processes based on graph metrics hence provide information to decision makers for developing improvement programs, audits and maturity assessments based on COBIT framework.

This paper is organized as follows. Section 2 provides literature on analysis on COBIT processes. Section 3 presents the method used to obtain the COBIT graph. Section 4 and its sub-sections provide the results of graph metrics used. The last section summarizes the main findings.

II. Related Work

Although COBIT and its related sources have been investigated widely such as comparison with other frameworks, detailed investigation of a specific area like security or project management etc. there are not much published papers concerning the inputs and outputs of COBIT processes. In paper [9] Tuttle and Vandervelde

978-83-60810-58-3/$25.00

examine the conceptual model of COBIT in an audit setting. They used data from COBIT assessments made by a panel and confirmed the internal consistency of COBIT. But their perspective was IT audits and used the data from the experts not from the COBIT itself. In [10] Bernroider and Ivanov investigated COBIT for specifically project management control (PO 10). They also used an empirical survey as data source similarly as Tuttle and Vandervelde.

Morimoto argues that COBIT is too general-purpose and requires expert knowledge to implement [11]. Morimoto was interested in only security part of framework and tried to create a new framework from existing frameworks. For that aim he used combination of ISO/IEC 12207 and ISO/IEC 27002 with COBIT. Morimoto is not the only one argued COBIT is very abstract and hard to implement and control objectives are fundamental examples. There are numerous papers mentioned that abstraction in such papers as [10], [12] and [13].

In [14] Abu-Musa has made an empirical survey using self-administered questionnaire among 500 Saudi organizations and 127 valid respond was collected. His findings claims that banks and financial organizations show more concerns to IT governance than other industries. Paper also provides us important processes over domain level according to the respondents. PO7 (Manage IT Human Resource) is subtracted being the most important process in the plan and organize domain. PO1, PO2, PO5, PO10 and PO11 are important processes, as well. AI1, AI2 and AI4 are the most important processes in acquire and implement domain. For delivery and support domain most important process is DS5. ME1 is important in monitor and evaluate domain. According to the results of paper it can be concluded that the importance of COBIT processes may change so far as industry and organizations' aims. That variation of importance may be observed in Kerr and Murthy investigation, also. On the other hand COBIT's provides e.g. management guidelines, control objectives, RACI charts and inputs/outputs section. There are some fundamentals information may be used during in case of an implementation.

In [12] inputs and outputs of processes in COBIT 4.1 investigated directly. They used the number of inputs and outputs as inputs and investigated the importance of processes according to these numbers. Their findings includes the most influential processes that are sending more artifacts to all other processes as PO4, PO6, PO7 and PO8. It is claimed that any improvement plan sequence should include PO4, PO6, PO7 and PO8 in its initial phase. They also produced some key analyses such as sensitivity that is sum of total inputs. In that sensitivity graph ME1 is the top process due to many inputs from other processes to monitor their performance. Besides calculating the summation of all inputs and outputs of a process interconnection is measured. This research`s aim is similar with our paper. But they just used the total numbers of inputs and outputs of processes. In our research, we convert the relationships between processes among by inputs and outputs into a graph differentiating the inputs and outputs. Neto, Fonseca and Webster's approach aligns with degree calculations in this research and outcomes are exactly similar based on degree. On the other hand we go further than that, using other graph metrics. These findings may help developing improvement programs, audits and maturity assessments based on COBIT framework to optimize resources and time.

III. GRAPH BASED ANALYSIS OF COBIT

A. COBIT Graph

COBIT contains Management Guidelines, including Maturity Models, Critical Success Factors, Key Goal Indicators and Key Performance Indicators for each of the 34 processes that are under four domains (see Appendix A) COBIT also provides inputs and outputs on the management guidelines section of each processes. For 34 processes, input output information of each process' control objective by two different tables is also presented For example, Table I. represents the input output relation for the PO1 (Define IT strategic Plan) process. However, it is difficult to obtain any holistic information from these tables, 34 processes and their input/output table turned into a relationship matrix. Then that matrix is converted to a graph to be able to investigate the overall framework.

Using Gephi [1] the matrix is converted to a graph as Figure 1. To create the graph 34 processes represented as vertices or nodes and their relationships as edges or arcs. Three IT requirements, going outside of COBIT, also represented as vertices as OTHER1, OTHER2, and OTHER3. One output to the outside of COBIT represented as a node as OUTSIDECOBIT. So in Figure 1 there are process-like items as OTHER1, OTHER2, OTHER3 and OUTSIDECOBIT. The definitions of external requirement or outputs as shown below.

- OTHER1 is Business strategy and priorities, an input for PO1 (Define a Strategic IT Plan).
- OTHER2 is Programme portfolio, an input for PO1 (Define a Strategic IT Plan).
- OTHER3 is Legal and regulatory compliance requirements, an input for ME3 (Ensure Compliance with External Requirements)
- OUTSIDECOBIT is Classification procedures and tools, an output from PO2 (Define the Information Architecture)

[1] Gephi is an open-source software for network visualization and analysis and provides built-in functions to explore, manipulate and analyze the data. The software is for Exploratory Data Analysis goals to make hypothesis, to discover patterns by using visuality.

TABLE I

INPUT&OUTPUT SECTION OF PO1 PROCESS.

From	Inputs	Outputs	To
PO5	Cost-benefits reports	Strategic IT plan	PO2, PO3, PO4, PO5, PO6, PO8, PO9, AI1, DS1
PO9	Risk assessment	Tactical IT plans	PO2, PO3, PO4, PO5, PO6, PO9, AI1, DS1
PO10	Updated IT project portfolio	IT project portfolio	PO5, PO6, PO10, AI6
DS1	New/updated service requirements; updated IT service portfolio	IT service portfolio	PO5, PO6, PO9, DS1
*	Business strategy and priorities	IT sourcing strategy	DS2
*	Programme portfolio	IT acquisition strategy	AI5
ME1	Performance input to IT planning		
ME4	Report on IT governance status; enterprise strategic direction for IT		

To observe all interactions over COBIT the external IT requirements are included. Finally, The Graph has 38 nodes which are processes basically, 311 edges which are relationships between processes. The graph is a directed and weighted graph which means there can be a path from PO1 to PO2 but not counter wise. Edges have weights that the numerical value of edge shows actual number of edges from one process to another. If PO1 has two different outputs to PO2 that means the edge between PO1 and PO2 has a weight of 2.

Table III demonstrates betweensess centrality metrics as size of nodes. Visualization is prepared by force-atlas algorithm provided by Gephi in layout section. Visualized graph seems to be understood easier. Processes that not include in framework are out of the heart-like shape. And strong relationships between all processes may be concluded directly. Using graph metrics will provide a deeper understanding of COBIT structure and may guide for IT governance implementation initiatives. For this purpose in the following section, the graph metrics and obtained results are presented.

IV. GRAPH METRICS BASED ANANLYSIS

A. Degree

In a graph *degree* is total number of the edges belongs to a specific vertex. Degree is essential and effective measure to decide the importance of a node. In a directed graph two types of degree comes out, in-degree and out-degree. In-degree for a vertex v is the number of edges that v is the terminal vertex. Similarly out-degree of a node v is the number of edges that v is initial vertex. Degree basically, shows the strength of relationships between processes. In COBIT I/O graph the average degree is 8.184 that means approximately any process can have relationship with 8 other processes. But the graph is directed and weighted so average weighted degree is 10.842. As shown some of vertices have high degree, some of them have low degree. According to average weighted degree lowly linked and highly linked nodes can be noticed.

For a full degree list table II can be analyzed. PO4, PO6, PO7 and PO8 are the top four processes based on degree. But our graph is directed and weighted so both degree and weighted degree values have to be considered, respectively in table II. PO4 has the highest rank in total degree but in weighted degree list PO6 is in the number 1 rank. Moreover comparing table II it can be concluded that plan and organize processes dominates degree list.

From high level perspective can be concluded at domain level. PO (Plan and Organize) has the largest number of output information. So PO domain may be considered as a good starting point for governance initiatives. Then DS (Delivery and Support) domain is in the second rank.

Moreover COBIT governance initiatives can start based on their strategies not only domain level but also process level. For example implementation can start with DS 5 (Ensure System Security) and then continue with PO7. While these 2 processes are in progress monitoring should be active. So that parallel works can be done such as PO1 (Define a Strategic Plan), DS1 (Define and Manage Service Levels) can be in parallel. Also PO4 and PO7 enactment programs can be simultaneous. Hence, from the viewpoint of initiating governance programs the question of groupings of process may arise.

B. Centrality Metrics

Centrality is an important structural attribute of a graph meaning a centrality score is about how a node fits within a graph overall. Vertices that have highest centrality are likely to be key conduits of information. Low centrality modes can be named as peripheral. Lower centrality can be associated with less work overload in an organization [15]. Centrality of a vertex is the relative importance within a graph. There are two common and widely used centrality metrics which are betweeness and eigenvector centrality.

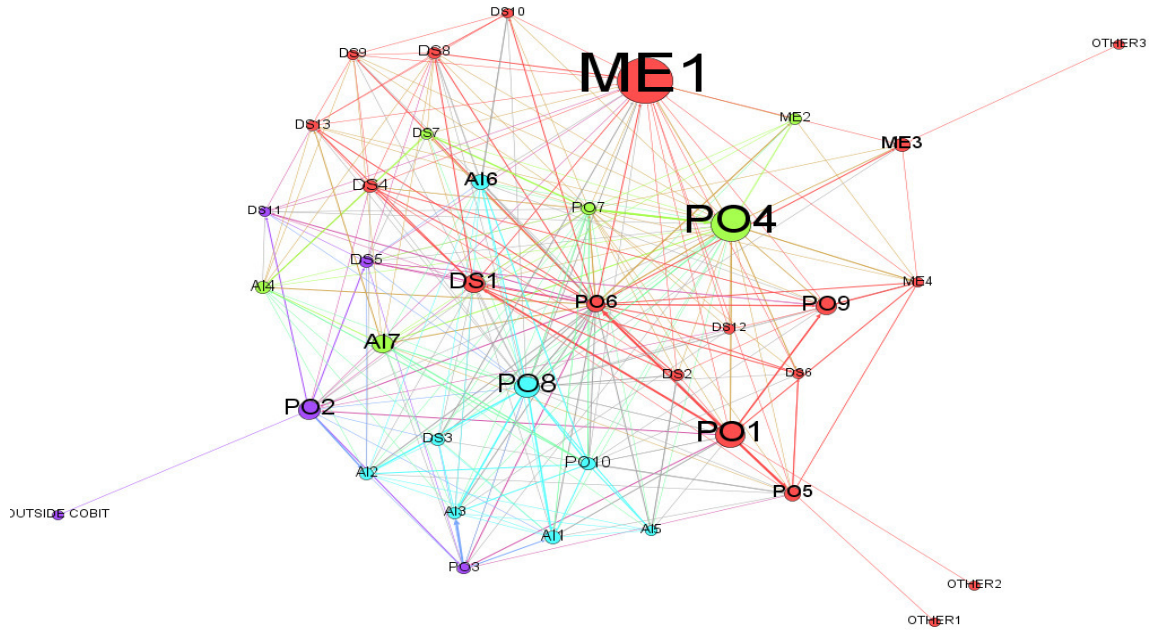


Figure 1 the Final Graph of I/O

C. Eigenvector Centrality

Eigenvector centrality or Gould index of accessibility [16] is calculated by assessing how well connected a node is to the parts of the network with the greatest connectivity. Eigenvector centrality is similar to the degree centrality but there is a difference that is eigenvector measures the importance of a node by the importance of its neighbors. A node receiving many edges does not mean receiver node has a high eigenvector centrality. Moreover, high eigenvector centrality node is not necessarily highly linked. That node may have a few edges to the other but they may be all important nodes.

Table III shows the eigenvector measurement. ME1 (Monitor and Evaluate IT Performance) has the highest score for eigenvector centrality. ME1 has a strong relationship with other processes as expected. But in the second rank is DS8 (Manage Service Desk and Incidents). DS8 gives outputs to AI6, DS10, ME1 and DS7. ME1 and AI7 have neighbors that have many connections and also takes inputs from 9 processes. As a conclusion DS8 is important because of its neighbor's importance. When top 5 processes are considered, they are similar in a manner. DS8 is for service desk, DS1 is for service levels, AI6 is to manage changes and finally PO9 is for risk management. These 4 processes are responsible to improve IT information criteria effectiveness and efficiency.

D. Betweenness Centrality

Betweenness centrality is a measure that is derived from shortest paths between nodes in a graph. The number of times a node acts as a cutpoint in the shortest point between two other nodes. In COBIT I/O graph that metric will show critical nodes to collaborate through all processes for spread of information. Algorithm in [17] is used to measure the centrality metric.

High betweenness nodes often don't have the shortest path to other nodes, but they have the greatest number of shortest path that have to go through them. Vertices that have high betweennes centrality metric, are critical to collaborate between other nodes. They are traders of information through the graph. Table III shows all nodes' betweeness centrality..

ME1 process has the highest betweenness centrality, PO4 follows it. DS6 (Delivery and Support domain – Identify and Allocate Cost) has the lowest betweenness centrality so that it can be inferred DS6 is an isolated process. DS6 is not a good point to maintain the spread of new information

TABLE III

FULL LIST OF DEGREE RESULTS

Process	Weighted Degree	Weighted In-Degree	Weighted Out-Degree	Degree	In-Degree	Out-Degree
PO6	75	9	66	39	6	33
PO4	53	15	38	42	9	33
PO8	52	7	45	39	6	33
PO7	46	7	39	37	4	33
PO1	42	15	27	25	12	13
ME1	33	26	7	30	23	7
DS1	32	14	18	23	11	12
PO10	30	10	20	20	7	13
PO2	24	11	13	17	9	8
PO5	24	16	8	19	12	7
PO9	23	14	9	19	11	8
AI7	23	14	9	18	10	8
AI6	23	16	7	19	12	7
AI1	23	16	7	17	10	7
AI3	23	16	7	16	9	7
AI2	22	16	6	16	10	6
PO3	21	11	10	14	8	6
AI4	20	12	8	17	10	7
DS8	20	15	5	17	13	4
DS4	19	11	8	17	9	8
DS5	17	11	6	15	9	6
ME4	17	11	6	13	8	5
DS3	16	8	8	14	7	7
DS13	16	13	3	13	10	3
AI5	16	13	3	12	10	2
DS2	15	12	3	11	8	3
DS7	14	12	2	10	8	2
DS9	13	8	5	12	7	5
DS10	13	9	4	11	8	3
DS11	13	11	2	11	9	2
DS6	13	11	2	9	7	2
ME2	11	7	4	10	6	4
ME3	9	6	3	8	5	3
DS12	9	8	1	8	7	1
OTHER1	1	0	1	1	0	1
OTHER2	1	0	1	1	0	1
OTHER3	1	0	1	1	0	1
OUTSIDE	1	1	0	1	1	0

TABLE II

FULL LIST OF CENTRALITY RESULTS

Process	Eccentricity	Betweenness Centrality	Eigenvector Centrality
OUTSIDE COBIT	-	0.00	0.05
PO4	2	181.18	0.42
PO8	2	87.23	0.32
PO6	2	44.07	0.27
PO7	2	19.66	0.17
PO1	2	115.34	0.52
ME1	2	280.42	1.00
AI1	2	27.50	0.44
PO10	3	24.90	0.31
DS1	3	65.24	0.58
PO9	3	59.79	0.54
AI7	3	61.91	0.43
ME4	3	5.17	0.32
ME2	3	13.26	0.31
ME3	3	34.00	0.14
PO5	3	34.29	0.49
DS3	3	15.64	0.30
AI6	3	42.55	0.53
PO3	3	11.15	0.34
DS4	3	22.79	0.43
DS5	3	16.71	0.42
DS9	3	5.90	0.29
DS8	3	16.13	0.60
DS2	3	10.73	0.37
DS10	3	1.53	0.36
DS7	3	5.14	0.38
DS13	3	6.77	0.44
OTHER1	3	0.00	0.00
OTHER2	3	0.00	0.00
DS6	3	0.40	0.30
DS11	3	3.04	0.41
DS12	3	1.86	0.30
PO2	4	67.04	0.46
AI3	4	15.64	0.37
AI2	4	18.08	0.44
AI4	4	23.51	0.42
AI5	4	6.39	0.43
OTHER3	4	0.00	0.00

E. Eccentricity

The eccentricity [18] is the distance of a starting node to the farthest node in a graph. COBIT graph here is weighted so that distance is a fundamental indice. In a graph minimum eccentricity value is its radius and maximum value is its diameter. In COBIT graph radius is 2 and diameter is 4. In Table III eccentricity values can be seen. PO4, PO6, PO7 and PO8 are in the list with value 2. But AI1 (Identify Automated Solutions), PO1 and ME1 are also in the list. Using eccentricity values it can be concluded that PO1, PO4, PO6, PO7, PO8, ME1 and AI1 are central vertices of the graph.

The overall results based on graph metrics are presented in Table IV.

V. CONCLUSION

Business dependency for IT services has been more crucial than ever. When the organization grows that dependency is increasing, too. To achieve organization's plans and business requirements IT governing the IT becomes tougher. Implementing an effective IT governance framework model becomes a necessity. Although providing a framework and toolset, it is hard to implement COBIT improvement initiatives. This study provides useful information about COBIT 4.1 framework from a holistic perspective. The findings presented in this paper can be used to develop a roadmap to plan the IT governance or audit initiatives.

Especially, plan and organize domain is found to be the most influential domain and processes PO4, PO6, PO7 and PO8 are particularly important. These processes produce many information items for the others, therefore these processes should be considered in first phase of the COBIT improvement initiatives. However it is important for an enterprise to decide based on specific objectives. The importance of the process implementation may change for the specific organization's approach. If the organization's aim may be to improve the current situation, then starting with PO2 process can be a good point, or the organization's aim may be to improve security then DS5 followed by AI6 and PO9 can be a starting point.

Also it is important to consider already enacted processes. Higher maturity level processes may produce mature outputs and in reverse processes that have lower maturity levels may be needed first as they may be central in the information flow. Using eccentricity values it can be concluded that PO1, PO4, PO6, PO7, PO8, ME1 and AI1 are central vertices of the graph.

As mentioned in the introduction, COBIT version 5 has been published. As a further research, we are planning to investigate COBIT5 in near future. In addition, usefulness of findings should be verified in practice.

VI. REFERENCES

[1] M. N. Kooper, R. Maes and E. Roos Lindgreen, "On the governance of information: Introducing a new concept of governance to support the management of information," *International Journal of Information Management,* vol. 31, pp. 195-200, 2011.

TABLE IV

OVERALL RESULTS

Metric	Description	Outcome
Eccentricity	The distance of a starting node to the farthest node	AI1,ME1,PO1,PO4,PO6,PO7,PO8
Closeness Centrality	Close center node can communicate with the other without need of many inder-mediaries	PO4,PO6,PO8,PO7
Betweenness Centrality	The number of times a node is in the shortest path between other two nodes	ME1
Eigenvector Centrality	The summation of the centrality values of a nodes that is connected to	ME1
Degree	Total number of edges of a node	PO4
Weighted Degree	Total number of edges of a node with weights in each edge	PO6
In-Degree	The number of edges that node is terminal itself	ME1
Out-Degree	The number of edges that node is initial itself	PO4,PO6,PO8,PO7
Weighted In-Degree	Total number of edges that node is terminal itself with weights in each edge	ME1
Weighted Out-Degree	Total number of edges that node is initial itself with weights in each edge	PO6

[2] P. Weill and J. W. Ross, IT govarnance How top performers manage IT decision rights for superior results, Boston: Harvard Business School Press, 2004.

[3] J. C. Henderson and N. Venkatraman, "Strategic alignment: Leveraging information technology for transforming organizations," *IBM Systems Journal,* vol. 32, no. 1, pp. 4-16, 1993.

[4] R. Hirschheim and R. Sabherwal, "Detours in the path toward strategic information systems alignment," *California Management Review,* vol. 44, no. 1, pp. 87-108, 2001.

[5] ISACA, COBIT 5: A Business Framework for the Governance and Management of Enterprise IT, Rolling Meadows: ISACA, 2012.

[6] J. Lainhart, "COBIT: a methodology for managing and controlling information and information technology risks and vulnerabilities," *Journal of Information*

[7] G. Bodnar, "What's new in COBIT 4.0.," *Internal Auditing,* vol. 21, no. 4, pp. 37-44, 2006.

[8] G. Hardy, "Using IT governance and COBIT to deliver value with IT and respond to legal, regulatory and compliance challenges," *Information Security Technical Report,* vol. 11, no. 1, pp. 55-61, 2006.

[9] B. Tuttle and S. D. Vandervelde, "An Empirical Examination of COBIT as an Internal Control Framework for Information Technology," *International Journal of Accounting Information System,* vol. 8, pp. 240-263, 2007.

[10] E. W. Bernroider and M. Ivanov, "IT Project Management Control and the Contol Objectives for IT and Related Technology(COBIT)," *International Journal of Project Management,* pp. 325-336, 2011.

[11] S. Morimoto, "Application of COBIT to Security Management in Information System Development," in *International Conference on Frontier of Computer Science and Technology*, Shangai, 2009.

[12] J. S. Neto, H. V. Fonseca and I. L. & Webster, "Importance of Inputs/Outputs Within COBIT Processes," *COBIT Focus,* vol. 3, pp. 6-9, 2009.

[13] S. J. Hussain and S. S. Muhammed, "Quantified Model of COBIT for Corparate IT Governance," in *First International Conference on Information and Communication Technologies*, Karachi, 2005.

[14] A. Abu-Musa, "Exploring the Importance and Implementation of COBIT processes in Saudi Organizations An Emprical Study," *Information Management & Computer Security,* vol. 17, no. 2, pp. 73-95, 2009.

[15] L. C. Freeman, "Centrality in Social Networks," *Social Networks,* vol. 1, no. 3, pp. 215-239, 1978-1979.

[16] P. Gould, "On the Geographical Interpretation of Eigenvalues," *Transactions of the Institute of British Geographers,* vol. 47, pp. 53-86, 1967.

[17] U. Brandes, "A Faster Algorithm for Betweenness Centrality," *Journal of Mathematical Sociology,* vol. 25, no. 2, pp. 163-177, 2001.

[18] P. Hage and F. Harray, "Eccentricity and centrality in networks," *Social NEtworks,* vol. 17, pp. 57-63, 1995.

APPENDIX A: COBIT PROCESSES

Plan and Organize(PO)	PO1 Define a Strategic IT Plan	DS1 Define and Manage Service Levels	Delivery and Support(DS)
	PO2 Define the Information Architecture	DS2 Manage Third-party Services	
	PO3 Determine Technological Direction	DS3 Manage Performance and Capacity	
	PO4 Define the IT Processes Organization and Relationships	DS4 Ensure Continuous Service	
	PO5 Manage the IT Investment	DS5 Ensure Systems Security	
	PO6 Communicate Management Aims and Direction	DS6 Identify and Allocate Costs	
	PO7 Manage IT Human Resources	DS7 Educate and Train Users	
	PO8 Manage Quality	DS8 Manage Service Desk and Incidents	
	PO9 Assess and Manage IT Risks	DS9 Manage the Configuration	
	PO10 Manage Projects	DS10 Manage Problems	
Acquire and Implement(AI)	AI1 Identify Automated Solutions	DS11 Manage Data	
	AI2 Acquire and Maintain Application Software	DS12 Manage the Physical Environment	
	AI3 Acquire and Maintain Technology Infrastructure	DS13 Manage Operations	
	AI4 Enable Operation and Use	ME1 Monitor and Evaluate IT Performance	Monitor and Evaluate(ME)
	AI5 Procure IT Resources	ME2 Monitor and Evaluate Internal Control	
	AI6 Manage Changes	ME3 Ensure Compliance With External Requirements	
	AI7 Install and Accredit Solutions and Changes	ME4 Provide IT Governance	

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1277–1284
DOI: 10.15439/2014F184
ACSIS, Vol. 2

Acquiring a Digital Audience for Theaters – Looking Through The Lenses of Customer Equity and Empirical Research

Paweł Kossecki
The Polish National Film, Television and Theater School
ul. Targowa 61/63
90-323 Łódź, Poland
Email: kossecki@poczta.onet.pl

Urszula Świerczyńska-Kaczor
Jan Kochanowski University, Żeromskiego 5, 25-369 Kielce Poland,
Email: swierczynska@ujk.edu.pl

Abstract— **The aims of this paper are to: 1) outline and discuss the framework for linking theater e-marketing with customer equity; 2) assess the impact of digital theater services on the metrics connected with Customer Lifetime Value. The results of empirical research suggest that art-oriented young Internet-users, who do not attend traditional theaters, can be attracted to digital theater services. Digital services can influence the potential patron's engagement in the theater's website, favorable word-of-mouth, and also their intention to visit traditional performances.**

INTRODUCTION

Marketing for performing arts organizations is mostly investigated through the lenses of the traditional marketing-mix, experience marketing or/and relationship marketing. As a new approach to marketing, based on customer equity, has been developing in the business sector (Blattberg & Deighton 1996; Rust et al. 2004; Hogan et al. 2002), it is tempting to start a discussion about the contribution of this approach to marketing of the non-profit cultural sector.

In this paper the problem of acquiring a digital audience is placed within the boarder context of customer equity. It aims to: 1) outline and discuss the framework for linking theater e-marketing with customer equity; 2) assess the impact of digital theater services on the metrics connected with Customer Lifetime Value (and therefore customer equity), such as a patron's intention to purchase theater services and to spread positive word-of-mouth referrals.

The paper is organized as follows. In the next section we discuss the research problem and outline the research framework. The third section refers to the results of the questionnaire study. The article concludes with the summarized results.

RESEARCH FRAMEWORK

A. The problem

Relationship marketing aims to build the loyalty of patrons, increase patron retention rate and to shift customers from the status of 'potential customer' to the level of 'partner' and is perceived as the base of theater strategy (Rentschler et. al 2002; Quero 2007). Relationship marketing differs from the customer equity approach. Customer equity emphasizes the role of profitability of the customer (not all relationships should be maintained from the company's perspective), and customer equity approach is also useful for the situations while building strong relationship between the company and the customer is difficult (e.g. FMCG market).

In the case of the theater audience, customer equity marketing[1] relies on:

- Relationships build with patrons. Classic theaters rely on building a relationship with their patrons, and the consumers often repeat their purchases for years. Therefore there is a possibility to evaluate Customer Lifetime Value (CLV) at an individual level or segment level (as opposed to the situation where single purchases cannot be traced e.g. buying washing powder, and only segment level data is available for evaluating customer equity)
- The valuation of the segment of regular theatergoers. The segment of frequent theatergoers is often the core audience (see e.g. - Instytut Teatralny "Raport – Badanie publiczności teatrów w stolicy" 2012; [http://www.instytut-teatralny.pl/projekty/raporty; 05.04.2014],

[1] Customer Equity Marketing is defined "[…] as a management approach for acquisition and retention, geared to individual lifetime values of current and future customers with the aim of continuously increasing Customer Equity" [Bayón et al. 2002, p. 214].

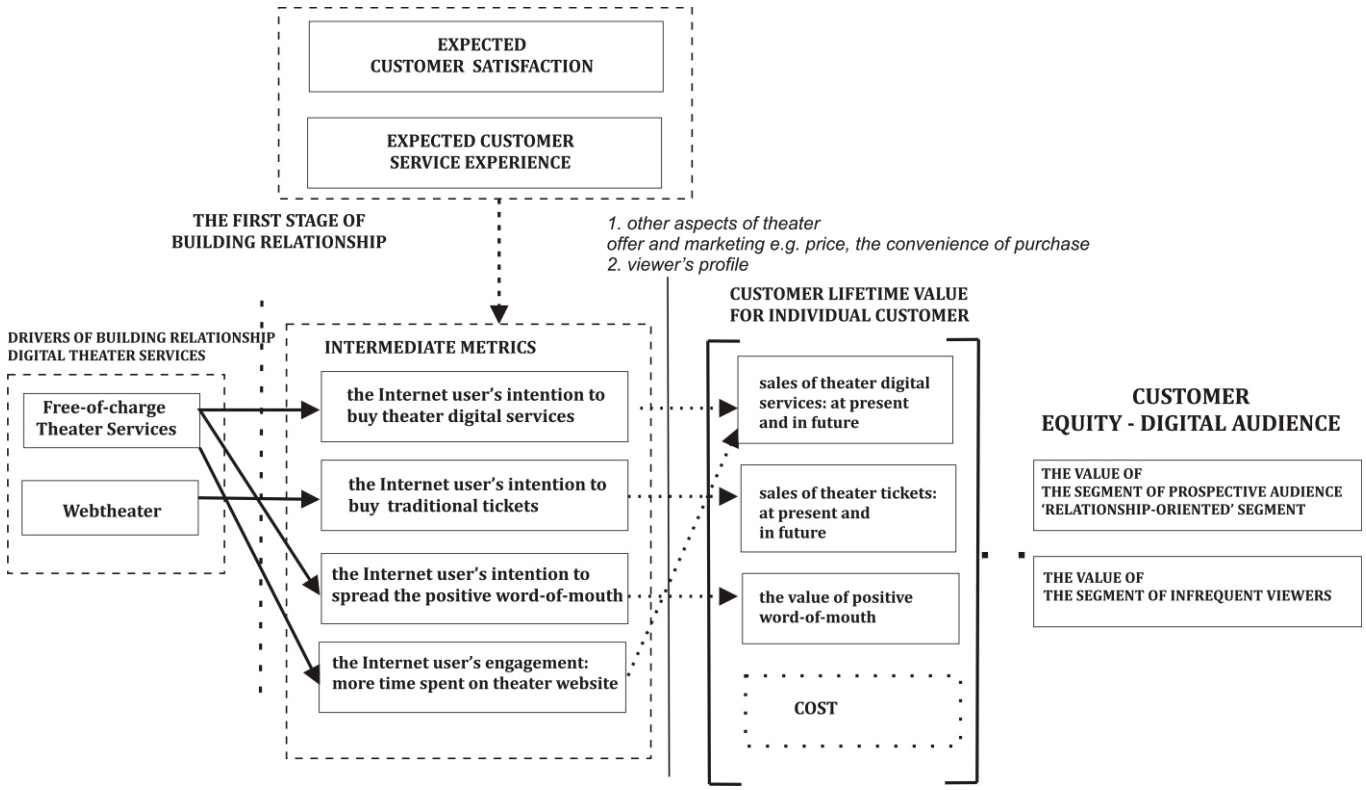


Fig. 1 The framework connecting theater e-marketing to customer equity

- Access to the data. The theater can often evaluate CLV at individual level (due to the data of subscribing patrons). The data referring to the market, such as number of viewers, number of performances, and cost of productions, is also often publicly available.

For overall theater strategy, customer equity management should be framed into boarder objectives such as objectives relating to artistic or curatorial quality, educational objectives, objectives relating to access for the public, objectives relating to knowledge, and social goals (Bakhshi & Throsby 2012 pointed out five dimensions for non-profit cultural institution). The idea of looking through the lens of maximizing customer equity may shape the theater's strategic decisions especially in other-than-product areas e.g. the selection of promotional channels, in the process of customer service before and after performance, or additional theater services. The advantage of applying customer equity lens lies in the shift from the product-oriented perspective to the customer-oriented perspective. The customer equity management (if implemented) would also 'force' theaters to investigate and manage the drivers of customer equity.

There are different approaches to the spectrum of drivers influencing customer equity. In the model developed by Rust, Lemon and Zeithml, brand equity is one of the drivers of customer equity along with value equity and relationship equity (Rust et. al 2004). Other researches point out that the concept of brand equity and customer equity are not the same constructs, but they can be seen as two sides of the same coin - marketing activities leading to the influence of brand equity also influence customer equity and vice versa (Bick 2009; Leone et al. 2006).

In this paper we raise the question whether the implementation of e-marketing tools can impact the variables influencing the first phase of building a relationship with prospective digital viewers, and therefore their Customer Lifetime Value and customer equity (as indicated in the research framework – Fig.1). We focus on the Internet as a tool of attraction strategy for a few reasons. Firstly, the Internet plays a significant role in theater e-strategy as:

1. The Internet facilitates innovation strategy for cultural organization in areas such as innovations in audience reach and in artform development, innovations in value creation and/or innovations in business management and governance (Bakhshi & Throsby 2012);
2. Websites or use of social media can ease the effect of constrained availability and uncertain outcomes of a 'difficult' theatrical brand. The Internet allows the establishment of 'persistent presence of the theater brand in patrons' mind' and the creation of 'small worlds - communities of performing art patrons' (see Preece & Johnson 2011);
3. The Internet can be effectively used in improving the responsiveness of theaters to different stakeholders, in enhancing the image, in attracting volunteers or sponsors and in selling tickets (Turrini et. al. 2012);
4. Social media can support marketing in the area of promotion and communication, word-of-mouth referrals, market research and innovation

management, and in reputation management (Hausmann & Poellmann 2013).

Moreover, as there is a steady decline in participation in art events and also with the problem of aging audiences (Turrini et al. 2012), the Internet allows theaters to reach young people, who are uninterested in traditional theater, but may become the theater audience within cyberspace. We assume that winning over a young audience in the competitive market of leisure activities needs 'the most untraditional approach'. Therefore in the empirical study we included the concept of theater services which are not only interactive, but break with the traditional perception of the role of the viewers. The tested concepts of new digital theater services are:

- free-of-charge digital products such as newsletters, ebooks about theater, webseminars about theater, webseminars about acting, gadgets connected with the theater, job posts, marketing research reports about theater;
- webtheater – the new artform which was described to each respondent as a possible option embedded on the website. So far, the live broadcasts of art performances (such as concerts, opera or even theater performance) lack the basic value of virtual communication, as they do not allow the spectator to co-create the product. The idea of the webtheater performance is based on co-creation of the play – the Internet-user can influence the flow of the performance by voting online on selected decisions e.g. the Internet-user can decide the fate of the main hero. Therefore the second part of the performance depends on the audience's decisions.

We assume that the Internet lowers the threshold of building the relationship between the prospective young viewers and the theater. For traditional performances "[t]he organization invests time, money and commitment to its mission to give the patron a rewarding experience. The audience also invests time, money and an emotional and social commitment in the arts organization." (Rentschler et. al 2002, p. 124). The Internet allows audiences to 'try' digital theater services, and makes theater attendance more convenient for young, digital audience. Is it 'the same theater' as traditional theaters from the aspect of the customer's experience? Probably not, but if traditional theatres do not attract young audiences, maybe 'digital theater' would meet their needs and enable the beginning of a young audience's relationship with traditional theater.

The customer equity approach raises questions about the drivers of customer equity for digital theater. We assume that the value of customers visiting the theater website (with developed interactive services) would be connected with at least:

1. financial value connected with direct purchases on the Internet e.g. tickets or other theater products (e.g. gadgets, ebooks);
2. indirect value connected with:
   a. enhancing the attractiveness of the website for advertisers (more engaged visitors stay longer on the website, probably paying more attention to its content including advertisements);
   b. Internet-user's referrals (performance, the theater website itself) to others;
   c. Internet-user's contribution, such as expressing and sharing opinions, which allows for improvement to the theater services (we did not capture this construct in the study).

B. The framework of empirical survey

The conducted survey was an exploratory study, and at this point of the research we focused on a selected problem - the link between online theater marketing and the rate of acquiring digital patrons. Therefore not all of the interrelations indicated in framework Fig. 1 were investigated.

We based the survey on a "what-if" scenario, while respondents were asked to what extent the new forms of free-of-charge services and interactive theater (webtheater) attracts their attention and influences their intention to purchase paid products.

In the study we measured the Internet-user's intention to purchase or consume the theater website services, assuming that strong intention leads to real purchase - if other aspects of the theater offer will be acceptable for the Internet user. The measured metrics, such as intention to buy theater services or tickets, the intention to recommend theater services or to be engaged on theater website, are linked with Internet-user's perception of service experience and expected satisfaction.

We want to emphasize that:

- we test the concept of digital theater services, not of developed and produced digital products;
- we do not include the costs, as the cost is 'unpredictable' at this point of research;
- the constructs of expected Internet-user' satisfaction, and perception of the expected experience are boarder than 'intentions' measured; although we presume that the user's intentions are linked with these constructs.

TABLE I.
THE QUESTION: NOWADAYS MOST THEATERS HAVE THEIR OWN WEBSITES, FACEBOOK, GOOGLE PLUS, TWITTER PAGE OR RUNNING BLOG. IF THE THEATER EMBEDDED THE FOLLOWING INTERNET SERVICES AND OFFERED THEM FREE-OF-CHARGE, HOW THIS WOULD AFFECT YOUR VISIT TO THE THEATER WEBSITE E.G. YOU SPEND MORE TIME ON THEATER WEBSITES ON A MONTHLY BASIS DUE TO VISITING MORE OFTEN OR TO CONSUMING MORE CONTENT OF THE WEBSITE? IF THESE SERVICES WHERE TO BE PAID FOR, HOW IT WOULD AFFECT YOUR ATTITUDE?

The theater offer: Group OT 'Out of target market n=37 Group AO 'art-oriented' n=45 Group FS 'filmschool students' n=44		**Free-of-charge**			**Paid content**		
		Negative 1-2	**Positive 4-5**	**Average (1-5)**	**Negative (1-2)**	**Positive (4-5)**	**Average (1-5)**
e-book about theater	OT	59.5%	29.7%	2.5	89.2%	2.7%	1.4
	AO	20.0%	62.2%	3.7	76.6%	15.6%	2.0
	FS	36.4%	40.9%	3.2	77.3%	11.4%	1.8
webseminar about theater	OT	64.9%	13.5%	2.1	83.8%	5.4%%	1.5
	AO	40.0%	31.1%	2.9	73.3%	13.3%	2.0
	FS	47.7%	29.5%	2.8	72.7%	13.6%	1.8
e-report about theater market	OT	54.1%	35.1%	2.6	89.2%	10.8%	1.6
	AO	40.0%	37.8%	3.0	77.8%	22.2%	2.0
	FS	50.0%	29.5%	2.7	88.6%	6.8%	1.6
job posts connected with theater	OT	32.4%	45.9%	3.1	73.0%	16.2%	1.9
	AO	24.4%	51.1%	3.4	75.6%	20.0%	2.1
	FS	9.1%	65.9%	4.0	88.6%	2.3%	1.5
gadgets	OT	32.4%	51.4%	3.3	70.3%	18.9%	2.1
	AO	17.8%	66.7%	3.9	51.1%	33.3%	2.6
	FS	29.5%	40.9%	3.2	72.7%	6.8%	1.8
webseminar about acting	OT	45.9%	35.1%	2.6	59.5%	21.6%	2.3
	AO	22.2%	64.4%	3.8	48.9%	44.4%	3.0
	FS	43.2%	31.8%	2.9	79.5%	11.4%	1.8
e-newsletter about theater	OT	100%	-	1.4	89.2%	2.7%	1.4
	AO	-	100%	4.5	62.2%	26.7%	2.4
	FS	20.5%	54.5%	3.5	81.8%	6.8%	1.7

C. The data and the sample

Although the construct of customer equity as a value of the customer base is agreed, there is no consensus on how to calculate it and what drivers influence customer equity and what intermediate constructs and metrics should be used to assess customer equity. Kumar & George (2007) point out that the approaches to calculate customer equity which appear in literature can be divided into two main streams:

- aggregate-level approaches evaluate customer equity using customer life time value (CLV) for a segment or "average CLV" in the company. Therefore these methods focus on segment- or firm- level, and on such metrics as the acquisition rate, the average average contribution margin or the average CLV from a sample;
- disaggregate-level approach is based on the sum of the calculated individual CLVs, therefore this approach needs to be based on detailed and individual level data.

In our study we collected the data from the sample of individual respondents who filled in the questionnaire. As we do not have the data connected with costs, we concentrated on assessing the customers' willingness to purchase the product and in spreading positive word-of-mouth recommendations.

Our approach to Customer Lifetime Value refers to the model proposed by Kossecki (2007, 2011) in which CLV for individual customer c is evaluated as:

$$CLV_C = \sum_{t=0}^{CL} \frac{P_t}{(1+r)^t}$$

P_t –profit per individual user in t-period. This profit should be understood as financial profit coming

from two sources: direct purchases and positive word-of-mouth
CL – customer lifetime
r – discount rate;

From an aggregated level perception the digital audience can be divided into segments according to their willingness to build a relationship with the theater. The customers who focus on 'one-time' contact with theater (e.g. interested in receiving free-of-charge theater digital gadgets) are less valuable than the relationship-oriented customers. Valuation of the customers of the 'one-deal' segment may be based on such metric as the profit per transaction, but the CLV approach is more suitable for the segment of 'relationship-oriented' customers (Kossecki 2007; Kossecki 2011).

In the survey we implemented a questionnaire which was distributed in January 2014 among 146 respondents, within two different groups of young adults (under the age of 30 years-old): the students of a filmschool in Poland (whom we regarded as experts in the field of theater production, n=44) and students of business studies (whom we regarded as non-experts, n=102). There is a significant difference between the filmschool students and the business students in attendance to classic theaters (test U Mann-Whitney, $p<0.05$): both in the variable 'when last theater visit occurred' and the variable 'frequency of visiting theater during last year'. Most of the filmschool students attended classic theaters more than once during the last year (20% of filmschool students attended more than 3 times). In contrast, 76% of business students did not take part in theater performances for the last year, and 17% attended only once during the last year. In the study this group of business students is considered as the group of 'potential audience', who rarely or not at all attend classic theater. In the study we did not capture the socio- and demographical profile of respondents, but the sample is homogenous with the age of respondents.

D. Hypothesis

The framework for research (Fig. 1) leads to following hypotheses:

• Hypothesis 1: • The free-of-charge Internet services embedded on a theater website enhance customer's engagement
• Hypothesis 2: Free-of-charge Internet services embedded on a theater website enhance the sale of paid products.
• Hypothesis 3: The free-of-charge Internet services embedded on a theater website enhance customer referrals
• Hypothesis 4: The new digital interactive artform 'webtheater' enhances the customer's willingness to buy traditional tickets

We arbitrarily agreed that the percentage of the 'prospective audience' which we would consider as 'significant enough' to justify theater efforts to develop digital services should be at least 51% of the sample for services offered free-of-charged and at least 30% of the sample for paid theater services.

THE RESULTS

A. *User's interest in* subscribing to free-of-charge newsletters about theater is an important factor in market segmentation.

In the questionnaire the respondents were asked if they would be interested in subscribing to a newsletter about the theater. Among the group of 'potential audience' (n=102) 36% declared that they would not be willing to subscribe the e-newsletter (1 o 2 on the scale 1-5), 20% expressed a neutral statement (3 on the scale), and 44% stated that they are willing to accept the newsletter (4-5 on the scale). Further analysis showed that this factor - acceptance of newsletter subscription - can be used as a segmentation criterion of the potential audience. The group with lowest and the group of the highest acceptance of subscription perceived theater services differently (Table I). Therefore among the group of 'potential young audience' we identified the segment called 'art-oriented segment' (AO), and the segment of young people who are uninterested in classic theater and seems to be outside the possible reach as the theater audience ('out of target market' - OT).

TABLE II.
THE CORRELATION BETWEEN THE INTENSIONS TO USE FREE-OF-CHARGE THEATER SERVICE AND INTENTION TO PURCHASE THE SAME PAID SERVICE – SPEARMAN, $p<0.05$

	Young adults (general) (n=102)	**Art-oriented segment (n=45)**	**Filmschool students (n=44)**
e-book about theater	0.28	0.30	0.61
webseminar about theater	0.44	0.45	0.64
e-report about theater market (e.g. audience market research)	0.28	-	0.33
job posts connected with theater	0.30	-	-
gadgets	0.39	0.36	0.46
webseminar of acting	0.54	0.66	0.57
e-newsletter about theater	0.39	-	0.45

TABLE III.
WOULD YOU RECOMMEND THE THEATER WEBSITE IF…

		Negative (1-2)	Neutral (3)	Positive (4-5)	Average
the theater offers free services (these as indicated previously)	Art-oriented segment (n=45)	17.8%	8.9%	73.3%	4.0
	Out-of-target (n=37)	29.7%	8.1%	62.2%	3.7
	Filmschool (n=44)	18.2%	18.2%	63.6%	3.8
the theater offers paid services (these as indicated previously)	Art-oriented segment (n=45)	64.4%	20.0%	15.6%	2.2
	Out-of-target (n=37)	73.0%	18.9%	8.1%	1.9
	Filmschool (n=44)	59.1%	22.7%	18.2%	2.3

As the respondents were asked about their intention to visit theater websites on a regular basis (more visits, more consumed content in the month), this 'art-oriented' customers can be also regarded as 'the relationship-oriented' segment of internet users.

B. Hypothesis 1: The free-of-charge Internet services embedded on the theater website enhance customer engagement

The results indicate that a young audience can be attracted to the theater website as the result of offering free-of-charge theater services (Table I). Not surprisingly young people value the services as being useful in their potential jobs, such as webseminars about acting (public presentation is part of the modern job) or posts about jobs. Free gadgets attract the majority of users to the theater website. There are differences between the perception of website content by fimschool students and business students. It can be explained by access to other information sources with similar content (e.g. easier access to other materials connected with general art by filmschool students).

The percentage of art-oriented users attracted by theater services differs depending on the particular product, e.g. webseminars about theater attract about 30% of the potential audience; webseminars about acting, about 60%, with similar percentages for free-of-charge gadgets and e-books about theater. Among the art-oriented segment the average percentage of Internet-users willing to use free-of-charge services is 52% (compared to 35% of the "out of the target" segment).

C. Hypothesis 2: Free-of-charge Internet services embedded on theater websites enhance the sale of paid products.

Although the young audience is not very likely to buy services offered by the theater (Table I), the percentage of art-oriented internet-users willing to buy theater services ranges from 13% (webseminars about theater) to 44% (webseminars about acting). For this segment the average percentage of users willing to purchase theater services is 23%.

The correlation between the user's interest in free-of-charge services and willingness to purchase one is quite

TABLE IV.
THE PERCEPTION OF INTERACTIVE THEATER –WEBTHEATER. QUESTION: DO YOU THINK THAT TAKING PART IN WEBTHEATER YOU FEEL…

	Art-oriented segment (n=45)	Out of target (n=37)	Young audience (general) (n=102)	Filmschool Students (n=44)
The sense of belonging to theater community	3.2	2.5	2.9	3.0
Being immersed in storytelling	3.3	3.0	3.1	3.0
Reflection connected with art performance	3.5	2.8	3.2	3.8
Being with creative people	3.0	2.8	2.9	2.9
Taking part in special event	3.5	2.6	3.1	3.3
Intensive esthetic experience	3.0	2.5	2.8	2.7
Spending free time in meaningful way	3.6	3.5	3.6	3.5
Intensive emotion	3.5	2.8	3.3	2.9
The feeling of discovering the world	3.3	2.8	3.0	2.7
Intellectual challenge	3.4	3.1	3.2	3.1
Personal growth	3.8	3.3	3.6	3.0
Feeling being together with others during event	3.6	2.8	3.3	3.0

TABLE V.
QUESTIONS: DO YOU AGREE WITH THE FOLLOWING STATEMENTS

	Young audience (n=102)	**Art-oriented users (n=45)**		**Filmschool Students (n=44)**
	average (1-5)	average (1-5)	positive opinion (4-5)	average
The idea of webtheater sounds interesting to me	3.3	3.6	64.4%	3.4
I would like participate in such event	3.3	3.6	55.6%	3.3
Webtheater allows me to take part in theatrical performance more often	3.6	3.8	62.2%	3.2
Webtheater is a good idea for art promotion	3.6	3.6	62.2%	3.9
Webtheater encourages me to watch the performance in traditional theater	3.4	3.6	66.7%	3.1

strong among filmschool students; weaker among other respondents (Table II). It is especially worth noticing that offering free gadgets does not necessary lead to the purchase of one (there is a weak positive correlation). As an exception we can point to webseminars – in this case using free-of-charge webseminars about actors performing can probably lead to buying a paid one.

D. Hypothesis 3: The free-of-charge Internet services embedded on theater websites enhance Internet-user referrals

About 73% of Internet users from the 'art-oriented' segment declared that they recommend the theater website to others if the website includes free-of-charge services. The percentage drops to about 16% in the case of offering paid products (Table III).

E. Hypothesis 4: The new digital interactive artform *'webtheater' enhances the customer's* willingness to buy traditional tickets

The digital theater (webtheater) was assessed by respondents in two areas: their perception of the expected experience connected with attending a digital theater (Table IV) and their general assessment of this idea (Table V).

The respondents assessed their expected experience in 12 dimensions (see - Walmsley 2011). The art-oriented viewers valued their experience highly in the following dimensions: personal growth, the feeling being together with others during the event, spending free time in a meaningful way, taking part in special event, intensive emotions and reflection connected with art performance. This aspect of digital webtheater needed to be explored more, and further, with much deeper research. For our purpose, the main conclusion is that the idea of a new digital interactive theater was not rejected by the young possible audience.

In the second area, respondents were asked if the

TABLE VI.
SUMMARIZED RESULTS

	Out-of-target segment	**Art-oriented users**	**Filmschool students**
H1: The free-of-charge Internet services embedded on theater website enhance customer's engagement [at least 51% of the sample expressed a positive opinion towards using particular free-of-charge service]	Confirmed for gadgets	Confirmed for e-book, gadgets, jobs posts, webseminar about acting	Confirmed for job posts, e-newsletters
H2: Free-of-charge Internet services embedded on theater website enhance the sale of the paid products [at least 30% of the sample expressed a positive opinion towards buying paid content and there is at least a medium positive correlation between using the free-of-charge content and the intention to purchase]	Rejected	Confirmed for webseminars about acting	Rejected
H3: The free-of-charge Internet services embedded on theater website enhance internet user's referrals [at least 51% of the sample expressed a positive opinion]	Confirmed	Confirmed	Confirmed
H4: The new digital interactive artfom 'webtheater' enhances the customer willingness to buy traditional tickets [at least 51% of the sample expressed a positive attitude about watching traditional performance]	Rejected	Confirmed	Rejected

webtheater influences their willingness to attend traditional theater performances. About 67% of the art-oriented users declared that the digital performance sparks their interest in visiting a traditional theater.

CONCLUSIONS

The summarized results of empirical study (Table VI) suggest that the theater strategy of audience broadening (attracting a new group of audience) can be based on offering free-of-charge services and offering the new artform of webtheater. The main conclusions are:

- The digital theater services mainly attracted the 'art-oriented' young people, and this segment of potential digital audience can also be perceived as the 'relationship-oriented' segment.
- The results underline the importance of the e-newsletter as a digital tool aiming to create a relationship with a patron.
- The sale of paid content is not driven by offering free-of-charge content. At this point of the research we may presume that customer equity for theater website may be based mostly on indirect values such as a patron's referrals and engagement.
- Free-of-charge digital content on the theater website can influence positive word-of-mouth referrals.

However, the presented results should be further investigated as our research has at least the following limitations:

- The data is based on the convenient sample, and therefore it is difficult to generalize the results for a larger population. On the other hand, the Internet rarely allows researchers to base a survey on a more controlled sample (with the possibility of verifying the identity of the user).
- As we tested the concept of digital services we did not include the cost connected with the digital offer.
- We based this study on the assumption that a declared intention to buy leads to a real purchase.
- We did not differentiate between the theater brand and the theater website brand, assuming that these constructs are very close related;
- We did not include socio- and demographical profiles of respondents into the analysis e.g. current consumption of art, the income or the place of living.

Customer equity management for the cultural sector is the under-researched area, in which deeper theoretical and empirical research is needed. We consider of particular importance, further investigation into the following problems:

- Developing the framework outlined in this paper, aiming at building a CLV model - which would be applicable for theaters.
- Identification of the spectrum of drivers of customer equity in theaters;
- Further research about the concept of webtheater. This new product – allowing the patrons to shape the digital theater performance - raises the question not only about the possible income connected with the sale of tickets, but also about the patron's new experience.
- The connection between Internet-audience equity and traditional theater-audience equity.

REFERENCES

[1] Bakhshi H., Throsby D. (2012), "New technologies in cultural institutions: theory, evidence and policy implementation", International Journal of Cultural Policy, Vol. 18, No. 2, March 2012, 205-222.

[2] Bayón T., Gutsche J., Bauer H. (2002), "Customer Equity Marketing: Touching the Intangible", European Management Journal, Jun2002, Vol. 20, No. 3, 213-222.

[3] Bick G. N. C. (2009), "Increasing shareholder value through building Customer and Brand Equity", Journal of Marketing Management, Vol. 25, No .1-2, 117-141.

[4] Blattberg R. C., Deighton J. (1996), "Manage Marketing by the Customer Equity Test", Harvard Business Review, Jul/Aug 1996, Vol. 74, Issue 4, 136-144.

[5] Hausmann A., Poellmann L. (2013), "Using social media for arts marketing: theoretical analysis and empirical insights for performing arts organizations", International Review on Public and Nonprofit Marketing, (2013), Vol. 10, 143-161.

[6] Hogan J. E., Lemon K. N., Rust R. T. (2002), "Customer Equity Management. Charting New Directions for the Future of Marketing", Journal of Service Research, Vol. 5, No. 1, August 2002, 4-12.

[7] Instytut Teatralny "Raport – Badanie publiczności teatrów w stolicy" 2012; [http://www.instytut-teatralny.pl/projekty/raporty; 05.04.2014].

[8] Kossecki P. (2007), „Wartość życiowa klientów i jej zastosowanie do wyceny przedsiębiorstw internetowych", Problemy *Zarządzania*, 3/2007 (17), 78-88.

[9] Kossecki P. (2011) „Kreowanie i pomiar wartości przedsiębiorstwa w świecie Internetu", Wydawnictwo Państwowej Wyższej Szkoły Filmowej, Telewizyjnej i Teatralnej im. L. Schillera w Łodzi.

[10] Kumar V., George M. (2007), "Measuring and Maximizing customer equity: a critical analysis", Journal of the Academy of Marketing Science, (2007), Vol. 35, 157-171.

[11] Leone R. P., Rao V.R., Keller K. L., Luo A. M., McAlister L., Srivastava R. (2006), "Linking Brand Equity to Customer Equity", Journal of Service Research, November 2006, Vol. 9, No. 2, 125-138.

[12] Preece S. B, Wiggins Johnson J. (2011), "Web Strategies and the Performing Arts: A Solution to Difficult Brands", International Journal of Arts Management, Vol. 14, No. 1, Fall 2011, 19-31.

[13] Quero M. J. (2007), "Relationship Marketing And Services Marketing: Two Convergent Perspectives For Value Creation In The Cultural Sector. Empirical Evidence On Performing Arts Consumers In Spain", International Review on Public and Non Profit Marketing, Vol. 4, No 1/2 (December 2007), 101-115

[14] Rentschler R., Radbourne J., Carr R., Rickard J. (2002), Relationship marketing, audience retention and performing arts organization viability", International Journal of Nonprofit and Voluntary Sector Marketing, Vol. 7, No. 2, 118-130.

[15] Rust R. T.; Lemon K. N.; Zeithaml,V. A. (2004), "Return on Marketing: Using Customer Equity to Focus Marketing Strategy", Journal of Marketing, Jan2004, Vol. 68 Issue 1, 109-127.

[16] Walmsley B. (2011), "Why people go to the theater: A qualitative study of audience motivation", Journal of Customer Behaviour, 2011, Vol. 10, No. 4, 335-351.

3rd Workshop on Information Technologies for Logistics

THE main purpose of the workshop is to provide a forum for researchers and practitioners to present and discuss current issues concerning use of ICT in logistic applications (hardware and software). There will be also an opportunity for hardware integrators, software developers and logistics companies to demonstrate their solutions, as well as achievements, in different logistic systems.

Topics

The topics of interest include but are not limited to:

- Innovations in information systems supporting logistics and its management (WMS, SCM, TMS, LIS, VMI, CRP, PLM, and others)
- Innovative technologies in warehouse management: RFID, Voice Picking, Image Recognition, Pick Radar, etc.
- Logistics process modeling, including influence of warehouse automatic
- Optimization of logistics processes:
  - optimal vehicle routing and management, boundary conditions
  - optimal picking routing (global optimization, fast search, collision prediction and prevention)
  - shared mobility systems
  - day-to-day dynamic traffic assignment models
  - effective methods of picking (multi picking, batch picking ect.)
  - relationships between picking efficiency and products decomposition in warehouse area
- Environmental protection (for example carbon-aware transportation)
- Artificial intelligence systems and decision support systems in logistics
- BI, data mining and process mining in logistics
- Quality management algorithms and methods
- Material Flow Theory and applications

Event Chairs

Gontar, Beata, Uniwersity of Łódz, Poland

Program Committee

Azavedo, Susana, University of Beira Interior, Portugal
Balicki, Jerzy, Gdansk University of Technology, Poland
Banaszak, Zbigniew, Warsaw University of Technology, Poland
Bobkowska, Anna, Gdansk University of Technology, Poland
Bruzda, Jaonna, Nicolaus Copernicus University, Poland
Duran-Grados, Vanesa, University Cadiz, Spain
Fosner, Maja, Faculty of Logistics, University of Maribor, Slovenia
Franczyk, Bogdan, Universitat Leipzig, Germany
Goh, Thong Ngee, National University of Singapore
Gontar, Zbigniew, University of Lodz, Poland
Jedliński, Mariusz, University of Szczecin, Poland
Korczak, Jerzy, Uniwersytet Ekonomiczny we Wrocławiu, Poland
Langviniene, Neringa, Kaunas University of Technology, Lithuania
Lent, Bogdan, University of Bern, Switzerland
Liao, Da-Yin, National Chi Nan University, Taiwan
Malavasi, Gabriele, University of Rome, Italy
Montemanni, Roberto, University of Applied Sciences of Southern Switzerland, Switzerland
Pamuła, Anna, University of Łódź, Poland
Patasiene, Irena, Kaunas University of Technology, Lithuania
Ricci, Stefano, SAPIENZA Università di Roma, Italy
Semenov, Louri, West Pomeranian University of Technology, Poland
Shinkevich, Алексей Иванович, Kazan National Research Technological University
Sitek, Pawel, Kielce University of Technology, Poland
Tipi, Nicoleta, University of Huddersfield, United Kingdom
Zielinski, Jerzy, University of Lodz, Poland

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1287–1291
DOI: 10.15439/2014F309
ACSIS, Vol. 2

Task Assignments in Logistics by Adaptive Multi-Criterion Evolutionary Algorithm with Elitist Selection

Jerzy Balicki

Gdańsk University of Technology,
Faculty of Electronics,
Telecommunications and
Informatics ul. Narutowicza 11/12,
81-230 Gdańsk, Poland
Email: balicki@eti.pg.gda.pl

Abstract— An evolutionary algorithm with elitist selection and an immunological procedure has been developed for Pareto task assignment optimization in logistics. A multi-criterion optimization problem has been formulated for finding a set of efficient alternatives. Some criteria have been applied for evaluation of solutions: bottleneck machine workload, a machine cost, and a system performance. Moreover, some numerical experiments have been performed and the machine constraints have been respected.

I. Introduction

Genetic algorithms, evolutionary algorithms and evolution strategies are the alternative evolutionary approaches to the other meta-heuristic multicriteria optimization methods such as simulated annealing, immunological systems [1, 6], tabu search [12], scatter search or Hopfield neural networks [4]. Evolutionary calculations process simultaneously a solution population, which permits finding a subset of P-optimal alternatives by one run as a replacement for several isolated runs of the other multiobjective optimization techniques [7]. From this reason evolution approaches are convenient, if we look for the subset of Pareto-optimal solutions [16, 17].

Experimental outcomes demonstrate that elitism can increase performance of multi-objective evolutionary algorithms radically [18]. Moreover, elitism avoids the damage of non-dominated alternatives, if they have been established. A concept of elitism for multi-criterion evolutionary algorithms is taken from evolution strategies regarding evolution strategy developed for combinatorial and multi-objective optimization problems [5]. The other evolution strategy with an archive for finding P-optimal solutions has been suggested in [13].

In this paper, a problem of task allocation in logistics has been verbalized as a combinatorial and multi-objective optimization question characterized by some partial criteria: a machine cost, a bottleneck machine workload, and the system performance. Moreover, three kinds of constraints have been considered. The first constraint sort is related with the assumption that each task should be allocated to the machine, and the second one – with assumption one and only one machine should be allocated to a place of the task performing. Furthermore, the resource constraints are respected.

II. Evaluation Criteria

We assume that some important logistic tasks are performed by some machines, automatically. A bottleneck machine is characterized by the heaviest logistic task load. A bottleneck workload is should be minimized as a critical factor that can balance a whole load among elements. The weight of the bottleneck machine is calculated, as below:

$$Z_{max}(x) = \max_{i\in\overline{1,I}}\left\{\sum_{j=1}^{J}\sum_{v=1}^{V} t_{vj}\, x_{vi}^{task} x_{ij}^{p} + \sum_{\substack{v=1\\ u\neq v}}^{V}\sum_{\substack{u=1\\ k\neq i}}^{V}\sum_{i=1}^{I}\sum_{k=1}^{I} \tau_{ikvu}\, x_{vi}^{task} x_{uk}^{task}\right\}, \quad (1)$$

where

t_{vj} – a time of the overhead performing for the task number v by the machine sort number j,

τ_{ikvu} – a time of a resource transport between the task number v at the place w_i and the task number u at the place w_k.

$$x = [x_{11}^{task},\ldots, x_{vi}^{task},\ldots, x_{VI}^{task}, x_{11}^{p},\ldots, x_{ij}^{p},\ldots, x_{IJ}^{p}]^{T},$$

$$x_{ij}^{p} = \begin{cases} 1 \text{ if } p_j \text{ is assigned to the } w_i, \\ 0 \text{ in the other case.} \end{cases} \quad i=\overline{1,I},\ j=\overline{1,J},$$

$$x_{vi}^{task} = \begin{cases} 1 \text{ if task } T_V \text{ is assigned to } w_i, \\ 0 \text{ in the other case,} \end{cases} \quad v=\overline{1,V},\ i=\overline{1,I}.$$

The machine cost is determined regarding the formula, as below:

$$K(x) = \sum_{i=1}^{I}\sum_{j=1}^{J} \alpha_j x_{ij}^{\pi}, \quad (2)$$

where α_j is the machine cost for the sort number j.

The total machine performance is calculated, as follows:

$$\Theta(x)=\sum_{i=1}^{I}\sum_{j=1}^{J}\beta_j x_{ij}^{p}, \quad (3)$$

where β_j is the logistic machine performance for its sort number j.

III. Mutiobjective Optimization Problem

An optimal configuration of task in a logistic system that can be modeled as a task assignment may reduce the total cost of a set of tasks execution or the workload of bottleneck machine. It can decrease the cost of machines because of the machine sort selection, too. A total amount of system performance is another measure that can be maximized by task scattering and by the machine sort selection. An advantage of the system with an optimal task assignment may exceed 50% value of any criterion for a system with a task scattering designed without an optimization technique [2]. The bottleneck machine load is assessment criterion of the system configuration that causes the load balance and also minimizes a response time [3].

In that problem, the admissible solution satisfies three classes of constraints. Because each unit is allocated to one node, the logistic task allocation constraints are devised, as below:

$$\sum_{i=1}^{I} x_{vi}^{task}=1,\ v=\overline{1,V}. \quad (4)$$

We assume one and only one machine should be allocated at each node. It implies the machine allocation constraints, as follows:

$$\sum_{j=1}^{J} x_{ij}^{p}=1,\ i=\overline{1,I}. \quad (5)$$

Each machine provides some resource capacities to perform some assigned tasks. Let some resources $z_1,...,z_l,...,z_L$ be required in a logistic system. We introduce μ_{jl} to represent the lth resource capacity in the machine p_j. We assume the task number v holds η_{vl} units of z_l. The values μ_{jl} and η_{vl} are nonnegative and limited.

The resource capacity limit in any machine in the ith place cannot be exceeded, what can be written, as bellows:

$$\sum_{v=1}^{V}\eta_{vl}x_{vi}^{task}\le\sum_{j=1}^{J}\mu_{jl}x_{ij}^{p},\ \ l=\overline{1,L},\ i=\overline{1,I}. \quad (6)$$

The multiobjective optimization problem may be established a triple $(\mathcal{X}, F, P)$ to find the Pareto representation of some optimal solutions, as follows [15]:

1) $\mathcal{X}$ - an admissible solution set

$$\mathcal{X}=\{x\in\mathcal{B}^{I(V+J)}\mid \sum_{j=1}^{J}x_{ij}^{p}=1,\ i=\overline{1,I};$$
$$\sum_{i=1}^{I}x_{vi}^{task}=1,\ v=\overline{1,V};$$
$$\sum_{v=1}^{V}\eta_{vl}x_{vi}^{task}\le\sum_{j=1}^{J}\mu_{jl}x_{ij}^{p},\ \ l=\overline{1,L},\ i=\overline{1,I}\}$$

$\mathcal{B}=\{0, 1\}$

2) f – a multi-objective optimization criterion

$$f:\mathcal{X}\to\mathcal{R}^3, \quad (7)$$

where $f(x) = [K(x), \Theta(x), Z_{max}(x)]^T$, $x\in\mathcal{X}$

3) P - the relationship of optimization preferences [14]

IV. Mutiobjective Optimization Algorithms

Some advanced evolutionary algorithms have been developed for several multi-objective optimization problems [8, 9, 11]. What is more, some of them have been tested for finding the set of Pareto-optimal task assignments [2, 3].

A ranking idea for non-dominated individuals was introduced to avoid the prejudice of the interior Pareto solutions [10]. Then, the first algorithm called NSGA with the ranking procedure has built on the ideas mentioned by Goldberg [15].

In a current population, some non-dominated individuals get a rank equal to 1. Then, the second level of non-dominated alternatives is assigned the rank 2. This assigning procedure is recurred until the population is preceded. It is worth to mention that all non-dominated individuals have the same reproduction fitness because of the equivalent rank.

Deb et al. have improved NSGA by introduction an elitist procedure [8]. In an evolutionary algorithm called NSGA-II, a selection of potential parents is based on a binary tournament. Both an offspring population and a parent population are combined to select a new parent population. If two chromosomes are characterized by the other ranks, it with smaller rank is preferred. If individuals have the same rank, there is preferred the logistic configuration of tasks in a less overcrowded region.

V. Adaptive Multi-Criterion Evolutionary Algorithm with Elitist Selection

Adaptive evolutionary algorithm with elitist selection called AMEA+ is noticed as an advanced optimization technique for multi-criterion logistic task assignment [3]. Figure 1 shows a diagram of AMEA+. The preliminary set of chromosomes is erected to satisfy constraints (4) and (5)

(Fig. 1, line 3). Generated individuals are constructed by integer coding, as below:

$$X = (X_1^{task},..., X_v^{task},..., X_V^{task}, X_1^p,..., X_i^p,..., X_I^p), \quad (8)$$

where $X_v^{task} = i$ if $x_{vi}^{task} = 1$ and $X_i^p = j$ if $x_{ij}^p = 1$. Besides, $1 \le X_v^{task} \le I$ and $1 \le X_i^p \le J$.

Fitness function is calculated for some admissible solutions (Fig. 1, line 4), as below:

$$F(x) = P_{max} - \lambda(x) + \lambda_{max} + 1, \quad (9)$$

where $\lambda(x)$ is the admissible solution rank, $1 \le \lambda(x) \le \lambda_{max}$.

1. BEGIN
2. φ:=0, enter ψ – the size of chromosome population, γ – the chromosome length, p_m:=$(\psi\ \gamma)^{-1}$;
3. Create a preliminary set **Pop**(φ), **M**(φ):= **Pop**(φ);
4. Compute ranks $\lambda(x)$ and values of fitness $F(x)$, $x \in$ **Pop**(φ)
5. work:=TRUE
6. WHILE work DO
7. BEGIN /* create a new population */
8. φ:= φ +1, $\mathbf{Pop}(\varphi) := \varnothing$
9. Estimate probabilities $p_s(x)$, $x \in \mathbf{P}(\varphi - 1)$
10. FOR ψ/2 DO
11. BEGIN /* reproduction cycle */
12. 2WT potential parent selection (**a,b**) from **P**(φ -1)
13. S-crossover of a pair (**a,b**) with the adaptive crossover rate $p_c := e^{-\varphi/T_{max}}$
14. S-mutation of an offspring pair (**a',b'**) with p_m
15. **P**(t):=**P**(t) $\cup$ (**a',b'**}
16. END
17. **P**(φ):= **M**(φ) $\cup$ **P**(φ)
18. Compute ranks $\lambda(x)$ and values of fitness $F(x)$, $x \in \mathbf{P}(\varphi)$
19. An elitist selection of ψ solutions with the largest F(x) in **P(t)**; if more than ψ items have the same rank, use the crowd measure to select ψ solutions;
20. IF ($\varphi \ge T_{max}$ OR **P**(φ) converges) THEN work:=FALSE
21. END
22. END

Fig. 1. A diagram of an adaptive multi-criteria genetic algorithm with elitist selection

The 2WT potential parent selection is the two-weight tournament because two times the roulette rule is made.

The chromosome crossover point is randomly selected and two offspring are formed regarding $p_c := e^{-\varphi/T_{max}}$. The first part of the parent **a** is concatenated with the second part of the parent **b**. Similarly, the first part of the parent **b** is concatenated with the second part of the parent **a**. Each pair of potential individuals is randomly chosen for crossover with the probability pc. If chromosomes are not taken for crossover, parents are transferred to a set of offspring. A crossover rate decreases during the progressing of evolution. So, the changes of a search area retire slowly.

The S-mutation is based on the integer random modification by another feasible value. For X_v^{task}, the set $\{1,..., I\}$ is considered, and for X_i^p, the set $\{1,..., J\}$ is adequate. A constant mutation rate is assigned.

A search space for the considered evolutionary algorithm consists of $I^V J^I$ elements. It can be proved that S-crossover and S-mutation give ability for obtaining each solution in the search space. So, we can expect to find non-dominated assignments after an exhausted search.

Some numerical experiments demonstrate that elitism may improve the quality of alternatives [13, 14, 18]. An improved elitist selection is carried out as follows. Let M(t) be an old population P(t-1) and P(t) be a new population created from M(t) by mating, crossover and mutation. Firstly, a sum of two populations M(t) $\cup$ P(t) is created. Secondly, ranks are calculated for an entire population M(t) $\cup$ P(t). Each non-dominated solution in the extended set is characterized by the fitness value regarding its rank. Finally, L solutions are qualified with the higher fitness f (x) to P(t).

Let T_{max} be a maximal number of new populations that is O(n), where n = max{I, V, J}, I is O(J), and V is O(J). Moreover, let a size L of population be O(n), too. Now, we can assess a complexity of the AMEA-II. Fitness (Fig. 1, line 18) is calculated $O(n^2)$ times what gives the complexity $O(n^6)$ for the AMEA-II applied to the multi-objective logistic problem.

VI. Convergence Measure

The convergence of the studied algorithm can be considered by measuring the quality of obtained logistic task assignments to the Pareto front. So, we introduce a closeness measure for the obtained efficient points to the given Pareto points $\{P_1, P_2,..., P_U\}$. They may be found by enumerative search for small instances. An evolutionary algorithm can find the set of sub-efficient points $\{A_1, A_2,..., A_{U'}\}$, where $U' \le U$. If there is an outcome $A_u=(A_{u1}, P_{u2}, A_{u3})$ with the same cost of computers as the uth Pareto result $P_u=(P_{u1}, P_{u2}, P_{u3})$, then the distance between these points is equal to $\sqrt{(P_{u1} - A_{u1})^2 + (P_{u4} - A_{u4})^2}$. If A_u is missed by the AMEA+, then the distance $\sqrt{\sum_{\substack{i=1 \\ i \ne 2}}^{3} (P_{ui} - A_{ui}^-)^2}$ is considered, where A_{u1}^- is the maximal load of the bottleneck machine, and A_{u3}^- is the minimal performance of machines, for the instance of the problem (7).

Some ξ initial populations are generated, and ξ values of the distances S_u^1 are calculated. A convergence criterion to the P-set is calculated, as below:

$$S = \frac{1}{\xi}\sum_{l=1}^{\xi}\sum_{u=1}^{U} S_u^1. \qquad (10)$$

The AMEA+ gives better results than AMEA without an elitist procedure. In the 200 iteration, S is 1.3% for the AMEA+ and 1.6% for the AMEA. Above algorithms have been run 30 times.

For the instance with 2 places, 10 tasks, and 5 machines types, 30 binary variables generate 1 073 741 824 solutions in the search space. Θ is a value from [200, 600], K is from [2, 10] [money unit], and Z_{max} is from [26; 75] [time unit].

A formula $I\,^{V}J\,^{I}$ permits to calculate an upper bound of the number of an admissible set. Figure 2 displays the two criteria space of evaluations. The ideal point y^o and the anti-ideal point y^- can be used for finding a compromise solutions.

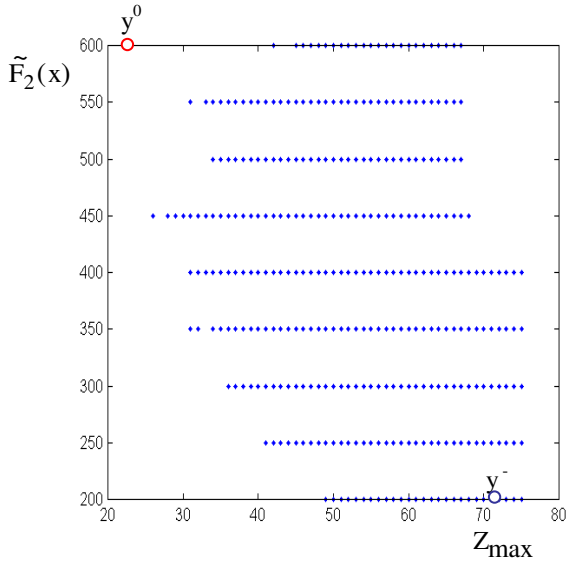


Fig. 2. Bi-criteria evaluation space

VII. Concluding Remarks

The AMEA+ is capable techniques for solving a multiobjective optimization problems focused on finding logistic task allocations that minimize the cost of machines, a workload of the bottleneck machine, and maximize performance of logistic system.

Every one of non-dominated solutions in the combined population is assigned a fitness based on the rank of solutions. Dominated solutions are assigned fitness worse than the worst fitness of any non-dominated solutions in a population. This assignment of fitness makes sure that the search is directed towards the non-dominated solutions, too.

Our future works will focus on finding the combination the multicriteria evolutionary algorithm with the immunological algorithm to include some ranking procedures to handle constraints and for improving the obtained Pareto-optimal task assignments.

References

1. Aragon V. S., Esquivel S. C.; Coello Coello C. A.: A modified version of a T-Cell Algorithm for constrained optimization problems. International Journal for Numerical Methods in Engineering, Vol. 84, Issue: 3, 2010, pp. 351-378.
2. Balicki J.: An Adaptive Quantum-based Multiobjective Evolutionary Algorithm for Efficient Task Assignment in Distributed Systems. Proceedings of The 13th WSEAS International Conference on Computers The 13th WSEAS Multiconference on Circuits, Systems, Communications and Computers, July22-26, 2009, Rodos Island, Greece, WSEAS Press, pp. 417-422
3. Balicki J.: Negative selection with ranking procedure in tabu-based multi-criterion evolutionary algorithm for task assignment. In: V.N. Alexandrov et al. (Eds.): Computational Science – ICCS 2006, Proceedings of the 6th International Conference on Computational Science, Reading, UK, May 28-31, 2006, Lecture Notes in Computer Science, vol. 3993, Springer-Verlag, Berlin Heidelberg New York, 2006, pp. 863-870
4. Balicki J., Kitowski Z., Stateczny A.: Extended Hopfield Model of Neural Networks for Combinatorial Multiobjective Optimization Problems. Proceedings of 1998 IEEE World Congress on Computational Intelligence, Anchorage, May 4-9, 1998, Vol. 2, pp. 1646-1651.
5. Binh T. T., Korn U., (1997) Multiobjective evolution strategy for constrained optimization problems. Proceedings of the 15th IMACS World Congress on Scientific Computation, Modelling and Applied Mathematics, Berlin, 357-362.
6. Canova A., Freschi F.: Multiobjective design optimization and Pareto front analysis of a radial eddy current coupler. International Journal of Applied Electromagnetics and Mechanics, Vol. 32, Issue: 4, 2010, pp. 219-236.
7. Coello Coello C. A., (1999) A comprehensive survey of evolutionary-based multiobjective optimization techniques, Knowledge and Information Systems. An International Journal, **1**, 269-308.
8. Deb K., Agrawal S., Pratap, A., Meyarivan T. (2000) A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. Technical Raport No. 200001, Indian Institute of Technology, Kanpur.
9. Freschi F., Repetto M.: VIS: An artificial immune network for multi-objective optimization. Engineering Optimization, Vol. 38 Issue: 8, 2006, pp. 975-996.
10. Goldberg D. E., (1989) Genetic algorithms in search, optimization, and machine learning, Addison-Wesley Publishing Company, Inc. Reading, Massachusetts.
11. Hiremath N. C., Sahu S., Tiwari M. K.: Designing a Multi Echelon Flexible Logistics Network using Co-evolutionary Immune-particle Swarm Optimization with Penetrated Hyper-mutation (COIPSO-PHM). Proc. of the 2nd International Conference on Mechanical and Aerospace Engineering, Bangkok, July 29-31, 2011, Book Series: Applied Mechanics and Materials, Vol. 110-116, 2012, pp. 3713-3719.
12. Jozefowska J., Waligora G., Weglarz J., (1998) Tabu search with the cancellation sequence method for a class of discrete - continuous scheduling problems. Fifth International Symposium on Methods and Models in Automation and Robotics, Miedzyzdroje, 1047 - 1052.
13. Knowles J., Corne D. W., (2000) Approximating the nondominated front using the Pareto archived evolution strategy, Evolutionary Computation, **8**, 2, 149-172,
14. Sinha A. K., Zhang W. J., Tiwari M. K.: Co-evolutionary immuno-particle swarm optimization with penetrated hyper-mutation for distributed inventory replenishment. Engineering Applications of Artificial Intelligence, Vol. 25, Issue: 8, 2012, pp. 1628-1643.
15. Srinivas N, Deb K., (1994) Multiobjective optimization using non-dominated sorting in genetic algorithms, Evolutionary Computation, **2**, 3, 221-248.

16. Ulungu E.L., Teghem J., (1994) Multi-objective combinatorial optimization problems: A survey, *Journal of Multi-Criteria Decision Analysis,* 3, 83-104.
17. Van Veldhuizen D. V., Lamont G. B., (2000) Multiobjective evolutionary algorithms: Analyzing the state-of-the-art, *Evolutionary Computation,* 8, 2, 125-147.
18. Zitzler E., Deb K., Thiele L., (2000) Comparison of multiobjective evolutionary algorithms: Empirical results, *Evolutionary Computation,* 8, 2, 173-195.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1293–1299
DOI: 10.15439/2014F226
ACSIS, Vol. 2

Using UAVs for Remote Study of ice in the Arctic with a View to Laying the Optimal Route Vessel

Alexey Lagunov
Northern (Arctic) Federal University
named after M. V. Lomonosov
ul. Severnaya Dvina Emb. 17,
Arkhangelsk, Russia
E-mail: a.lagunov@narfu.ru

Dmitry Fedin
Northern (Arctic) Federal University
named after M. V. Lomonosov
ul. Severnaya Dvina Emb. 17,
Arkhangelsk, Russia
E-mail: d.fedin@narfu.ru

Anatoliy Tyagunin
Northern (Arctic) Federal University
named after M. V. Lomonosov
ul. Severnaya Dvina Emb. 17,
Arkhangelsk, Russia
E-mail: av.tyagunin@yandex.ru

***Abstract*—Arctic Development is one of the strategic goals of Russia and other circumpolar countries. Solution of this problem is faced with the problem: most of the coastal seas and the Arctic Ocean is covered with ice. For large vessels and icebreakers problem is solved by applying remote sensing of ice with spacecraft and aircraft. This solution has a very high cost and low efficiency. The authors propose to use for this purpose unmanned aerial vehicles (UAVs). This paper describes the solution of choice of UAV. This paper describes the solution of choice UAV. The features of the use of equipment for photo and video and microwave sensing of ice.**

I. Introduction

THE Arctic is extremely important military-strategic significance for Russia. Mission of Northern (Arctic) Federal University named after M.V. Lomonosov - formation and development of competitive human capital in the North-West Federal District through the creation and implementation of innovative services, development from the perspectives of the development of the Russian North and the Arctic. To achieve this mission NArFU has highly skilled specialists, high-tech equipment, the ability to introduce new developments in the production and achieve the desired result. Russian Government commissioned a comprehensive research in the Arctic is the NArFU

Arctic exploration is largely hampered by the presence of ice in the Arctic Ocean and adjacent seas. The problem of predicting the state of the ice: its thickness, density, size of the ice fields - is one of the most important issues of today. Methods of direct examination by the state of the ice landing on the surface is not always applicable for reasons of safety. For forecasting ice conditions are more commonly used methods of distance study.

Remote sensing - sensing the Earth's surface is on board aircraft or equipment from space using the properties of electromagnetic waves emitted, reflected or scattered by the sensed objects, for the purpose of providing information for disaster management, improve management of natural resources land use management and environmental protection.

Remote Sensing - a set of methods for study of the atmosphere, the earth's surface, the oceans, the top layer of the earth's crust by air aerospace methods based on the decoding of images obtained at a distance from the aircraft.

The result is:

1. Aerial photographs obtained from a height of preferably 500 m up to 10 km, but not more than 30 km;
2. Space images - from a height of 150 km.

Operating regions for the filming freeze commonly use the following satellites:

1. The radar KA RADARSAT-1 and ENVISAT-1 to provide a guaranteed all-weather shooting at any time of day.
2. Highly optical satellites EROS A/B for a detailed analysis of regions ice and flooding, SPOT multispectral optical unit 4 for the analysis of ice on rivers, ice crossings and in areas of potential ice blocks.

All remote sensing techniques for observing the Earth's surface use electromagnetic (EM) radiation in the visible/near-infrared (VNIR), thermal infrared (TIR) or microwave bands of the electromagnetic spectrum. There are three principal measurement techniques applicable to ocean and sea ice observations:

1. Measurement of the part of the incoming solar radiation that is reflected at the surface of the Earth (visual and near-infrared remote sensing).
2. Measurement of the thermal radiation from the surface (thermal infrared and passive microwave remote sensing).
3. Measurement of the return signal from an active source, such as a microwave radar, using several types of instruments that measure backscattered radiation from the surface.

The sea ice surface can be considered to have three macroscopic components:

1. Open water in leads, polynyas and melt ponds on top of the ice in summer.
2. Ice with varying amounts of salt water intrusions (salt pockets).
3. Snow on top of the ice.

978-83-60810-58-3/$25.00

The following properties of the macroscopic components have impact on the remote sensing measurements:

1. Percentage and distribution of the three components.

2. Temperature of the components.

3. Salinity of the components and the distribution of salt intrusions in the ice.

4. Crystal structure of the ice and the snow.

5. Occurrence of snow and ice layers, and rough surfaces (e.g., floes, ridges, frost flowers).

In addition, there are variables defined by the remote sensing instrument that also have significant impact on the measurements:

1. Frequency (wavelength) of the radiation.

2. Angle of incidence of the radiation.

3. Polarization of the radiation if the view is not from directly above.

The main disadvantages of the methods at work in the Arctic:

1. The use of space or air assets has a high cost, so it can be used only for large vessels (nuclear icebreakers, military ships, tankers). For smaller vessels, which often carry out scientific research and commercial activities (fishing, harvesting of marine animals, environmental activities) use of these funds because of high prices is impossible.

2. Low levels of efficiency: for taking pictures requires considerable time. Receive data in real time is difficult.

3Files aerial and satellite images are quite large, making it difficult to transmit this information via communication channels. In the Arctic, there are problems with communication, mainly by satellite, which is unsustainable. This leads to the fact that the communication channels are quite narrow and has a high data transfer rates [1].

We propose to explore the question of the use of unmanned aerial vehicles (UAV).

II. UNMANNED AERIAL VEHICLES

Stimulus to the development of unmanned aircraft in the world was the successful and widespread use of UAVs by the armies of the United States and Israel during the military operations (Persian Gulf, Yugoslavia, the Middle East, the Arab-Israeli wars). While drones have proven to be an effective means of intelligence, combat support, as false targets for the detection of enemy air defense systems, delivery of goods, to perform other missions.

UAV - in general, this aircraft without crew on board. The notion of an aircraft includes a large number of types, each of which has its analogue drone. In this article the definition of UAV misses the following concept: aircraft without crew on board, using the principle of creating aerodynamic lift by fixed or rotary wing (UAV aircraft and helicopter type), equipped with an engine and having a payload and flight duration, sufficient to perform specific tasks.

One of the major problems with UAVs - management of the unit.

There are the following control methods:

1. Manual control of the operator (or remote piloting) with remote control within optical observability or specific information from the forward-looking video camera. This control operator primarily solves the problem of piloting: maintaining a desired course, altitude, etc.

2. Automatic control provides the ability to fully autonomous UAV flight along a predetermined path at a predetermined height at a given rate and the stabilization of the orientation angles. Automatic operation via onboard software devices.

3. Semi-automatic control (or remote control) — flight is performed automatically without human intervention using the autopilot initially set parameters, but the operator can make changes to the route online. Thus, the operator has the ability to influence the outcome of the operation that is focused on the task of piloting.

Manual control can be one of the modes for the UAV, and may be the only way to manage. UAVs, deprived of any means of automatic flight control.

Last two methods are currently the most popular among users of unmanned systems, as impose the least training requirements and ensure safe and efficient operation of UAV systems Fully automatic control can be an optimal solution for a given section of aerial tasks when you need to shoot at a great distance from home base is in contact with the ground station. At the same time, as the responsible person for the flight, the launching, the opportunity to influence the flight from the ground station can help avoid emergency situations.

To perform specific tasks, particularly for aerial photography and remote sensing, UAV should be considered in conjunction with its instrumentation and payload, which introduced the term unmanned aircraft system (UAS).

UAS, in addition to the UAV consists of onboard control complex, payload and ground control station.

1. Onboard complex:

- Integrated Navigation System;
- Satellite Navigation System Receiver;
- Autopilot. Autopilot problem:

a) Piloting: automatic flight on a given route, automatic takeoff and landing approach, maintaining the desired altitude and airspeed, the stabilization of the orientation angles, forced landing in the event of engine failure or other serious problems.

b) Software control onboard systems and payloads, such as camcorder stabilization and synchronization time and coordinates of the camera shutter, release the parachute.

- Drive flight information.

2. To payload for remote studies of ice refers digital camera, can be used as addition video camera, thermal imager, infrared camera, microwave complex.

3. Ground control functions: Flight tracking, Data receiving, Transfer commands.

Formulate a series of signs to determine the UAV used for the remote research of ice:

1. Type: UAV aircraft or helicopter type.

2. Control mode: automatic or semi-automatic.

3. UAV remote research of ice must have on board a complete autopilot, capable of withstanding the shooting parameters (route angles equipment, the percentage of longitudinal and transverse floors, height, etc.) even at low weight machine in a wide range of weather conditions.

4. Payload: calibrated automatic digital camera or microwave module (possibly as an adjunct video camera, a thermal imager and infrared camera), no excessive payload needed for military drones.

5. For today it should be models flying at low altitudes (in the class G airspace with a height of up to 4.5 km in unpopulated areas, within which it is planned to introduce a notification procedure for small and fly unmanned aircraft). Obtaining permission to fly in Class A and C while it is possible only by the military.

4. Commercially available - stood the test flights and entered production.

5. The model performed photogrammetric projects that have links on the site, or based on projects released article. The company's website is an indication that the main or one of the purposes is aerial photography.

III. MICROCOPTER

In 2006, the German enthusiasts Holger Buss and Ingo Busker created a microcopter (MC) around which gathered a large community of enthusiastic people - modelers, designers and programmers. Already in mid-2007, through the efforts of the founders and the entire community of this project MC has hovered steadily and quickly moved through the air. Microcopter is a radio-controlled flying platform with 4, 6, 8, 12 brushless motors with propellers. In flight platform takes horizontal position relative to the earth's surface, can hover over a certain place, move left, right, forward, backward, up and down. Nowadays, thanks to the additional equipment developed microcopter has the opportunity to actually perform semi-autonomous flight.

Along with other quite successfully implemented similar projects MC project has a very extensive support that will be needed and the novice and the experienced modeler that uses all the features of this unit.

MC flight can be considered an example hexacopter (option unit with 6 motors), sets with different number of motors are controlled in the same way. Hexacopter has propeller pairs which rotate in the opposite direction. Machine Diagram is shown in Fig.1.

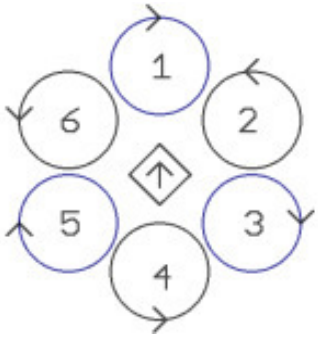


Fig. 1. Schematic rotation of hexacopter propellers.

As shown in the diagram, 1, 3, 5 hexacopter propellers rotate clockwise and 2. 4, 6 counterclockwise. To soar, all propellers must rotate at the same speed. Since the same number of propellers rotate in opposite directions at the same time, the force between them is equalized and allows to the hexacopter stably float in the air, not rotating around its axis. Photo of the machine shown in Fig.2.

Fig. 2. Hexacopter photos during the flight.

To fly in any direction, hexacopter should come out of balance, which increases the rotational speed of the propeller, opposite the direction of movement required. It makes hexacopter bend and move in this direction. For example: to fly forward, rear propeller must spin faster ("nick"). The word "nick" is used as the name for the movement of the machine forward or backward. Left or right movement called "rolls". Turn on the spot called "yaw". To implement "yaw", need to upset the balance of speed between the front / rear and left / right propellers: to turn the unit around its axis in a clockwise direction, front / rear propeller must rotate faster, and left / right propellers bit slow. Thus, carried hexacopter turn clockwise keeping the height at which the unit is located.

To ensure sustained flight controller is required which main task is to manage the flying platform stabilization in the air in a horizontal position by supplying control signals engines. It uses data from multiple sensors and calculates a speed for each propeller. The controller also compensates for external influences, such as wind. The controller works in a special program.

Key design MC elements: Baseboard (FlightCtrl), Brushless motor controllers (BL-Ctrl), Brushless motors (BLmotors), Chassis, LiPo battery, Propellers, Radioreceiver, Transmitter, Connecting wires, connectors.

Additionally microcopter can be equipped with the following elements:

1. Height sensor (barometer) - with the help of the device acquires the ability of flight at a constant altitude;

2. Magnetic compass (magnetometer) - with the help of the machine is able to hold the longitudinal axis in a constant direction;

3. GPS-receiver, which determines the coordinates of the vehicle in space and, therefore, it can provide a securing position for the given coordinates, i.e. fly on the specified points on the routes, and return to the starting point (this arrangement can also provide additional data for such telemetry as altitude, speed, direction,);

4. Sound locator (sonar), with which it is possible more precise hold altitude (it is used at an altitude of up to 6 m as an alternative barometer; moreover, it is used in the automatic takeoff and landing);

5. Optical sensor that allows precise positioning of the aircraft vertically to a height of 2.5 m

To use MC also needed: Chargers for LiPo batteries, Set of tools and accessories.

Issues requiring resolution:

1. Increased flight time - in research requires more flight time, at least 45 minutes;

2. Increase in all weather and range use microcopter;

3. Automation of some operations use multicopter;

4. Improved safety - a safe landing when connection is lost, the protection of propellers, etc.

A configuration microcopter was decided to use the UAV with six rotors (Hexacopter). Choice is due to the aerodynamic scheme balanced multi rotor aircraft. Unlike quadrocopter, failure of one or two rotors can be controlled landing. Despite lower gas-dynamic efficiency due to a smaller propeller group wheelbase engines selected design is more secure.

Depending on the task many rotary UAV can:

- lifted to a height of the payload of up to 2,000 m (the load is fixed at the central portion of the machine, around which are located coaxially or from 3 to 12 engines);
- «hang" on a user specified height with the ability to smoothly zoom in and outперемещаться во всех направлениях с полезной нагрузкой со скоростью до 50 км/ч;
- be in flight up to 30 minutes (flight time the machine depends on its configuration and payload);
- operated in a wide range of ambient temperatures and wind speed of 20-25 m / s (operation in rainy weather and snowfall during the difficult, but still possible);
- perform autonomous flight route specified coordinates on the navigation map, with stops at the desired points on the operator specified time;
- perform at stops aerial photography area, holding a given operator height and position using GPS coordinates;
- back to a place of take-off from any point on the route and from any position held by a signal operator;
- upon loss radio implement decrease with pre-programmed speed or to return to the starting point of a GPS;
- have a mobile structure (about the size of 30 × 30 × 50 cm when folded), allowing to bring the machine to a working state from traveling for 1-5 minutes;
- provide guidance onboard aerial camera television system to the subject.

Unmanned complex has on board three independent radio telemetry and positioning. One of the channels displayed on the computer, the other is connected to a standalone module indicating that lie directly in the pilot. The third channel is built into the radio equipment control. Telemetry provides continuous monitoring of all critical flight parameters and operating modes of complex systems. Management also dubbed by second post connected to a personal computer via radio. It records all flight data on the memory card, which allows for the completion of the flight to make his "analysis." Onboard video transmitter broadcasting provides high quality images to a ground monitor in real time to provide accurate guidance photos and video. The complex is equipped with a gyroscopic stabilization system with camera tracking mode. It provides image stability and vibration-free even under difficult conditions, aerial photography. Video transmitted to the monitor board operator who controls the position of the camera during the aerial survey.

IV. Features with UAV aerial data

Aerial UAV is not fundamentally different from shooting with "large aircraft", but has certain features that we continue to consider. UAV flight, usually made with a cruising speed of 70-110 km / h (20-30 m / c) at altitudes of 300-1500 m. Usually used for shooting non-metric camera household with matrix size 10-20 megapixels. Focal length of the chamber is typically about 50 mm (35 mm equivalent), which corresponds to pixel size in the terrain (GSD) of from 7 to 35 cm. Often images are processed with UAV simple lax methods (affine transformation pictures to the plane). As a result, the user receives a block layout, which, in addition to low precision circuits may contain gaps at the joints adjacent shots.

In this article, we start from a rigorous photogrammetric data processing, as a result of which we can expect the accuracy of the results (usually orthophoto mosaics) about one GSD. When the parameter values taken, the above

results correspond to the scale accuracy orthophoto from 1:500 to 1:2000 depending on the height of the shooting.

For rigorous photogrammetric aerial data and obtain the most accurate results to images in the same route had a triple overlap, and the overlap between adjacent strips in areal recording is at least 20%. In practice, when taking UAV these parameters are not always maintained. Affect the flight of the UAV gusts, turbulence, and other confounders. If shooting with conventional aircraft plan overlapped along the route 60%, and between 20-30% of the routes, the design should be shot with UAV along with overlapping routes 80%, and between routes - 40% [2].

To improve the quality of the images is recommended to install on the flight chamber lenses with fixed focal length. When shooting should exhibit focuses on infinity and disable the "AF".

Pictures of digital cameras, both amateur and professional, have a rectangular shape. "Advantageous" position the camera so that the long side of the image extends across the flight - it can take a large area with the same length of the route. Survey should be carried out with the highest quality - with the smallest jpeg compression or in RAW, if the latter is possible.

The present level of navigational aids allows for measurement of exterior orientation directly in the process of shooting. Typical accuracy of such measurements reach a few centimeters over the spatial coordinates X, Y and Z and 0.005 degrees angles roll, pitch and yaw for the most accurate systems. Often this is enough to make processing without the use of reference points. In any case, the presence of such data greatly simplifies the handling and enables some of the processing steps in fully automatic mode. Recent advances in microelectronics allow to collect the mechanical (or rather MEMS - electro-mechanical) gyroscope package size of a few mm, costing from $ 250. Such precision gyroscopes do not give a professional, have considerable care (about one degree per hour) during operation, but greatly simplify subsequent data processing. Passport accuracy GPS devices is 10 mm + 1.5 mm × B (B - removal from the base station in km) horizontal and 20 mm + 1.5 mm × B adjustment. Unfortunately, usually on board the UAV mounted cheaper GPS receivers and shall not establish IMU sensors. Data centers in the projection images are shot through telemetry data NMEA protocol and have in this case, the accuracy of 20-30 m, and the angles of pitch, roll and yaw are calculated from the velocity vector of GPS measurements. Yaw angle accuracy in such telemetry data is low and can exceed 10 degrees and the values themselves contain systematic errors, which complicates the subsequent data processing.

When shooting with a dual-band GPS receiver used in differential mode (or PPP data processing GPS), then the minimum number of control points for the most accurate results processing. Usually enough points 100 1-2 shots, in some cases the treatment can be carried out without GCP. In the case where there is no accurate projection centers, the requirements for horizontal and vertical justification of the standard, one horizontal and vertical point on 6-10 shooting bases.

V. Microwave Sensing

Work systems radiophysics diagnostics and control interference environments based on an analysis of the reaction medium under investigation by the probe signal.

Task adequate description of the interaction of electromagnetic waves with the probed environment characterized by complex permittivity (ε*), is one of the most pressing. This is due to the fact that the material of the medium of sensed are complicated dielectric structure. In reality, these media are in constant contact with a variable temperature field and the water in its various states of aggregation. These variables determine the components, mainly the dielectric properties of such media.

In solving problems of radiowave diagnostic status and properties of layered media is necessary to consider the spatial distribution of ε*. Data on their core distribution ε * can be obtained either from a priori data or using approximate theoretical models or experimentally. Analyze the value of the reflection coefficient of plane waves vertical and horizontal polarization at oblique incidence on a controlled environment. Sharing the results of sensing vertically and horizontally polarized waves can extract information on the dielectric properties of sensed layers.

Of free space (ε * = 1, μ * = 1) on the layered dielectric medium falls flat electromagnetic wave at different angles Θ (Fig.3) [3].

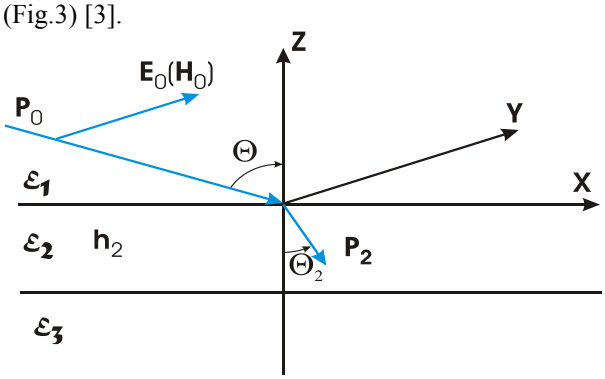


Fig.3.Geometry of the problem, the vertical and horizontal-E0-H0 polarization

Required to determine the value of the reflection coefficient K of the investigated medium, depending on the angle of incidence of the waves of horizontal and vertical polarization in the case of a finding on the medium surface of the dielectric layers. The upper and lower layers have a semi-infinite thickness, and the thickness of the relatively thin second layer - and the variable commensurate with the

wavelength. In this case, the first layer - the air, the second layer - the ice, and the third - the water. ε* values of the second and third layer are changed during the experiment..

To carry out numerical simulation environment with specialized distribution ε * can be represented as a multilayer system. In this case, ε * is a function of the coordinates of Z, and at the boundaries between the layers, this function may be discontinuous. Dependence of ε * (Z) within each layer is given by the numerical values at certain points Zi. To simplify the calculations, we set ε * between points Zi and Zi +1 is a constant and uniform in X and Y directions in layers.

Reflectance of a multilayer medium can be determined by the recurrence formula [1]:

$$K_{1,n} = \frac{K_{1,2} + K_{2,n} e^{-j\frac{4\pi h_2}{\lambda\sqrt{\varepsilon_2}}}}{1 + K_{1,2} K_{2,n} e^{-j\frac{4\pi h_2}{\lambda\sqrt{\varepsilon_2}}}} \quad (1)$$

$$K_{i,i} = 0, \quad K_{i,i+1} = \frac{\sqrt{\varepsilon_{i+1}} - \sqrt{\varepsilon_i}}{\sqrt{\varepsilon_{i+1}} + \sqrt{\varepsilon_i}} \quad (2)$$

$$K_{i,k} = \frac{K_{i,i+1} + K_{i+1,k} e^{-j\frac{4\pi h_{i+1}}{\lambda\sqrt{\varepsilon_{i+1}}}}}{1 + K_{i,i+1} K_{i+1,k} e^{-j\frac{4\pi h_{i+1}}{\lambda\sqrt{\varepsilon_{i+1}}}}} \quad k \neq i, k \neq i+1 \quad (3)$$

Using the formulas (1-3), we find the formula for the reflection coefficient K1-3 in the case research we have adopted the model:

$$K_{1,3} = \frac{K_{1,2} + K_{2,3} e^{\gamma_1}}{1 + K_{1,2} K_{2,3} e^{\gamma_1}}, \quad \text{де} \quad \gamma_1 = -j\frac{4\pi h_2}{\lambda\sqrt{\varepsilon_2}} \quad (4)$$

then for horizontal polarization:

$$K_{1,2} = \frac{\sqrt{\varepsilon_1}\cos\Theta - \sqrt{\varepsilon_2 - \varepsilon_1(\sin\Theta)^2}}{\sqrt{\varepsilon_1}\cos\Theta + \sqrt{\varepsilon_2 - \varepsilon_1(\sin\Theta)^2}} \quad (5)$$

$$K_{2,3} = \frac{\sqrt{\varepsilon_2}\cos\Theta_2 - \sqrt{\varepsilon_3 - \varepsilon_2(\sin\Theta_2)^2}}{\sqrt{\varepsilon_2}\cos\Theta_2 + \sqrt{\varepsilon_3 - \varepsilon_2(\sin\Theta_2)^2}}, \quad (6)$$

$$\text{де}\ \Theta_2 = \arcsin\left(\frac{\sin\Theta}{\sqrt{\varepsilon_2}}\right)$$

for vertical polarization:

$$K_{1,2} = \frac{\varepsilon_2\cos\Theta - \sqrt{\varepsilon_1(\varepsilon_2 - \varepsilon_1(\sin\Theta)^2)}}{\varepsilon_2\cos\Theta + \sqrt{\varepsilon_1(\varepsilon_2 - \varepsilon_1(\sin\Theta)^2)}} \quad (7)$$

$$K_{2,3} = \frac{\varepsilon_3\cos\Theta_2 - \sqrt{\varepsilon_2(\varepsilon_3 - \varepsilon_2(\sin\Theta_2)^2)}}{\varepsilon_3\cos\Theta_2 + \sqrt{\varepsilon_2(\varepsilon_3 - \varepsilon_2(\sin\Theta_2)^2)}} \quad (8)$$

By (4-8) for different states of the environment were calculated modules reflection coefficients for horizontal waves | KH | and vertical | KV | polarizations at various angles of incidence Θ on sensed environment. In this varied thickness of the thin layer h2, set different values ε * thin layer and the third layer. For clarity, the thickness of the thin layer was set in relative units and normalized with the wavelength in the medium

$$\gamma_1 = -j\frac{4\pi h_2}{\lambda\sqrt{\varepsilon_2}} = -j\frac{4\pi H}{\varepsilon_2}, \quad \text{де}\ H = \frac{h_2}{\lambda_{cp.}}, \lambda_{cp.} = \frac{\lambda}{\sqrt{\varepsilon_2}} \quad (9)$$

We simulated the situation. One of the results for the horizontal and vertical polarization is shown in Fig.4 and Fig.5.

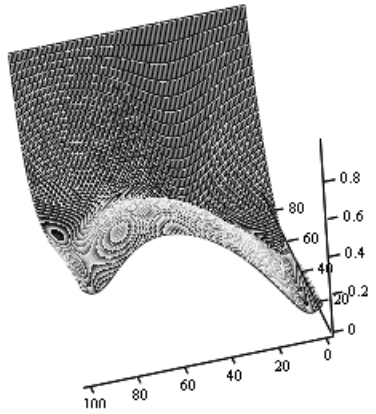


Fig. 4. Model for vertical polarization

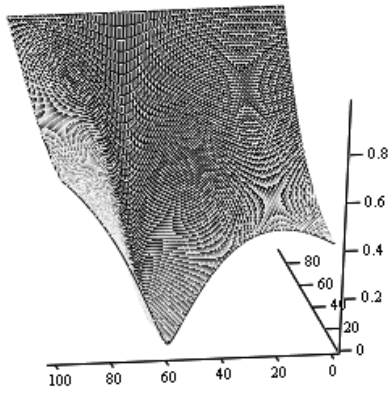


Fig. 5. Model for horizontal polarization

The X axis shows the angle Θ, on the Y-axis the thickness of the ice along the axis Z - the reflection coefficient. Chart analysis shows that horizontal polarization wave we have higher reflectance. Therefore, we will use the antenna, allowing to receive horizontal polarization. For microwave module to reduce the weight of the flight apparatus we used stroboscopic method.

VI. Experimental research

Experimental study was conducted in the delta of the Northern Dvina and dry sea near the island Mudyug.

UAV showed good performance on the stability of the en-route. Photo and video possible to determine accurately enough to crack the ice and to estimate the size of ice fields.

Microwave sensing results are shown in Table I. Analysis of the results showed that the error in determining the ice thickness is about 15%.

TABLE I.
EXPERIMENTAL RESULTS

Measurement point	Ice thickness (m)	
	Microwave sensing	Drilling
Point number 1 r. Northern Dvina 64°26'59.27"N 40° 49'2.7"E	0,21	0,38
Point number 2 r. Northern Dvina 64°28′10.23″N 40°48′42.93″E	0,58	0,42
Point number 3 r. Northern Dvina 64°27′1.93″N 40°51′5.07″E	0,43	0,4
Point number 1 Dry sea 64°52′25.32″N 40°18′10.27″E	0,61	0,48
Point number 2 Dry sea 64°55′3.55″N 40°18′39.94″E	0,57	0,47
Point number 3 Dry sea 64°54′22.63″N 40°23′8.14″E	0,6	0,49

In our data detected by sensing was carried out expeditionary ground survey of ice cover. Ground survey data confirmed the presence of breaks plots autumn ice ships that were detected on the river was probed image even in the presence of ice on the surface of the snow layer thickness of 35-40 cm.

Fig. 6. Route Map for the Northern Dvina River

This happened due to the nature of hummocky surface and metamorphosed ice structure, consisting of the frozen broken ice, slush and slush interspersed with lots of ice and fine sand particles. Due to scattering of light has broken-ice has a characteristic white color, and so it is sometimes referred to as white ice (as opposed to dark crystal clear - "black" - ice).This type of ice texture in contrast to dark ("black") reflecting intense electromagnetic waves in the centimeter range. The presence of intense white spots and stripes on the was probed image is an important deciphering sign of milled icebreakers of ice cover. Such assessment of ice covered with a layer of snow, it is impossible according aerorazvedki or visual ground survey, as well as methods of remote sensing in the optical wavelength range. The study traced the route on the Northern Dvina River (Fig.6).

VII. CONCLUSION

In the study of the possibilities of remote ice research, we concluded that the application of classical methods: spacecraft and aircraft uneconomical and has low efficiency. In addition, using these tools, we are faced with the problem of transmitting data through a narrow channel.

To solve this problem, we selected unmanned aerial vehicles (UAVs), which lack the above drawbacks, but have weight limits of the equipment used. We conducted an analysis of existing UAV and concluded that the best fit for our problem microcopter.

UAV have the following advantages:

1. Low cost.
2. Small size and weight that allows you to place equipment even on small vessels.
3. Data may be transmitted from UAV quickly in real time.
4. With a small UAV flight range can be used for broadband data transmission channel. This allows you to use equipment with high resolution.
5. Can be used in disaster areas without risk to life and limb of the operator (debacle, flooding).

Disadvantages of using UAVs:

1. Low level of accuracy of the measurements.
2. The limited size of the payload.
3. Dependence on weather conditions.

To capture an image, we propose to use a digital camera with a fixed focal length.

For remote sensing of ice, we propose to use the frequency range 1-2 GHz and microwave sensing stroboscopic method.

Experiments have shown the possibility of using UAVs for remote sensing of ice. Error in determining the thickness of the ice was not more than 15%.

REFERENCES

[1] Anpilogov VR, AA Afonin Attenuation in the satellite channels Ku-and Ka-band [electronic resource] - Access mode: http://www.tssonline.ru/articles2/practicum/zatyhanie-v-spytnikovih-kanalah-ku-i-kadiapazonov#sthash.CRQJPZ21.dpuf. Retrieved 11/21/2013

[2] Scoobyev SI, Research and Production Institute of Land and Information Technology of the State University of Land "Zeminform" (Russia), The use of unmanned aerial vehicles for the purposes of cartography. Abstracts of the X Jubilee International Scientific and Technical Conference "From imagery to map: digital photogrammetric technologies." Gaeta, Italy, 2010

[3] L.M.Brehovskih, Waves in sandwich mediums. – M.: Pub. AS USSR, 1956G.

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 1301–1310

DOI: 10.15439/2014F360
ACSIS, Vol. 2

Visual Enhancement of Service Maps in Logistics Clouds

Michael Glöckner, Björn Schwarzbach,
Andreas Barton, André Ludwig
University of Leipzig
Grimmaische Strasse 12,
04109, Leipzig, Germany
Email: {gloeckner, schwarzbach, barton,
ludwig}@wifa.uni-leipzig.de

Bogdan Franczyk
University of Leipzig,
Grimmaische Straße 12, 04109, Leipzig, Germany
Email: franczyk@wifa.uni-leipzig.de
and
Wrocław University of Economics
Komandorska 118/120, 53-345 Wrocław, Poland

***Abstract*—Logistics and its involved parties are nowadays faced with demanding challenges, in order to fulfill their customers' needs. Hence logistics service providers are constrained to cooperate with each other, which leads to the challenge of 'understanding' each other's service descriptions and integrating the strongly differing IT-systems. An emerging approach to solve this problem is the operation of cloud platforms. Main tasks are service retrieval and composition. However, a suitable visualization is needed to generate a high user acceptance. With the service map concept a first step is taken, that certainly needs further improvement. This paper briefly gives an introduction to general information visualization and analyzes the suitability of several approaches for improving the service map concept with regards to different scales of measurement. After their comparison a general guideline for fostering the visualization concept is derived. Objective is the increase of information content while keeping an intuitive usability.**

I. Introduction

TRADITIONALLY, the field of logistics is mainly dominated by logistics service providers, which are characterized as small and medium-sized enterprises (SME) [1]. Fig. 1 shows the distribution of logistics companies in the European Union (EU) categorized by the number of employees, for the year 2011.

In the last few years, but especially since the establishment of the European Union (EU), these SME are constantly facing new challenges. The abolition of traditional trade barriers and the simplification or even elimination of customs regulations led to a specialization of certain geographical regions in the field of production. Another problem for these companies is the strong wage gap between Asian countries and the EU and North American countries, which results in the manufacturing industry migrating resources to these countries. This wage gap is also existent within the EU itself, caused by constant expansion and integration of new EU member states. For instance, in 2004 ten new member States (Cyprus, the Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Slovakia and Slovenia) joined the EU. At that time these new

The work presented in this paper was co-funded by the German Federal Ministry of Education and Research under the project LSEM (BMBF 03IPT504X) and the CENTRAL EUROPE programme co-financed by the ERDF under the projects LOGICAL and ESSENCE.

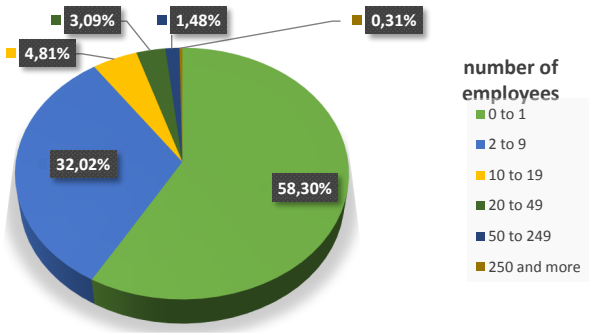


Fig. 1. Distribution of logistics companies in Europe categorized by number of employees[1]

member states had significantly lower average wages than the other incumbent member states. As a result, a non-minor number of manufacturing companies moved their locations to the south-eastern borders of the EU. This - in combination with other factors - leads to the fact that the supply chain of individual companies and company networks is stretched and thus less predictable. One of the major resulting challenges for LSP is to meet the needs of their customers furthermore [2], [3].

To cope with this challenge the division of labor and outsourcing are suitable concepts, as they provide flexibility as well as the possibility for SME to participate in larger networks. Hence, LSP have to cooperate with each other to fulfill their customers needs [4]. The most basic level of reliable networking is to publish the offered services, in order to combine them to more complex logistics services. However, by publishing the services further challenges arise. As LSP usually got a varying quality and style of service description (i.e. based on different description languages, emphasizing different aspects) as well as a wide range of implemented IT-systems (e.g. from SAP down to handmade MS Excel-sheets), an integration of information flow is hard to realize in an IT-supported way [5], [6]. In order to overcome this challenge

cooperating partners should make use of a common platform or cloud platform, respectively.

Project ESSENCE, an project within the Central Europe program - co-financed through the European Regional Development Fund (ERDF), aims to create an information and communication technology (ICT) platform. The general objectives of this platform are an increased effectiveness of transports, reducing its environmental burden, and the neutralizing of current deficiencies in SME to increase their productivity and competitiveness. ESSENCE strives to establish an ICT network that SME are encouraged to use - free of charge - to manage their logistics and optimize their supply chain by designing their own business networks. [7] Finding the right partners for fulfillment of atomic process steps is a difficult task. In the past, several scientific approaches to this problem have been developed.

To be a connecting tool between the participating LSP, the main functions of such a platform comprise retrieval and composition functions for services. Those functions are to be customized in an easy-to-understand and intuitive way in order to gain a high user acceptance. Visualization is outlined in literature as important aspect to aid users in exploring, understanding, and analyzing data through progressive, iterative visual exploration, especially in data analysis applications of business informatics related fields (e.g. visual analysis of business data, scientific data, images and videos, auction data, search results) [8]. Hence, the importance is given of visualizing service retrieval, composition and management in heterogeneous logistics clouds in a sophisticated way. One strategy is a combined approach of catalog and modular service construction system - the logistics service map (SM) [9]. Since the state of the art of the SM concept is still in draft mode, experiences with SMs still contain potential for improvement. The current structure is quite simple and the offered range of functions and displaying possibilities is way too small.

After the introductory part the basic principles of information visualization are being presented. Section three gives an overview on the current state of cloud platforms and provides basic definitions of terms such as service, process activity and process model within the context of project ESSENCE. The reader gets an impression on how the ICT platform and underlying principles are working together. In addition to that, the concept of SMs will be introduced to the reader together with an adaption to the logistics sector (logistics SM). In section four several approaches on increasing the information content of the SM by applying different combinations of scales of measurement and visual strategies are proposed and analyzed. Finally a comparison of the approaches leads to a guideline for an enhanced visualization for SMs. The paper is concluded by a short summary and future research prospects.

II. Information Visualization

As already outlined information visualization is an important aspect in business information systems. A comprehensive design study *methodology* is proposed in [10] with different phases from preliminary general preconditions (e.g. general design science knowledge acquirement, casting of all stakeholders) over core activities (e.g. problem analysis, developing alternatives, implementation and deployment) to outward-faced validations. As they assume the design process for information visualization done by specialized domain-foreign designers and the authors only focus on core activities and claim to have sufficient knowledge in the logistics and cloud domains from previous research and development projects with partners from industry and service sector, the most important and influencing parts for our purpose are the following. When developing information visualization the essential points are at first a sufficient range of possible solutions to be considered (i.e. different visual encoding and data abstraction). Afterwards those solutions are to be filtered down to a narrow proposed selection. Further [10] advise to refer on generally accepted design principles and guidelines. The major points are briefly introduced in the following paragraphs.

Commonly accepted and generally applied aspects on *design principles* for information visualization are outlined by [11]. First, he summarizes 'perception and purpose' as important aspects with appropriate visual affordance (e.g. differentiating contrast of brightness easier for human eye than contrast of colors or getting a common basis for optical comparison), identifying the right objectives, goals and tasks (e.g. setting right basis for comparison) and aesthetics (e.g. complementary colors or sectio aurea (golden ratio)). As these aspects often collide Dix further highlights 'interaction' as a solution. For this aspect drill down and hyperlinks, overview and context are highlighted as well as dynamization (i.e. changing parameters and representations, temporal fusion of different information views). Moreover, the 'visualization mantra' of [12] is mentioned with: "Overview first, zoom and filter, then details on demand". 'Information scent' [13] is a term describing the intuitiveness of handling information. In order not to click through every possibility in a matter of trial and error even if total information is not available, the user always needs a kind of clue to 'scent' on where to seek next for suitable, more detailed information. With focusing on the wider context, the perspective on different stakeholders and the resulting effects of information visualization shall be taken into account (e.g. analysts, customers, senior management). Other literature analyzes the capacity limits of attention and the subsequent influence on information visualization [14]. They conclude grouping similar objects in some cases as a powerful guideline. Moreover a proper legend or key, as well as a reduction of categories actually displayed help working with visualized information. With regard to capacity limits of attention the number of nominal categories should be dedicated to data with high dynamic range or different categories. Generally they tend to a smaller number of categories and colors to keep the user focused.

In [8] an extensive literature review on *challenges of information visualization* is conducted. As a conclusion of their paper they summarize the following five major technical challenges:

1) Usability: As the user of a system is inevitably involved in the visualization and toolkit system to accomplish its analysis, retrieval and composition tasks, the visualization shall support the user in an intuitive and efficient way.

2) Visual Scalability: Following the former mentioned visualization mantra [12], the capability of effectively displaying large data sets is an important challenge. The core part is constituted in data reduction techniques [15], as the amount of data easily exceeds the display capacities.

3) Integrated Analysis of Heterogeneous Data: The aspect of handling data from multiple sources is marked as one of the biggest challenges.

4) In-Situ Visualization: This point mainly focuses on the effective visualization of streamed data (e.g. in social networks like twitter) and how to share same processor and memory space in order to synchronize data processing and visualization task. As the concepts presented in this paper do not deal in any aspect with streamed data, this point does not hold any relevance for our purpose.

5) Errors and Uncertainty: Errors in data sets (e.g. noisy and inconsistent social media data, imprecise data from sensors) and/or resulting errors by data transformation are to be displayed to the end-user as well, in order to strengthen the truthfulness of visualization. As homogeneous data from a proven source is assumed, this point does not apply big relevance to our purpose.

The visualization pipeline is a technical infrastructure fostering information visualization is introduced by [16] with reference to [17]. However, they outline the concept, which is applied for visualizing service-oriented architectures (SOA), the concept could be abstracted and thus, also be transferred to other similar context. The four sections of the pipeline contain a typical ETL-process (i.e. the first three sections) extended by a reporting in terms of a final graphical representation. At first, heterogeneous data from different sources (e.g. UML, BPMN, XML) is extracted. Secondly, a transformation to a consistent and semantically correct form is conducted to gain the relevant meta data. Thirdly these data are loaded in a central repository as well as the pre-processed data from the repository is loaded to the next section, tailored by stakeholders' configuration (e.g. pre-defined views for different stakeholders are possible). Fourthly and finally, the graphical representation is processed from the pre-processed data from the repository taking the stakeholders' configuration into account.

Summarizing, the phases of developing alternatives and their evaluation are important and to be solved using general design guidelines. The challenges of usability and scalability are the emphasized designated requirements for the visualization approaches of the SM. As the paper focuses on visualization more in cognitive than in a technical way, only the last two sections of the visualization pipeline are considered.

III. Logistics Clouds and Service Maps - State of the Art

After giving a briefly discussing information visualization and its importance in business information systems the paper now proceeds by introducing a current approach of a cloud platform that could be used in the logistics sector. After introducing the SM concept as well, general visual disadvantages of both concepts are outlined, that are to be solved during the latter progress of the paper.

A. Logistics Cloud

At this point, the focus is on describing an existing approach - as a representative for cloud platforms in general - as detailed as possible within the restricted frame of this paper.

1) Working Definitions of ESSENCE: The ESSENCE platform [7] encourages SME in exploring the use of customized eServices that give logistics support their business. Supply chains are modeled by application of three different entities. Those entities are named: process activity, process model and service. In order to fully understand the operation principle of ESSENCE the aforementioned terms have to be defined in a short way.

Process activities are the smallest units in the ESSENCE supply chain. They are single atomic actions, described by the values 'Name', 'Description' and 'Type'. While the meaning of 'Name' and 'Description' are self-descriptive, 'Type' subdivides the activities into four distinct categories. 'Standard activities' are to be fulfilled at the side of the activity creator. It's counterpart - the 'At customer'-activities are done at the customers side. 'Latency activities' are fulfilled either at creators or customers side. Each of those latency activities has a predecessor whose completion is obligatory. The fourth category is being shaped by the so called 'custom process activities'.

Further, a *process model* results from the combination of one or more of those process activities. Similar to the aforementioned process activities these models are specified by a 'Name' and a 'Description' which are free-text fields. The values 'Duration', 'Quantity' and 'Cost (EUR)' are assigned additionally to each process activity enclosed in a process model. Furthermore, process activities within a model can be sequenced by using the field 'Sorting'.

The superior entity *service* consists of one or more process models and is specified by free-text fields containing 'Name' and a 'Description' (each twice, both in local and in common platform language 'lingua franca'). In addition to that a service could also contain 'Functional parameters', a 'Classification', 'Attachments' and 'Dependencies'.

2) ESSENCE Operation Principle: After having introduced the terms service, process model and activity in the context of the ESSENCE platform this section shows its current operation principle [7]. The workflow for designing process models is shown in a roughly simplified mainstream manner without considering any exceptions or irregularities, which is illustrated in Fig. 3.

After the initial registration process, which is approved by the platform provider, the user is able to log in. In order to publish its company's service offer on the platform the service list view is used. Each service needs a name and a short description in local language and lingua franca as well.

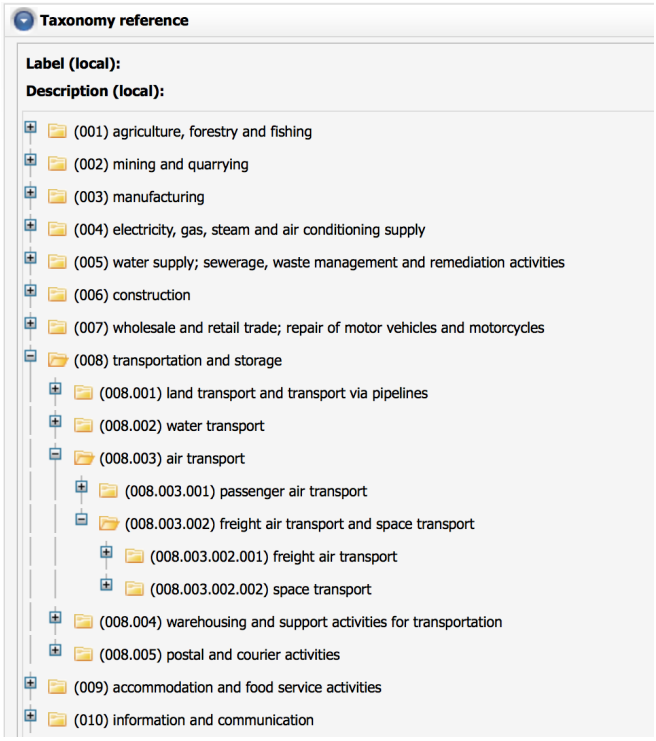


Fig. 2. Example of the taxonomy used by the ESSENCE platform.

In addition a service can have one or more parameters, e.g. number of packages of the shipment. Those parameters are used during process modeling to transfer information from one activity to another. Services can also be categorized by the ESSENCE taxonomy. This taxonomy is a tree with industrial sectors as nodes and leaves. Fig. 2 shows an example of the taxonomy tree. In addition services can have attachments and depend on other services.

To create process models the user first needs to create atomic activities in the activities list view. Therefore the user has to set a name and a description for the particular activity. After creation of this activity the user has to assign one or more supplier services to this activity. Such supplier services are alternatives, i.e. during process model execution one of the alternatives is selected.

After defining the process activities a process model is created as a orchestration of the process activities. An example of a process model consisting of three activities is shown in Fig. 4 with local language Hungarian.

During the whole process model creation the most time consuming and important step is the selection of the service suppliers and their services for the activities. To add a service of a service supplier to an activity the user first has to select the supplier and then the service of the supplier. There is no possibility to find similar services of different suppliers or filtering of services for some criteria yet. Nonetheless such a functionality would be very helpful and result in a higher usability and subsequently user acceptance. A first attempt towards such a functionality was done by implementing the taxonomy of services mentioned before. This taxonomy enables the selection of services according to their purpose. The main disadvantage of this approach is that although the taxonomy is very fine-grained it is not flexible as it offers just one dimension to categorize the offered services.

As the number of services within a network is growing the challenge of finding and selecting appropriate services is growing more than linear. Hence, an advanced approach for visualizing, filtering, revealing and combining the services of a cloud platform is needed.

B. Service Map Concept

As already outlined in the context of collaborating logistics service providers, participants are faced with the major challenge of a lack in standards for service descriptions (activity descriptions, respectively) and IT-systems in logistics [5]. Hence, in a network of logistics service providers a wide range of description types with differing wording and formats arises. Representative survey conducted within a network of small and medium logistics enterprises motivating this topic can be found in [6]. To solve this problem cloud platforms - like the introduced ESSENCE platform of the former subsection - have been developed, to support the main functions of presenting the available services (or activities, resp.) of a network (catalog function) and combining them to complex composite services (modular service construction system function), in a way that is commonly accepted by all network participants. As current approaches lack in providing a suitable visualization (e.g. the ESSENCE platform, see also Fig. 2) sophisticated approaches are needed. This issue is solved by the concept of the service map (SM). As the concept is based on another set of wording definitions the according terms of the ESSENCE platform are always given in parentheses.

In literature exists a wide variety concerning the SM concept. Either (1) the term 'service map' is used and also the functionality meets (partly) the requirements mentioned above or (2) the term is used, but a different substantial functionality is addressed or (3) the term is not used, but the described concept (partly) includes functionality for the mentioned purpose.

The approaches of (1) [18], [19], [20], [21] are situated in a wide range of scientific fields. The highest relation to the described context could be found in [18]. The SM provides an overview of existing service portfolio of a company (in the field of financial industry). Main objective is the merging and outsourcing of different companies' portfolios or business models and the related IT-systems, respectively. The catalog-related objective of service retrieval and the construction system-related objective of creating atomic and composite services are not addressed. In [19] the authors categorize mobile services and apps with the help of a scale involving dimensions of 'customer needs'. The revelation of 'empty spaces' within the maps helps identifying yet not satisfied combinations of different customer needs that could be covered by innovative new services or apps. The authors of [20] propose an XML-

Fig. 3. BPMN diagram of the ESSENCE workflow for designing process models.

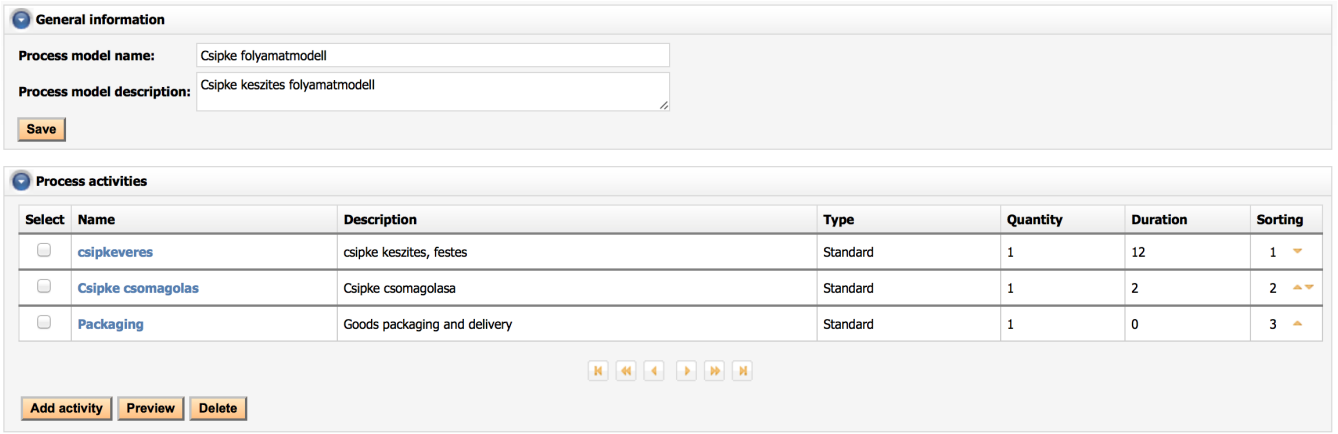


Fig. 4. Example of an ESSENCE process model with 3 activities.

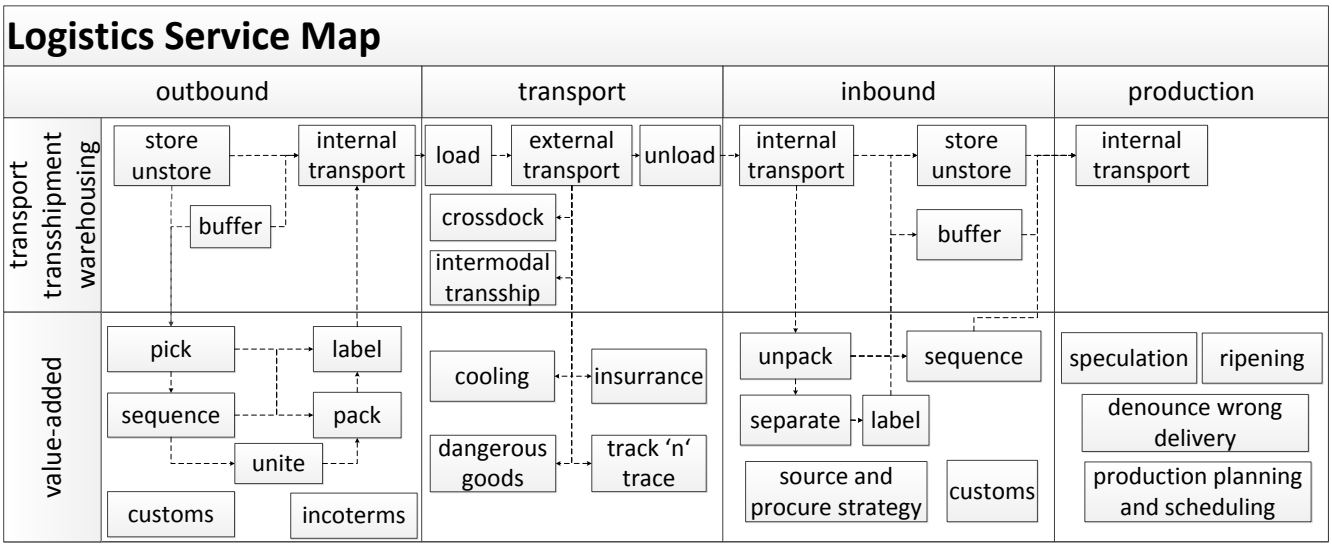


Fig. 5. Exemplary service map catalog with two dimensions: 'classic logistics function vs. value-added' and 'stage-specific'. Dashed arrows mark suggested services for composition.

based approach for structuring and categorizing services with new association and combination operators based on XML-tags. However, the approach lacks in maturity because a clear usage concept and a proper visualization is missing. [21] present a mobile data management approach. With obtaining a detailed view of available networks and their inherent capabilities, attributes and offered services in the surrounding of a mobile device. However, their categorization pattern is spatial-based and the combination functionality is missing.

The case (2) [22] addresses a SM for mapping or matching, respectively, of Quality of Service (QoS)-classes. The approach deals with data quality in heterogeneous networks consisting of several network technologies. The goal is a mapping of performance parameters of the different technologies. Neither overview of services nor building composite services is focus of the approach.

The first concept of (3), the 'service portfolio management framework' [23] combines service science with portfolio management. Hence its focus lies on strategic service management and the construction system functionality is missing. In [24] the 'business capability map' is introduced. Like [18] it focuses on comparing different companies concerning their

capabilities of serving special tasks in order of outsourcing but does not provide retrieval or construction function.

Summarizing, an approach combining both, the catalog and the construction function, is not a part of the above mentioned concepts. Subsequently, the concept of the SM introduced by [9] addresses the challenge by combining the two functions. First a catalog of all available services and process activities is provided, that could be categorized by the user's needs in different abstraction layers. As shown in Fig. 5 a graphical representation with two spatial dimensions for the user-chosen categories simplifies the interaction for the users when searching for services or process activities. Hence, service retrieval is enhanced and done in an intuitive way. Further, the concept includes a modular service construction system in order to combine atomic activities to composite services. With this approach, the network participants are supported in retrieving services in different use cases. Firstly, adding a new service provider (supplier) to the network and matching its offered services to the existing set of services by just adding the new service provider to the provider list of the particular services. Secondly, developing a new composite service (process model) to meet a specific customer's need by selecting and composing services from the SM. Service-specific information and attributes can be displayed when changing the selected granularity to a more detailed level to foster planning and monitoring. Moreover, the unique standard of the used set of services within a network and the visualization foster a precise mediation and communication between all stakeholders during the whole service life-cycle. Thirdly, finding compensational service (activity) or provider (supplier), when realizing the urgency for replanning or elimination of errors because of unpredictable disturbance in the network. Consequently, a SM is considered to be a core element of a service-oriented engineering and management platform or cloud-based service platform, respectively.

As the SM concept is an still evolving draft, certainly further development is needed to exploit its full potential. Especially, service retrieval is in the focus as this function is limited by two spatial dimensions and realized by 'highlighting categories' (1. displaying them at all, 2. reduction of displayed services by adding another category). The following section gives a more detailed introduction of the shortcomings and potential approaches for improvement.

IV. Visual Enhancement of Service Maps

The idea of a SM shown in the previous section has two points with potential for improvement. This section introduces these two ideas in detail and concludes with a brief guideline for the visual enhancement of SM concepts.

A. Preliminary Consideration

The first enhancement could be provided with the possibility of having more than only two dimensions displayed in the SM at a time. This point gains more importance with an increasing number of services available in a network that are to be displayed. Since existing approaches of SMs are only two-dimensional, the benefit of using a SM is limited in terms of category quantity. Especially with a high amount of services available in a network, the service representation lacks in clarity due to this limitation. The possibilities of integrating further categories by adding more dimensions are needed to improve the work with the SM.

Another shortcoming is that the SM categorization is fixed and cannot be changed on demand in runtime. An enhancement could be reached with the possibility of adapting the categories of the SM on the fly according to the current user needs. This would provide the user with an even more intuitive experience on the service retrieval and composition.

As summarized in the section concerning information visualization, the paper follows the methodology of [10] and focuses on developing alternatives, which are evaluated with reference to the main requirements 'usability' and 'scalability' [8]. As categories are taken into account for service retrieval and each category follows a particular type of scale, the following four scales of measurement [25] are considered. A *nominal* scale is able to display distinct named values with no natural order at all. An *ordinal* scale capable of displaying nominal values that can be brought to a natural order but have no measurable distance metrics between two values. A *interval* scale is capable of displaying quantitative numbers, which can be naturally ordered and own a distance metric between the values and a *ratio* scale as a special characteristic of a interval scale that includes a origin of ordinates.

B. Adding Further Dimension

Via brainstorming the authors and partly the ESSENCE consortium figured out the following attempts, that are compared and discussed in detail: (1) spatial dimension, (2) colored dimension and (3) shape dimension.

1) Spatial Dimension: As the current version of the SM only displays two dimensions, adding another spatial dimension is an obvious option. [16] figured out a lack in understanding and representing SOA (service-oriented architectures) (i.e. 3 layer: business-process, service interface, application) as well as the inherent intra-layer and cross-layer connections and dependencies. Reason to this is an unappropriate visualization. Outlined disadvantages of two-dimensional visualization are either overlapping connection lines or a lack in displaying different diagram types simultaneously. Hence, they summarize a loss of context with a growing number of information units and thus, propose a three-dimensional approach to visualize processes and their related services modules and classes. Drawing an analogy between the visualization of a SOA and a SM, a common evidence can be found for the requirement of an appropriate visualization with a growing number of services and inter-relations to be displayed and hence, adapt this idea to the SM.

Adding a spatial dimension to the SM leads to two possible approaches: either bringing the three-dimensional SM on a two-dimensional output device like in computer games (i.e. ordinary screen) or an three-dimensional output device (i.e. virtual reality environment, 3D display). As the latter one puts

high requirements on technical infrastructure it proves rather unsuitable for the usage in SME. As a consequence, the usage of a two-dimensional output device is proposed. Advantages are the intuitive usage and the possibility of appealing views. Whereas the main disadvantages are a difficult navigation within the SM, a high development effort and the limitation of only one further dimension.

2) Colored Dimension: Another approach of adding further dimension is visualizing categories by colors which means that the services that are shown in the SM will be colorized with an corresponding color.

Tufte states, "The fundamental uses of color in information design [are]: to label, to measure, to represent or imitate reality, to enliven or decorate." [26]. Further, he provides some basic guidelines on how to use colors for visualization of information. Small color spots on light gray background highlight the information. Such spots should be of no more than 20 to 30 colors found in nature, else negative returns would be produced by the colors. It is also important to use colors of the same hue and with maximum differentiation (e.g. yellow, red, blue, black).

"Color itself is subtle and exacting. And, furthermore, the process of translating perceived color marks on paper into quantitative data residing in the viewer's mind is beset by uncertainties and complexities. These translations are non-linear (thus gamma curves), often noisy and idiosyncratic, with plenty of differences in perception found among viewers (including several percent who are color-deficient)." [26] As already stated before [14], the usage of color as a further dimension shall be taken very carefully.

The first step for using color coding is to identify the scale of measurement of a dimension. Because of the different characteristics of the different scale of measurement it is necessary to use different colorization models for different scales of measurement. A commonly used approach for color coded scales is the *color gradient* as shown in Fig. 6. A color gradient either consists of two colors for a minimum and a maximum value or includes intermediary colors for intermediary values. To determine the exact color of a specific characteristic service within a category it is necessary to compute the characteristic's distance to the end points of the scale. The color of the service is proportional to the distance to the ends of the color gradient. Since it is necessary to compute a distance such a color gradient can only be applied to interval and ratio scales. A widely used color gradient coding is the red to green gradient where red stands for poor values and green stands for appreciated values, e.g. referring to a logistics delivery schedule whereas red stands for *all deliveries out of schedule* and green stands for *all deliveries in time*, for intermediary ratio of *50% of deliveries in time* a yellow colorization would be used as it is situated in the center of the gradient. Red-green-gradients are only suitable for assessment category dimensions referring to the maxima poor and optimal, because people always perceive red as bad and green as good due to the extensive common usage of this coding. Hence, for other categories it is obligatory to select other colors as endpoints of the scale. Categories presenting the intensity of one distinct characteristic in interval or ratio scales are likely to use a gradient from 'non-colorful' color like white to a colorful one. Again, the usage of green and red is to be considered carefully.

Another important step is the selection of the color space for the gradient. RGB and HUV are well used and well known color spaces. The main problem with these color spaces is that the perceived distance between two colors is not the same as the computed distance, hence RGB and HUV are not a good choice for the gradient. A color space addressing this problem is the CIE-LUV color space where the distance of all colors is proportional to the perceived distance [27] For dimensions of ordinal scale it is possible to use a color gradient with uniform partitioning. This results in a clear visualization of the order of the categories to the user. For nominal scales there is no method to assign special colors to categories of the dimension. If the CIE-LUV color space is applied the colors of all the categories should have a maximum distance to each other to ensure an optimal discriminability. To provide the user of the SM with a basis for interpretation of the categorization color-coding it is necessary to show a legend giving information on the meaning of the colors and the scale.

However, the main problem with color coding of categories and their particular characteristics is a lack of a common way of perception. Every person perceives colors in a different way. There are people with color vision deficiency, e.g. monochromacy, red-green or blue-yellow. Those people will possibly not be able to perceive the information that is color coded. Hence, an ideal solution provides an option to disable color coding or to use another coding instead.

3) Shape Dimensions: Another approach is to add a dimension by dedicating distinct shapes to different categories. In the beginning, the proposed forms related to geometrical shapes (e.g. circle, ellipse, rectangle, polygon, star-shaped, heart-shaped, cloud-shaped and arrow-shaped). For instance, one approach included a matching of shape and semantic of a service (e.g. transport services were presented with an arrow, while warehouse services were illustrated by pentagons in house shape). Unfortunately, preliminary tests revealed an enormous lack in usability. In other cases it was much more difficult to illustrate dimensions with appropriate shapes e.g. presenting a characteristic with a n-cardinality with the assistance of an n-sided polygon or visualizing dimensions with an ordinal scale. Further disadvantages are an unfavorable behavior in line break of labels as well as a tricky cognitive differentiation when a category with a multitude of characteristics had to be represented by a multitude of different shapes (high range of different n-shaped polygons). However, as the matching of shape and service semantics appears to be an appealing and useful approach, the decision was made to transform this idea to a more practical level. The classical rectangle-shape for services was retained and the shape dimension is integrated by reducing forms to pictograms that are to be displayed within or next to the rectangles as shown in Fig. 7. Another finding integrated in the concept is that this dimension is only useful

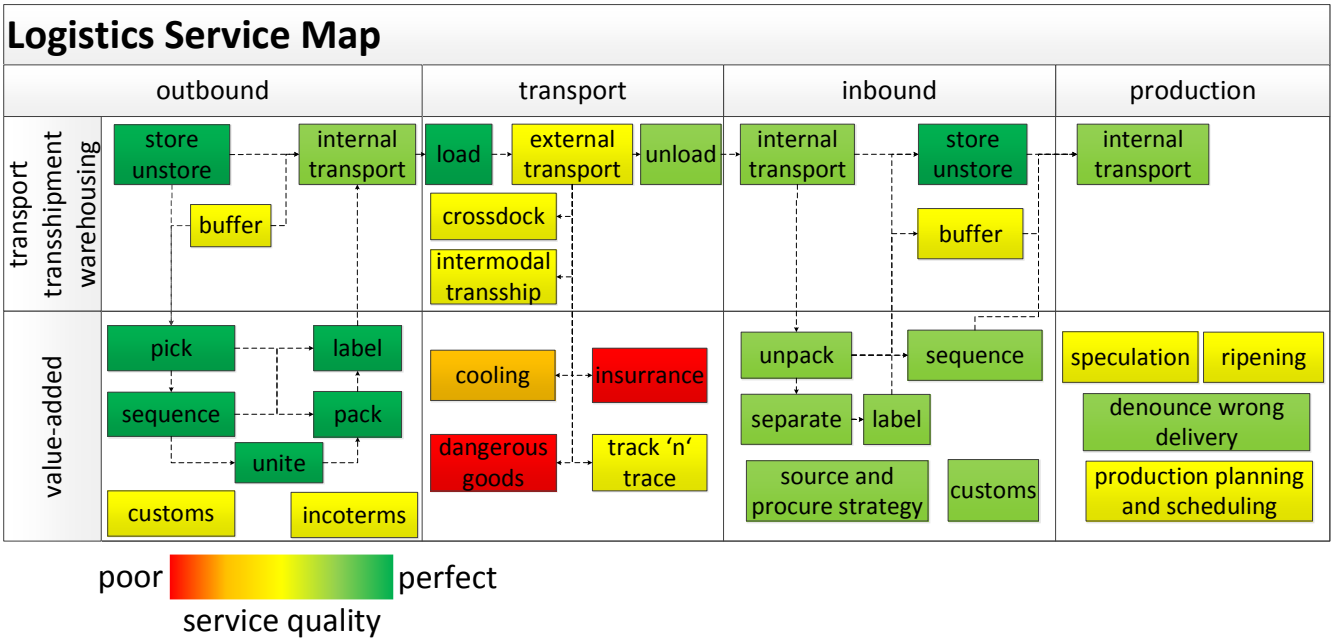


Fig. 6. Example of color gradient dimension for service map catalog.

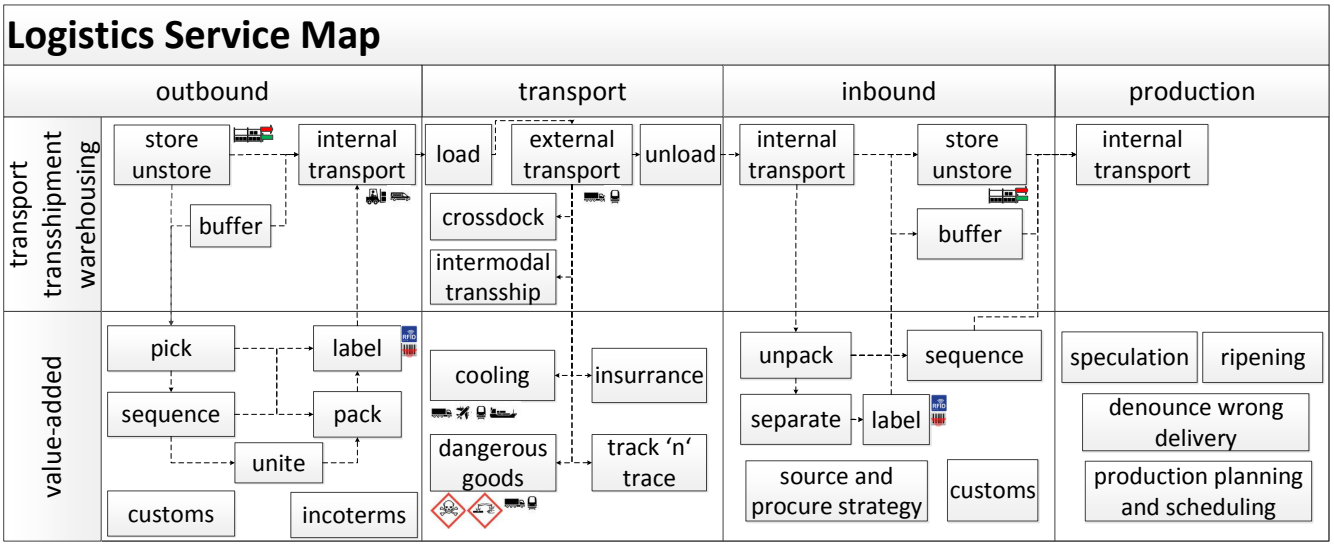


Fig. 7. Example of pictogram dimension for service map catalog.

for categories, which rely on a nominal scale.

Results of short interviews—conducted in our research group and partly the ESSENCE consortium—for *comparison* are shown in Table I. *Spatial dimensions*, i.e. 3D-SMs, were rated very poor. The only possible usage is for nominal and ordinal scales. The navigation within the 3D-space of the SM was the main problem for most of the participants in the interview. Spatial dimension do not scale well with higher number of services since they just add space in the third dimension but this space is limited for every characteristic because of clarity. Most of the participants of the interviews stated that *color dimensions* are suitable for ordinal, interval and ratio scales because they are easy to understand and can be processed pre-attentively which leads to intuitive usage. Color dimensions are also suitable for displaying nominal scaled categories but only with limited number of characteristics, as a too colorful impression only leads to user confusion. All participants agreed that *pictograms* are very suitable for nominal scales. As pictograms are not very capable in displaying order or metric, other scales than nomial are not suitable. Pictograms are appropriate for small as well as for high numbers of services. A possible problem of pictogram

TABLE I
COMPARISON OF USABILITY AND SCALABILITY OF PROPOSED APPROACHES WITH VALUES FROM 0 (POOR) TO 10 (OPTIMAL) (N—NOMINAL SCALE, O—ORDINAL SCALE, I—INTERVAL SCALE, R—RATIO SCALE)

Approach	Usability				Scalability				Advantages	Disadvantages
	N	O	I	R	N	O	I	R		
Spatial	6	7	2	2	-	-	-	-	intuitive usage, adds space for categories instead of using the exitsing space	high development effort, only one further dimension, difficult navigation
Color	5	8	9	8	6	7	8	8	easy to understand, pre-attentive processing	problems with color blind people, unclear if used with nominal scale with many categories
Pictogram	10	2	0	0	9	2	0	0	ideal for nominal scales, easy to understand, pre-attentive processing	not suitable for other scales than nominal scale, high development effort

dimensions is that there is a need for particular pictograms for every characteristic of the scale.

The paper concludes the following *guideline*: the usage of pictograms for nominal scales and color dimensions for ordinal, interval and ratio scales is proposed. If there is the need of having multiple ordinal or nominal scales then a spatial dimension could be taken into consideration. However, with two spatial dimensions and further added color dimension and pictograms there are four dimensions in total displayed in the SM. Regarding the capacity limit of attention, four dimensions are sufficient as too much categories (and the related dimensions, respectively) are counter-productive and reduce clarity and usability.

Referring to the visualization pipeline concept [16], the integration of this guideline is proposed for the third section where a stakeholder could choose options for the resulting visualization due to its requirements.

C. Dynamization

Despite the possibilities of codings for further dimensions of services, the other problem remains: if there are too many services in one category to be displayed all at once. To address this problem the visual information seeking mantra "Overview first, zoom and filter, then details-on-demand" [12] can be applied. Hence, an aggregation of the individual services (and their characteristics and visualization, resp.) is needed. The result is an overview of this category, which is a compromise between information content and possible perception. Best practice depends on the type of the chosen dimension coding. If color coding applies, one approach for n distinct characteristics in a category is to have n equally sized parts of the category with a distinct color. Another approach is to size the colored parts proportional to the ratio of the characteristics of the same color. In case of the shape coding with pictograms it is feasible to show all the pictograms of all the service once. In addition it is possible to size the pictograms or to adjust their opacity according to the ratio of characteristics for each pictogram. If the user filters services or zooms in so that the size of the displayed part is sufficient, the view changes to the small service boxes without names but with colors or small pictograms. If the user zooms further, the services will be shown normally. If the user wants to know the details of a service he can hover or click the service and the details will be shown in an overlay or a pop-up. This way it is possible to have a multitude of services shown in one category in a perceptible manner.

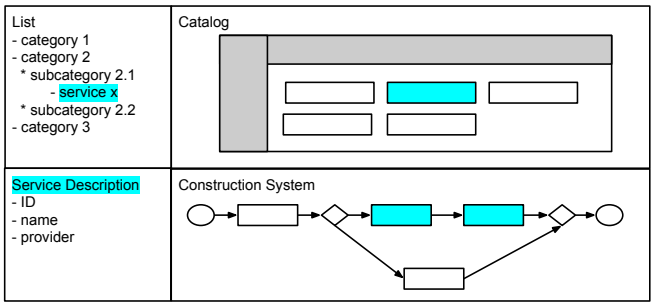


Fig. 8. Rough sketch of temporal fusion for service map concept.

Another approach for enhancing the usability of the SM concept is temporal fusion, shown in Fig. 8. By a simultaneous and identical marking of the same aspect in different views, context-specific relations can be visualized.

V. CONCLUSION AND FUTURE PROSPECTS

This paper presented an improvement of the information visualization for the service map (SM) concept by adding further dimensions. Starting with an introduction to information visualization a profound basis for the enhancement is given with an appropriate methodology, commonly accepted design guidelines, general challenges and a technical approach for implementing a visualization pipeline. Afterwards, the state of the art of the ESSENCE platform, as an representative example of a logistics cloud, was introduced as well as its current shortcomings, i.e. poor clarity in overview and difficulties in orchestration because of a poor user interface. Hence, the paper further introduced the SM concept as a combination of a catalog and a modular service construction system, in order to improve the current cloud platforms' disadvantages. Nevertheless, there were also some visual shortcomings for the SM concept.

The particular contribution of the paper is a profound analysis of information visualization aspects and a derivation of an enhanced visualization approaches for the SM concept. Further 'dimensions' for the SM are realized by the possibility of adding categories represented by colors or pictograms. Finally, the paper determines suitable approaches by giving a guideline on choosing a particular visualization approach dependent on a given scale of measurement.

Future research prospects will address challenges of technically integrating the SM concept in logistics clouds like the ESSENCE platform. Furthermore, the realization of the introduced approach of temporal fusion will be implemented and analyzed towards its usability. Further research prospects comprise sophisticated visualization approaches. Potential ideas are situated in advanced field of free spatial service positioning in a graph theory oriented manner, whereas edge lengths could be related to similarity measures between services.

REFERENCES

[1] (2014) eurostat: Your key to european statistics. [Online]. Available: http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/

[2] B. Adomavičius and Z. Lydeka, "Cooperation among the competitors in international cargo transportation sector: key factors to success," *Engineering Economics*, vol. 51, no. 1, pp. 80–90, 2007.

[3] Cruijssen, Franciscus Cornelis Andreas Maria, "Horizontal cooperation in transport and logistics," Ph.D. dissertation, Universiteit van Tilburg, Tilburg, 2006.

[4] R. Handfield, F. Straube, H.-C. Pfohl, and A. Wieland, *Trends and Strategies in Logistics and Supply Chain Management: Embracing bal logistics complexity to drive market advantage*, ser. Trends and strategies in logistics and supply chain management. Hamburg: DVV Media Group, 2013.

[5] L. Terry. (2014) 2014 third-party logistics study: The state of logistics outsourcing: Results and findings of the 18th annual study. [Online]. Available: http://www.es.capgemini.com/resource-file-access/resource/pdf/3pl_study_report_web_version.pdf

[6] U. Arnold, J. Oberländer, and B. Schwarzbach, "Logical - development of cloud computing platforms and tools for logistics hubs and communities," in *Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on Computer Science and Information Systems*, 2012, pp. 1083–1090.

[7] (2014) Essence: Easy eservices to shape and empower sme networks in central europe. [Online]. Available: http://www.essence-project.eu

[8] S. Liu, W. Cui, Y. Wu, and M. Liu, "A survey on information visualization: recent advances and challenges," *The Visual Computer*, 2014.

[9] M. Glöckner and A. Ludwig, "Towards a logistics service map: Support for logistics service engineering and management," in *Pioneering solutions in supply chain performance management: Proceedings of the Hamburg International Conference of Logistics (HICL) 2013*, ser. Reihe: Supply chain, logistics and operations management, T. Blecker, W. Kersten, and C. Ringle, Eds. Eul, 2013, vol. 17, pp. 309–324.

[10] M. Sedlmair, M. Meyer, and T. Munzner, "Design study methodology: Reflections from the trenches and the stacks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2431–2440, 2012.

[11] A. Dix, "Introduction to information visualisation," in *Information Retrieval Meets Information Visualization*, ser. Lecture Notes in Computer Science, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, M. Agosti, N. Ferro, P. Forner, H. Müller, and G. Santucci, Eds. Springer Berlin Heidelberg, 2013, vol. 7757, pp. 1–27.

[12] B. Shneiderman and C. Plaisant, *Designing the user interface: Strategies for effective human-computer interaction*, 4th ed. Boston: Pearson/Addison Wesley, 2004.

[13] P. Pirolli, *Information foraging theory: Adaptive interaction with information*, ser. Oxford series in human-technology interaction. Oxford and New York: Oxford University Press, 2007.

[14] S. Haroz and D. Whitney, "How capacity limits of attention influence information visualization effectiveness," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2402–2410, 2012.

[15] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler, "Challenges in visual data analysis," in *Tenth International Conference on Information Visualisation (IV'06)*, 2006, pp. 9–16.

[16] S. Eicker, T. Spies, and C. Kahl, "Software visualization in the context of service-oriented architectures," in *2007 4th IEEE International Workshop on Visualizing Software for Understanding and Analysis*, 2007, pp. 108–111.

[17] H. Schumann and W. Müller, *Visualisierung*. Berlin and Heidelberg: Springer Berlin Heidelberg, 2000.

[18] F. Kohlmann and R. Alt, "Aligning service maps - a methodological approach from the financial industry," in *Proceedings of the 42nd Annual Hawaii International Conference on System Sciences*, R. H. Sprague, Ed. IEEE Computer Society Press, 2009, pp. 1–10.

[19] J. Kim, S. Lee, and Y. Park, "User-centric service map for identifying new service opportunities from potential needs: A case of app store applications," *Creativity and Innovation Management*, vol. 22, no. 3, pp. 241–264, 2013.

[20] S. Vaddi, H. Mohanty, and R. Shyamasundar, "Service maps in xml," in *Proceedings of the CUBE International Information Technology Conference*, V. Potdar, Ed. ACM, 2012, pp. 635–640.

[21] D. Kutscher and J. Ott, "Service maps for heterogeneous network environments," in *MDM 2006 Japan*. IEEE Computer Society, 2006.

[22] Mi Sun Ryu, Hong-Shik Park, and Sang-Chul Shin, "Qos class mapping over heterogeneous networks using application service map," in *Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies, 2006. ICN*, 2006, p. 13.

[23] T. Kohlborn, E. Fielt, A. Korthaus, and M. Rosemann, "Towards a service portfolio management framework," in *ACIS2009 - Australian Conference on Information Systems*, 2009, pp. 861–870.

[24] J. Fleischer, M. Herm, U. Homann, K. Peter, and K.-H. Sternemann, "Business capabilities als basis fähigkeitsorientierter konfigurationen," *ZWF - Zeitschrift für wirtschaftlichen Fabrikbetrieb*, vol. 100, no. 10, pp. 553–557, 2005.

[25] S. S. Stevens, "On the theory of scales of measurement," 1946.

[26] E. R. Tufte, *Envisioning information*, 5th ed. Cheshire and Conn: Graphics Press, 1995, c1990.

[27] M. Tkalcic, J. F. Tasic *et al.*, "Colour spaces: perceptual, historical and applicational background," in *Eurocon*, 2003.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1311–1319
DOI: 10.15439/2014F90
ACSIS, Vol. 2

Modeling enablers for sustainable logistics collaboration integrating – Canadian and Polish perspectives

Katarzyna Grzybowska
Poznan University of Technology,
Strzelecka 11, 60-965 Poznan, Poland
Email:
katarzyna.grzybowska@put.poznan.pl

Anjali Awasthi
Concordia University,
1455 De Maisonneuve Blvd. West
Montreal H3G 1M8, Canada
Email:
awasthi@ciise.concordia.ca

Mohammad Hussain
Concordia University,
1455 De Maisonneuve Blvd.
West
Montreal H3G 1M8, Canada
Email:
whozane@gmail.com

***Abstract*—Collaboration planning is vital for achieving sustainable logistics. In this paper, we present an ISM based approach for modeling enablers for sustainable logistics collaboration integrating Canadian and Polish perspectives. Enablers can be defined as the key elements (or drivers) for achieving successful collaboration. A comprehensive literature review is conducted to identify 17 enablers for sustainable logistics collaboration. Based on these enablers, a questionnaire study is conducted with 20 logistics experts in Canada and Poland to identify their importance. Based on the aggregated expert ratings, an Interpretive Structural Model (ISM) is developed to identify the relationships among the various enablers. The results of our study show that not all enablers to sustainability collaboration between logistics partners require the same amount of attention. This classification will help Supply Chain managers to help them to focus on those variables that are most important for the transformation of collaboration between logistics partners.**

I. Introduction

COLLABORATION is vital to achieving success in sustainable logistics operations. Modern logistics operators are under increased pressure and administrative regulations in order to fulfill environmental objectives, reduce congestion and make parking space available for public space. For example, vehicle timing, access and sizing regulations are limiting the areas, timing and size of delivery vehicles. Likewise, tax rebate policies may encourage the use of clean energy vehicles or energy efficient goods distribution practices. Under these conditions, collaboration seems a logical and viable strategy for many logistics operators to achieve operational performance as well as successfully meet environmental targets. Several researchers emphasize the importance of collaboration in value creation in supply chains [1]; [2]; [3]. Langley [4] advocates that with the complexity and dynamic nature of today's rapidly evolving business world, any firm stands to lose if trying to "go it alone." [5]. Barratt [6] addresses supply chain collaboration through customer buying behavior and service needs, identifying elements that make up supply chain as well as investigate the interrelationships among the cultural, strategic and implementation elements of supply chain. Cassivi, Garner Group, Ovalle and Márquez (2003) present e-collaboration tools for supply chains [7], [8], [9]. Holweg et al. [10] classify collaboration initiatives using conceptual water-tank approach; and discuss their dynamic behaviors and key characteristics. Kale et al. [11] investigate internet-based collaborative transportation networks. Muckstadt et al. propose guidelines for collaborative supply chain system design and operation [12]. Zhou et al. investigate strategic alliance in freight consolidation [13].

Supply chain management is defined as "the systemic, strategic coordination of the traditional business functions and the tactics across these business functions within a particular company and across businesses within the supply chain, for the purposes of improving the long term performance of the individual companies and the supply chain as a whole" [14]. Supply chain structure defines the way various organizations within the supply chain are arranged and related to each other. The supply chain structure falls into four main types: Convergent: each node in the chain has at least one successor and several predecessors. Divergent: each node has at least one predecessor and several successors. Conjoined: which is a combination of each convergent chain and one divergent chain. Network: which cannot be classified as convergent, divergent or conjoined, and is more complex than the three previous types [15, 16, 17].

In this paper, our research objectives are:

- to identify and rank the enablers for collaboration between logistics partners in the context of Poland and Canada
- to find out the relation and interaction among identified enablers using ISM
- to suggest for future research.

The paper is organized as follows. In section 2, we will present a list of enablers aimed at the identification of cooperation between business partners. A brief characteristic of questionnaire, the ISM methodology along with its application will be covered in sections 3-5. The MICMAC analysis will be presented in section 6. Finally, the article will

This work is supported by Poznan University of Technology and Concordia University

end with a recapitulation which includes the most important conclusions and directions for future works in sections 7-8.

II. Identification of enablers for collaboration between logistics partners

An enabler is defined as "as one that enables another to achieve an end" where enable implies to make able; give power, means, competence, or ability to (Merriam-Webster). An enabler is considered as a variable that enables (ability to) the attainment of the Sustainable Supply Chain. This definition is consistent with the use of the term enabler in ISM models [18], growth enablers in construction companies [18], information technology (IT) enablement in the Supply Chain [20], enablers of reverse logistics [21], supply chain performance measurement system implementation [22], modelling the barriers of global supply chain [23], supply chain sustainability – analysing the enablers [24].

On the basis of the literature as well as the experience and knowledge of experts, enablers were proposed for collaboration between logistics partners. 17 important variables (enablers), have been differentiated, which, in the opinion of experts, are of significance in business cooperation. Enablers, discussed and selected for analysis, are presented below.

A. Information sharing

Information sharing among logistics partners can take place about anything that leads to amelioration of operational efficiency and attainment of environmental objectives. For example, customer demands, vehicle resources, warehousing capacity, goods inventory or technological know-how. Information sharing leads to visibility in supply chains which in turn leads to cooperation among supply chain partners.

B. Coordination

Coordination can be defined as alignment of project objectives and resources in order to achieve the successful collaboration. Effective coordination among all enterprises cooperating with one another in supply chain is essential to its success [25]. Lee and Whang suggested that information is a basic enabler for tight coordination [26]. Coordination allows for the efforts and the aims of the individual enterprises to be unified. The coordinating actions are fundamental, those that (1) stimulate the supply chain through the creation of a supply chain growth concept, (2) regulate the supply chain by redistributing the possessed resources (3) integrate the supply chain by linking resources, monitoring and an assessment of the actions [27].

C. Trust

In a supply chain, trust is one of the key cooperation factors (e.g. trust that a supplier or a subcontractor perform their duties according to specifications; trust that a supplier with which the enterprise did cooperate previously, will supply a product of a proper quality; trust that the customer will pay within agreed period of time and will not cause a payment gridlock, etc) [28]. It is during cooperation when complex trust-based reactions occur, since one entity's gains depend on the other. The risk and uncertainty connected with trust and cooperation develop as the number of participants increases.

D. Willingness to collaborate

Willingness to collaborate is vital in achieving successful collaboration. Disinterested partners lack commitment and may leave the collaboration anytime leading to waste of resources, time, money and personnel.

E. Communication

According to Hahn et al. effective communication among all elements of supply chain. Clearly communicated goals across members of all hierarchy in the organization leads to efficient realization of planned objectives under given time, pressure and resources [25].

F. Common business goals

Common business goals are one of the main reasons behind any organization's interest to collaborate with other partner organizations. Similar business goals lead to common practices, techniques, and efficient sharing of knowledge leading to win-win situation for all the organizations participating in collaboration.

G. Responsibility sharing

Responsibility allocation and sharing among the participating organizations leads to increased trust, information sharing, and commitment from the involved partners and fosters strong collaboration.

H. Planning of supply chain activities

Efficient and timely planning of supply chain activities reduces waste of time, resources and money arising from last minute changes in customer demands or unstable market conditions. For successful collaboration, right planning also leads to efficient realization of goals, sharing of resources, and profit allocation.

I. Flexibility

Organizations participating in any collaboration do not always have the same experience, culture, technological readiness, or brand image. Under these situations, it is vital for participating organizations to be flexible to adapt to others needs for joint success. Richey et al. identify technology and flexibility as key enablers for logistics collaboration [29].

J. Benefit sharing

Benefit sharing among participants is vital for retaining loyalty towards collaboration. Benefit sharing can be done

equally among partners or depending upon the stakes of major contributing organizations after mutual consensus.

K. JOINT DECISION MAKING

All the organizations involved in collaboration should work together and perform joint decision making to achieve operational and environmental goals. This will lead to increased trust and commitment which is essential for long term success of any collaboration.

L. ORGANIZATIONAL CULTURE

Organizational culture is very important in sharing of common vision and goals for any collaboration. Participating organizations should have a good understanding of other participant organizations culture to avoid any misconceptions and gaps that can serve as barriers in realization of project objectives.

M. ORGANIZATIONAL COMPATIBILITY

Organizational compatibility in terms of product-service types, size, location, strategy, employee culture, technology, management commitment, budget, resources etc. aids in developing successful collaborations among the participating organizations.

N. RESOURCE SHARING (INTEGRATION)

Once organizations enter into collaboration, it is important to address the goals of all participants efficiently, uniformly and in time either through partial and/or complete sharing of resources. This will also help in achieving successful project co-ordination and completion.

O. TOP MANAGEMENT SUPPORT

Commitment from management includes an effort and financial backing from the upper management to implement sustainability in logistics operations. Top management commitment retains employee interests in implementing sustainability practices and continuous improvement goals [30]. In order to achieve long term success, management support and commitment is very important and should be accompanied with employee rewards and training programs.

P. TECHNOLOGICAL READINESS

Use of IT tools to monitor the supply chains and sharing information among the partners leads to visibility in supply chain, thereby providing better cooperation among different levels of the supply chain [30]. Electronic data interchange and internet have enabled partners in supply chains to act upon same data rather than rely on distorted and noisy data that emerges in an extended supply chain [26, 31]. Swafford et al emphasize the role of IT integration and flexibility in achieving supply chain agility [32].

Q. TRAINING

Training helps employees with expertise to perform their tasks efficiently. Company's power comes from the physical and mental strength of their workers. Organizing employee trainings and maintaining occupational safety and health are among the main functions of human resources management departments. These two functions interact and they both serve the aim of protecting employee's physical, psychological and social health [30].

III. QUESTIONNAIRE-BASED SURVEY

In order to determine the relative importance of enablers, we conducted a questionnaire study. The main objective of the questionnaire-based survey was to facilitate experts in developing a relationship matrix as a first step towards developing an ISM-based model. In this survey the respondents were asked to indicate the importance of 17 listed enablers on a five-point Likert scale [18]. On this scale, 1 and 5 correspond to 'very low importance' to 'very high importance', respectively. In total, questionnaires were sent to 20 experts in Poland and Canada. All of them were analyzed.

TABLE I.
ENABLERS ACCORDING TO RESEARCHERS FROM POLAND

lp.	Enablers	Mean score
1	Communication	4,6
2	Information sharing	4,6
3	Coordination	4,4
4	Willingness to collaborate	4,3
5	Planning of supply chain activities	4,1
6	Trust	4,0
7	Responsibility sharing	3,9
8	Common business goals	3,8
9	Benefit sharing	3,6
10	Flexibility	3,4
11	Joint Decision Making	3,4
12	Resource sharing (integration)	3,4
13	Organisational compatibility	3,3
14	Organizational culture	3,3
15	Top management support	3,1
16	Technological readiness	2,5
17	Training	2,5

Tables 1-2 show the difference between perceiving the relevance of the enablers in relation to the collaboration between business partners in Poland and Canada. According to researchers from Poland (Tab. 1), the most significant enabler is communication (4,6). The lack of information often leads to misunderstandings, and these, in turn, cause uncertainty, ambiguity, difficulty and failure in communication, and in effect failure in collaboration between business organizations.

TABLE III.
ENABLERS ACCORDING TO RESEARCHERS FROM CANADA

lp.	Enablers	Mean score
1	Trust	4,4
2	Coordination	4,3
3	Information sharing	4,3
4	Common business goals	4,3
5	Flexibility	4,1
6	Willingness to collaborate	4,1
7	Organizational culture	4,0
8	Organisational compatibility	3,9
9	Joint Decision Making	3,9
10	Responsibility sharing	3,9
11	Technological readiness	3,9
12	Communication	3,7
13	Benefit sharing	3,7
14	Resource sharing (integration)	3,7
15	Training	3,7
16	Planning of supply chain activities	3,6
17	Top management support	3,4

TABLE IIIII.
ENABLERS APPLIED IN THE RESEARCH

lp.	Enablers	Mean score
1	Information sharing	4,45
2	Coordination	4,35
3	Trust	4,20
4	Willingness to collaborate	4,20
5	Communication	4,15
6	Common business goals	4,05
7	Responsibility sharing	3,90
8	Planning of supply chain activities	3,85
9	Flexibility	3,75
10	Benefit sharing	3,65
11	Joint Decision Making	3,65
12	Organizational culture	3,65
13	Organisational compatibility	3,60
14	Resource sharing (integration)	3,55
15	Top management support	3,25
16	Technological readiness	3,20
17	Training	3,10

While researchers from Canada (Tab. 2) deemed trust (4,4) as key in proper collaboration. Practically each commercial transaction contains an element of trust. Trust has its practical, real, economic value, since it results in the increase in the effectiveness of the system and allows for more goods to be produced.

The explanation of the occurring differences is cultural differences - different behaviors, social norms or a different system of values. Despite the fact that social, political, economic changes as well as technological achievements have made inter-cultural contacts become an everyday matter, cultural differences still take place and play a significant role in the 21st century.

Table 3 includes the joint list of enablers, which was accepted in the research and analysis.

IV. METHODOLOGY

Interpretive Structural Modeling (ISM) is defined as a process aimed at assisting the human being to better understand and clearly recognize what one does not know [33]. The ISM process transforms unclear, poorly articulated mental models of systems into visible and well defined models. ISM is an interactive learning process. In this technique, a set of different directly and indirectly related elements are structured into a comprehensive systematic model. The model so formed portrays the structure of a complex issue or problem in a carefully designed pattern implying graphics as well as words [18, 34].

Interpretive Structural Modeling was first proposed by Warfield [35]. It enables individuals or groups to develop a map of the complex relationships between many elements involved in a complex decision situation [22].

A. THE IMPORTANT CHARACTERISTICS OF ISM

The important characteristics of ISM are:

• This methodology is interpretive, as the judgment of the group decides whether and how the different elements are related.

• It is structural on the basis of mutual relationships as the overall structure is extracted from the complex set of elements.

• It is a modelling technique, as the specific relationships and overall structure are portrayed in a digraph model [34].

B. THE VARIOUS STEPS INVOLVED IN THE ISM TECHNIQUE

The steps involved in the ISM are represented in the form of a flow diagram (see Fig. 1).

The various steps involved in the ISM technique are as follows [18]:

Step 1: Different enablers (or variables), which are related to defined problems, are identified.

Step 2: A Structural Self-Interaction Matrix (SSIM) is developed for enablers. This matrix indicates the pair-wise relationship among enablers of the system. This matrix is checked for transitivity.

Step 3: A Reachability Matrix (RM) is developed from the SSIM.

Step 4: The RM is partitioned into different levels.

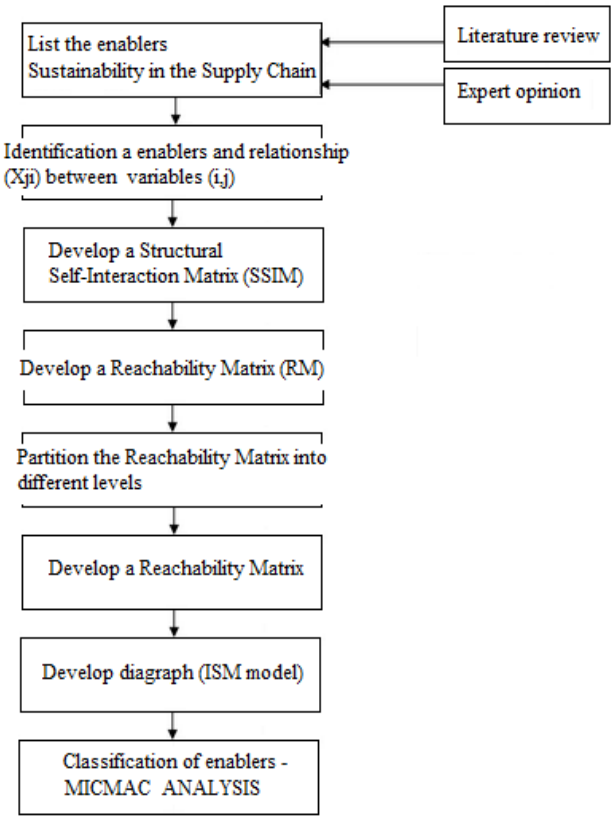


Fig. 1 Flow diagram for constructing an Interpretive Structural Modeling; based on [35]

Step 5: The Reachability Matrix is converted into its conical form, i.e. with most zero (0) elements in the upper diagonal half of the matrix and most unitary (1) elements in the lower half.

Step 6: Based upon the above, a directed graph (digraph) is drawn and transitivity links are removed.

Step 7: Digraph is converted into an ISM model by replacing nodes of the elements with statements.

Step 8: The ISM model is checked for conceptual inconsistency and necessary modifications are incorporated.

C. The Formation of Structural Self-Interaction Matrix (SSIM)

After identifying the 17 enablers through the review of literature and expert opinions, the next step is to analyze these enablers. For this purpose, a contextual relationship of "leads to" type is chosen. Bearing the contextual relationship for each enabler in mind, the existence of a relation between any two enablers (i and j) and the associated direction of this relation has been decided [18].

From the enablers identified in step 1, a contextual relationship is identified among enablers with respect to which pairs of variables would be examined. This step transforms the list into a matrix and marks dependencies using expert opinions. After resolving the enablers set under consideration and the contextual relation, a Structural Self-Interaction Matrix (SSIM) is prepared.

Four symbols are used to denote the direction of relationships between the enablers (i and j):

- V: for the relationship from enabler i to enabler j and not in both directions;
- A: for the relationship from enabler j to enabler i and not in both directions;
- X: for both the directional relationships from enabler i to enabler j and j to i;
- O: if the relationships between the enablers did not appear valid (enablers i and j are unrelated) [18].

TABLE IVV.
STRUCTURAL SELF-INTERACTIVE MATRIX (SSIM)

Ip.	Enablers	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
1	Information sharing	X	A	V	X	X	V	A	O	O	V	X	A	X	V	X	X
2	Coordination	A	A	V	X	X	V	O	O	X	X	A	O	X	V	O	
3	Trust	A	O	V	X	X	V	X	A	O	O	O	O	O	V		
4	Willingness to collaborate	A	O	V	V	X	V	X	X	O	X	A	A	A			
5	Communication	X	X	V	X	X	V	A	O	V	X	A	O				
6	Common business goals	A	O	V	V	X	V	A	A	O	O	X					
7	Responsibility sharing	A	V	X	X	X	V	X	X	V	V						
8	Planning of supply chain activities	A	A	V	V	X	V	V	O	X							
9	Flexibility	A	A	V	V	X	V	O	O								
10	Benefit sharing	O	O	V	V	X	V	X									
11	Joint Decision Making	A	O	V	X	X	V										
12	Organizational culture	A	O	O	A	X											
13	Organisational compatibility	O	X	V	X												
14	Resource sharing (integration)	A	X	V													
15	Top management support	A	V														
16	Technological readiness	O															
17	Training																

Based on the review of literature and expert's responses, the SSIM is constructed as shown in Tab. 4. The following statements explain the use of symbols in Structural Self-Interaction Matrix, e.g. [18]:

- Symbol V is assigned to cell (1, 15) as enabler 1 influences or reaches enabler 15.
- Symbol A is assigned to cell (2, 17) as enabler 17 influences the enabler 2.
- Symbol X is assigned to cell (5, 16) as enablers 5 and 16 influence each other.
- Symbol O is assigned to cell (6, 16) as enablers 6 and 16 are unrelated.

The next step is to develop the Reachability Matrix (RM) from the Structural Self-Interactive Matrix. This is obtained in two sub-steps.

In the first sub-step, the Structural Self-Interaction Matrix is trans-formed into a binary matrix (see Tab. 5), called the initial reachability matrix by substituting V, A, X, O by 1 and 0 as per the case. The rules for the substitution of 1s and 0s are as follows:

- If the (i, j) entry in the SSIM is V, then the (i, j) entry in the reachability matrix becomes 1 and the (j, i) entry becomes 0.

• If the (i, j) entry in the SSIM is A, then the (i, j) entry in the reachability matrix becomes 0 and the (j, i) entry becomes 1.

• If the (i, j) entry in the SSIM is X, then the (i, j) entry in the reachability matrix becomes 1 and the (j, i) entry also becomes 1.

• If the (i, j) entry in the SSIM is O, then the (i, j) entry in the reachability matrix becomes 0 and the (j, i) entry also becomes 0 [18].

TABLE V.
INITIAL REACHABILITY MATRIX

Ip.	Enablers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	Information sharing	1	1	1	1	1	0	1	1	0	0	0	1	1	1	1	0	1
2	Coordination	1	1	0	1	1	0	0	1	1	0	0	1	1	1	1	0	0
3	Trust	0	0	1	1	0	0	0	0	0	0	1	1	1	1	1	0	0
4	Willingness to collaborate	0	0	0	1	0	0	0	1	0	1	1	1	1	1	1	0	0
5	Communication	1	1	0	1	1	0	0	1	1	0	0	1	1	1	1	1	1
6	Common business goals	1	0	0	1	0	1	1	0	0	0	0	1	1	1	1	0	0
7	Responsibility sharing	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
8	Planning of supply chain activities	0	1	0	1	1	0	0	1	1	0	1	1	1	1	1	0	0
9	Flexibility	0	1	0	0	0	0	0	1	1	0	0	1	1	1	1	0	0
10	Benefit sharing	0	0	1	1	0	1	1	0	0	1	1	1	1	1	1	0	0
11	Joint Decision Making	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	0	0
12	Organizational culture	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
13	Organisational compatibility	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
14	Resource sharing (integration)	1	1	1	0	1	0	1	0	0	0	1	1	1	1	1	1	0
15	Top management support	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0
16	Technological readiness	1	1	0	0	1	0	0	1	1	0	0	0	1	1	0	1	0
17	Training	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1

In the second sub-step, the final reachability matrix is prepared (see Tab. 6). The concept of transitivity is introduced so that some of the cells of the initial reachability matrix are filled by inference. The transitivity concept is used to fill the gap, if any, in the opinions collected during the development of the SSIM [18].

TABLE VI.
FINAL REACHABILITY MATRIX

Ip.	Enablers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	Information sharing	1	1	1	1	1	0	1	1	0	0	0	1	1	1	1	0	1
2	Coordination	1	1	0	1	1	0	0	1	1	0	0	1	1	1	1	0	1*
3	Trust	0	0	1	1	0	0	0	0	0	1*	1	1	1	1	1	1*	0
4	Willingness to collaborate	0	0	1*	1	0	0	0	1	0	1	1	1	1	1	1	0	0
5	Communication	1	1	0	1	1	0	0	1	1	0	0	1	1	1	1	1	1
6	Common business goals	1	0	0	1	0	1	1	0	0	0	1*	1	1	1	1	1*	0
7	Responsibility sharing	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0
8	Planning of supply chain activities	0	1	0	1	1	0	0	1	1	0	1	1	1	1	1	0	0
9	Flexibility	0	1	0	0	0	0	0	1	1	0	0	1	1	1	1	0	0
10	Benefit sharing	0	0	1	1	0	1	1	0	0	1	1	1	1	1	1	0	1*
11	Joint Decision Making	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	0	0
12	Organizational culture	0	0	1*	0	0	0	0	0	0	0	0	1	1	0	0	0	0
13	Organisational compatibility	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
14	Resource sharing (integration)	1	1	1	0	1	0	1	0	0	0	1	1	1	1	1	1	0
15	Top management support	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0
16	Technological readiness	1	1	1*	0	1	1*	0	1	1	0	0	0	1	1	1*	1	0
17	Training	1	1	1	1	1	1	1	1	1	1*	1	1	0	1	1	0	1

TABLE VII.
ITERATION 1

No.	Reachability set	Antecedent set	Intersection set	Level
1	1,2,3,4,5,7,8,12,13,14,15,17	1,2,5,6,7,11,13,14,16,17		
2	1,2,4,5,8,9,12,13,14,15,17	1,2,5,7,8,13,14,16,17		
3	3,4,11,12,13,14,15	1,3,10,11,13,14,17		
4	4,8,10,11,12,13,14,15	1,2,3,4,5,6,7,8,10,11,13,17		
5	1,2,4,5,8,9,12,13,14,15,16,17	1,2,5,7,8,10,13,14,16,17		
6	1,4,6,7,12,13,14,15	6,7,10,11,13,17		
7	1,2,4,5,6,7,8,9,10,11,12,13,14,15, 16	1,6,7,10,11,13,14,15,17		
8	2,4,5,8,9,11,12,13,14,15	1,2,4,5,7,8,9,13,16		
9	2,8,9,12,13,14,15	2,5,7,8,9,13,16,17		
10	3,4,6,7,10,11,12,13,14,15,17	3,4,7,10,11,14,16,17		
11	1,3,4,5,6,7,10,11,12,13,14,15	3,4,6,7,8,10,11,13,14,17		
12	12,13	1,2,3,4,5,6,7,8,9,10,11,12,13,14,17	12,13	I
13	1,2,3,4,5,6,7,8,9,10,11,12,13,14,1 5,16	1,2,3,4,5,6,7,8,9,10,11,12,13,14,16		
14	1,2,3,5,7,11,12,13,14,15,16	1,2,3,4,5,6,7,8,9,10,11,13,14,16,17		
15	7,15,16	1,2,3,4,5,6,7,8,9,10,11,13,14,15,16, 17		
16	1,2,3,5,6,8,9,13,14,16	3,5,6,7,13,14,15,16		
17	1,2,3,4,5,6,7,8,9,11,12,14,15,17	1,5,17		

In the present case, the 17 enablers are presented in Tables 7–8. The level identification process of these enablers is completed in four iterations as shown in Tables 7–8.

TABLE VIII.
ITERATIONS 2-5

No.	Reachability set	Antecedent set	Intersection set	Level
15	7,15,16	1,2,3,4,5,6,7,8,9,10,11,14,15,1 6,17	7,15,16	II
3	3,4,11,14	1,3,4,10,11,14,17	3,4,11,14	III
4	8,10	1,2,5,6,8,10,17	8,10	IV
8	2,5,8,9	1,2,5,8,9	2,5,8,9	IV
9	2,8,9	2,5,8,9,17	2,8,9,	IV
14	1,2,5	1,2,5,6,8,9,1017	1,2,5	IV
1	17	6,17	17	V
2	17	17	17	V
5	17	17	17	V
6	6	6,17	6	V
7	6	6,17	6	V
10	17	17	17	V
11	6	6,17	6	V
13	6	6	6	V
16	6	6	6	V
17	6,17	17	17	V

V. BUILDING THE ISM MODEL

The diagraph for interpretive structural modelling is drawn. Having identified the levels of the elements, the relations between the elements is drawn with the help of an arrow. The level I enablers are in the top level in the hierarchy. The enablers of the same level are kept on the same level of hierarchy [36]. The diagraphs give information about the hierarchy between the elements of enablers for the collaboration between logistics partners (see Fig. 2). It can be seen in Fig. 2 that the most important enablers (level I) are 'Common business goals' and 'Training'.

The purpose of Cross-Impact Matrix Multiplication Applied to the Classification analysis (MICMAC) is to analyse the drive power and dependence power of enablers. This is done to identify the key enablers that drive the system in various categories [18]. The variables are classified into four clusters (see Fig. 3).

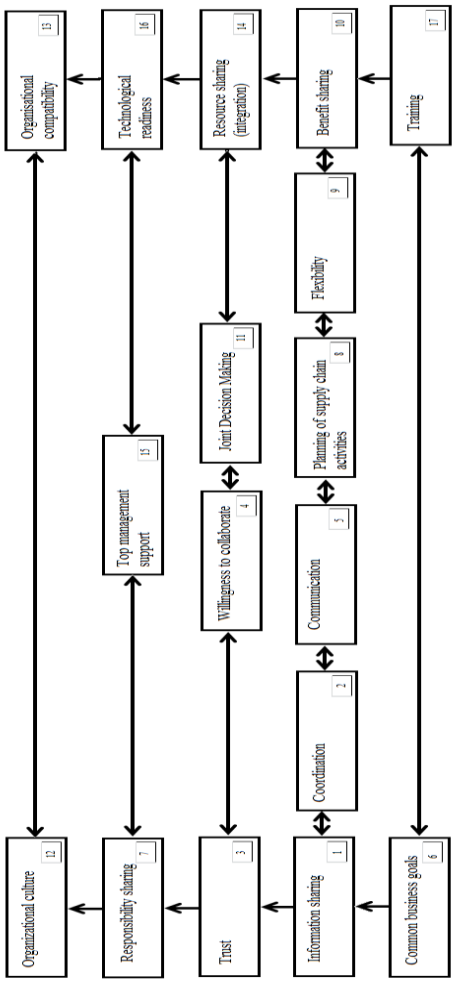


Fig. 2 The ISM based model for enablers for collaboration between logistics partners

VI. MICMAC Analysis

They have been classified into four categories:

• The first cluster consists of autonomous variables (Autonomous enablers). These enablers have a weak drive power and weak dependence. They are relatively disconnected from the system, with which they have few links, which may be very strong.

• The second cluster consists of the dependent variable (Linkage enablers). These have strong drive power as well as strong dependence. They are also unstable. Any action on them will have an effect on others and also a feedback effect on themselves.

• The third cluster has the linkage variables (Dependent enablers). This category includes those enablers which have a weak drive power but strong dependence power.

• The fourth cluster includes the independent variables (Independent enablers). These have a strong drive power but a weak dependence power. It is generally observed that an enabler with a very strong drive power, called the 'key enabler' falls into the category of independent or linkage enablers [18, 20, 24].

Enablers is constructed as shown in Fig. 3 (The driver power-dependence Matrix).

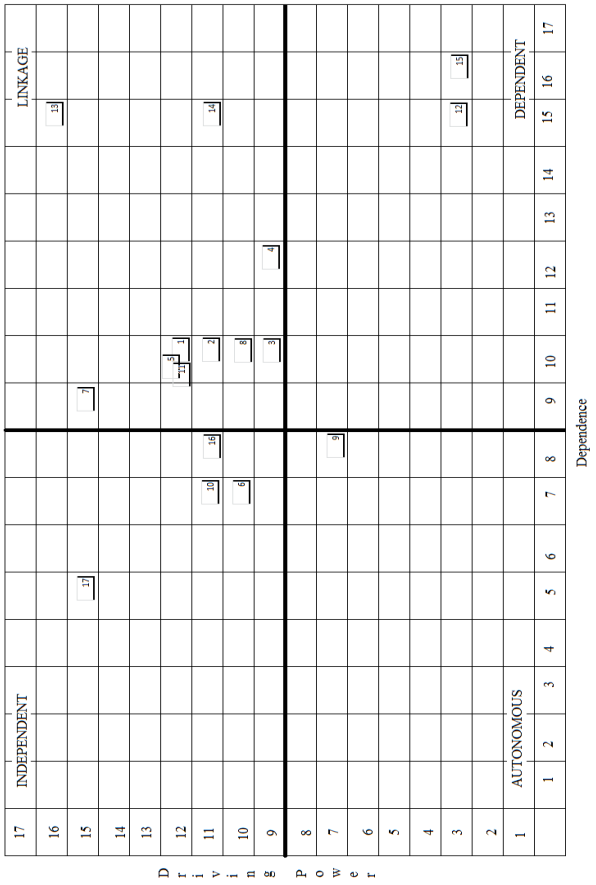


Fig. 3 Clusters of enablers in the implementation of collaboration between logistics partners

VII. Description of the Most Important Implementation Elements

In this research, an ISM-based model has been developed to analyse the interactions among different enablers [18]. The main objective of this research is to analyse the effectiveness of various enablers which help in the implementation of collaboration between logistics partners. The methodology proposed here identifies the hierarchy of actions to be taken for handling different enablers' ability to implement sustainability collaboration between logistics partners.

The managers can gain insight into these enablers and understand their relative importance and interdependencies. The driver dependence diagram gives some valuable insight about the relative importance and interdependencies among the collaboration sustainability enablers between logistics partners.

Some of the important implications emerging from this study are as follows: Figure 10 shows that in Autonomous enabler is one enabler: 'Flexibility' (enabler 9). Autonomous

enabler is weak driver and weak dependent and do not have much influence on the system.

Two enablers are Linkage enablers. Linkage enablers are 'Information sharing' (enabler 1), 'Coordination' (enabler 2), 'Trust' (enabler 3), 'Willingness to collaborate' (enabler 4), 'Communication' (enabler 5), 'Responsibility sharing' (enabler 7), 'Planning of supply chain activities' (enabler 8), 'Joint Decision Making' (enabler 11), 'Organisational compatibility' (enabler 13), 'Resource sharing (integration)' (enabler 14). They have a strong driving power as well as high dependencies [18]. If they are implemented in a proper way they can create a positive environment for the successful implementation of collaboration between logistics partners.

Enablers' 'Common business goals' (enabler 6), 'Benefit sharing' (enabler 10), 'Technological readiness' (enabler 16), and 'Training' (enabler 17) are Independent enablers. They have a strong driving power and weak dependency on other enablers.

'Organizational culture' (enabler 12) and 'Top management support' (enabler 15) are Dependent enablers. These enablers are weak drivers but strongly depend on one another. The managers should take special care to handle these enablers.

VIII. Conclusions and Future Works

This model proposed for the identification of enablers of sustainability collaboration between logistics partners can help in deciding the priority to take steps proactively. The results of this research can help in strategic and tactical decisions for a company wanting to create sustainability collaboration between logistics partners.

The main strategic decision relies on 'Common business goals' and 'Training'. These enablers, which are at the bottom of the ISM-based model and are the most important enablers that initiate strategic activities.

The analysis reveals that enablers 'Common business goals', 'Benefit sharing', 'Technological readiness' and 'Training' are ranked as Independent enablers as they possess the maximum driver power. This implies that these variables are key barriers in the successful implementation of sustainability collaboration between logistics partners. The most important among them are 'Common business goals' and 'Training'.

The ISM-based model provides a very useful understanding of the relationships among the enablers. The present model can be statistically tested with use of structural equation modelling (SEM) which has the ability to test the validity of such models [18].

References

[1] L. Horvath, Collaboration: key to value creation in supply chain management, Supply Chain Management, Vol. 6, no. 5, 205-7, 2001. http://dx.doi.org/10.1108/13598540810860994

[2] T.M. Simatupang, R. Sridharan, The collaborative supply chain, Logistics Management (MCB UP Ltd) Vol. 13, pp. 15-30, 2002. http://dx.doi.org/10.1108/09574090210806333

[3] T.M Simatupang, A.C. Wright, R. Sridharan, Applying the Theory of Constraints to Supply Chain Collaboration, Supply Chain Management: An International Journal, Vol. 9, No. 1, pp. 57-70, 2004. http://dx.doi.org/10.1108/13598540410517584

[4] C. Jr. Langley, (2000), 7 Immutable laws of Collaborative Logistics.

[5] Adetiloye, Taiwo Olubunmi, Collaboration Planning of Stakeholders for Sustainable City Logistics Operations. Masters thesis, Concordia University, 2012.

[6] M.A. Barratt, Understanding the meaning of collaboration in the supply chain, Supply Chain Management, Vol. 9, no. 1, 30-34, 2004. http://dx.doi.org/10.1108/13598540410517566

[7] L. Cassivi, Collaboration planning in a supply chain, Supply Chain Management: An International Journal, Vol 11, no. 3, 249 -258, 2006. http://dx.doi.org/10.1108/13598540610662158

[8] Garner Group, Collaborative - commerce: The new arena for business applications, Technical Report, Gartner, Inc. Stamford, Connecticut, 1999.

[9] O.R. Ovalle, A.C. Márquez, The effectiveness of using e-collaboration tools in the supply chain: an assessment study with system dynamics." Purchasing and Supply Management, Vol. 9, 151-163, 2003.

[10] M. Holweg, S. Disney, J. Holmström, J. Småros, Supply Chain Collaboration: Making Sense of the Strategy Continuum, European Management Journal, Vol. 23, no. 2, 170–181, 2005.

[11] R. Kale, P.T. Evers, M.E. Dresner, Analyzing private communities on internet - based collaborative transportation networks, Transportation Research E, Vol. 43: 21–38, 2007.

[12] J.A Muckstadt, D.H Murray, J.A. Rappold, D.E. Collins, Guidelines for Collaborative Supply Chain System Design and Operation, Technical report No. 1286, School of Operations Research and Industrial Engineering, College of Engineering, Cornel University ITHACA, NY 14853-3801, 2011.

[13] G. Zhou, Y.V. Hui, L. Liang, Strategic Alliance in Freight Consolidation, Transportation Research E, vol. 47, no. 1,pp. 18-29, 2011. http://dx.doi.org/10.1016/j.tre.2010.07.002

[14] A. Awasthi, K. Grzybowska, S. Chauhan, Goyal S K ., Investigating Organizational Characteristics for Sustainable Supply Chain Planning Under Fuzziness", Kahraman, Cengiz, Öztayşi, Başar (eds.), Supply Chain Management Under Fuzziness, Studies in Fuzziness and Soft Computing 313, Springer- Ferlag Berlin Heidelberg 2014.

[15] J. Mula, D. Peidro, M. Díaz-Madroñero, E. Vicens, Mathematical programming models for supply chain production and transport planning, European Journal of Operational Research, Vol. 204, (3), pp. 377–390, 2010. http://dx.doi.org/10.1016/j.ejor.2009.09.008

[16] P. Sitek, J. Wikarek, Cost optimization of supply chain with multimodal transport, Federated Conference on Computer Science and Information Systems (FedCSIS), 2012, pp. 1111–1118.

[17] P. Sitek, J. Wikarek, A hybrid method for modeling and solving constrained search problems, Federated Conference on Computer Science and Information Systems (FedCSIS), 2013, pp.385-392.

[18] T. Raj, R. Shankar, M. Suhaib, (2008), An ISM approach for modelling the enablers of flexible manufacturing system: the case for India, International Journal of Production Research, Vol. 46, No. 24, pp. 6883–6912. http://dx.doi.org/10.1080/00207540701429926

[19] S. Bhattacharya, K. Momaya, K. C. Iyer, Enablers of sustaining competitiveness: a case of growth strategies of top international construction companies, Global business review, Vol. 10, No. 1, pp. 45-66, 2009. http://dx.doi.org/10.1177/097215090801000103

[20] S. Jharkharia, R. Shankar, IT enablement of supply chains: modeling the enablers, International Journal of Productivity and Performance Management, Vol. 53, Issue 8, pp. 700–712, 2004. http://dx.doi.org/10.1108/17410400410569116

[21] V. Ravi, R. Shankar, Analysis of interactions among the barriers of reverse logistics, Technological Forecasting and Social Change, Vol. 72, Issue 8, pp. 1011–1029, 2005. http://dx.doi.org/10.1016/j.techfore.2004.07.002

[22] P. Charan, R. Shankar, R. K. Baisya, (2008), Analysis of interactions among the variables of supply chain performance measurement system implementation, Business Process Management Journal, Vol. 14, No. 4, pp. 512-529. http://dx.doi.org/10.1108/14637150810888055

[23] K. Grzybowska, Modelling the barriers of global supply chain, Management of Global and Regional Supply Chain – research and concepts, K. Grzybowska, Publishing House of Poznan University of Technology, pp.41-58, 2011

[24] K. Grzybowska, Supply Chain Sustainability – analysing the enablers, Environmental issues in supply chain management - new trends and applications, P. Golinska, C. A.Romano (Eds.), Springer, pp. 25-40, 2012. http://dx.doi.org/10.1007/978-3-642-23562-7_2

[25] C.K Hahn, E.A. Duplaga, J.L. Hartley. Supply chain synchronization: lessons from Hyundai Motor Company, Vol. 30, Issue 4, pp. 32-45, 2000.

[26] H. L. Lee, S. Whang, Information distortion in a supply chain, International Journal of Technology Management, Vol. 20, Issue 3-4, pp. 373-387, 2000.

[27] K. Grzybowska, G. Kovács, Logistics Process Modelling in Supply Chain – algorithm of coordination in the supply chain – contracting, 2014

[28] K. Grzybowska, "Creating trust in the supply chain", New insights into supply chain, K. Grzybowska (ed.), Publishing House of Poznan University of Technology, pp. 9-21, 2010

[29] R. G. Richey, F. G. Adams, V.Dalela, Technology and Flexibility: Enablers of Collaboration and Time-Based Logistics Quality, Journal of Business Logistics, Issue 33, pp. 34–49, 2012. http://dx.doi.org/10.1111/j.0000-0000.2011.01036.x

[30] M. Hussain, Modelling the enablers and alternatives for sustainable supply chain management. Masters thesis, Concordia University, 2011.

[31] A. Agarwal, R. Shankar, M.K. Tiwari, Modeling agility of supply chain, Industrial Marketing Management, 36 (4), pp. 443–457, 2007, http://dx.doi.org/10.1016/j.indmarman.2005.12.004

[32] P. Swafford, S. Ghosh, N. Murthy, Achieving supply chain agility through IT integration and flexibility. International Journal of Production Economics, Vol. 116, pp. 288-297, 2008. http://dx.doi.org/10.1016/j.ijpe.2008.09.002

[33] D. R. Farris, A. P. Sage, On the use of interpretive structural modeling for worth assessment, Computers & Electrical Engineering, Vol. 2, Issue 2-3, pp. 149-174, 1974. http://dx.doi.org/10.1016/0045-7906(75)90004-X

[34] M. D. Singh, R. Shankar, R. Narain, A. Agarwal, An interpretive structural modeling of knowledge management in engineering industries, Journal of Advances in Management Research, Vol. 1, No. 1, pp.28-40, 2003.

[35] J. Warfield, Societal Systems: Planning, Policy and Complexity, John Wiley & Sons, Inc., New York, 1973.

[36] S. K. Sharma, B. N. Panda, S. S. Mahapatra, S. Sahu, Analysis of Barriers for Reverse Logistics: An Indian Perspective, International Journal of Modelling and Optimization, Vol. 1, No. 2, pp. 101-106, 2011. http://dx.doi.org/10.7763/IJMO.2011.V1.18

[37] N. Kumar, R. Prasad, R. Shankar, K. C. Iyer, Technology transfer for rural housing: An interpretive structural modeling approach, Journal of Advances in Management Research, Vol. 6, Issue 2, pp. 188-205, 2009. http://dx.doi.org/10.1108/09727980911007208

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1321–1329
DOI: 10.15439/2014F75
ACSIS, Vol. 2

Sustainable Supply Chain - Supporting Tools

Katarzyna Grzybowska
Poznan University of Technology,
Strzelecka 11, 60-965 Poznan, Poland
Email:
katarzyna.grzybowska@put.poznan.pl

Gábor Kovács
Budapest University of Technology and
Economics, Műegyetem rkp 3, 1111
Budapest, Hungary
Email:
gabor.kovacs@logisztika.bme.hu

Abstract— The most important topic for researchers is supply chain, that takes into account environmental factors and social aspects. That is why top managers prefer taking into account key performance indicators currently. Harmonization of social, environmental and economic components makes development of supply chains sustainable. This document is based on environmental protection; it details the main features of sustainable supply chain. It presents supporting tools of collaboration in sustainable supply chains. The main examined areas: system identification, order picking, inventory control systems, city logistics, intermodal logistics processes, routing, and logistics processes of earthwork. The tools: neural network, simulation, genetic algorithm, ant colony algorithm.

The paper is structured as follows: First chapter defines sustainability and Sustainable Supply Chain (SSC). The second chapter presents supporting tools of collaboration in sustainable supply chains.

I. Introduction, Environmental Protection, Sustainable Supply Chain

Environmental sustainability depends on the interaction between organizations in supply chain, products and ecosystems [1]. Sustainability is playing an increasingly significant role in planning and management within organizations and across supply chains [2], [3], [4].

Supply chain management is defined as "the systemic, strategic coordination of the traditional business functions and the tactics across these business functions within a particular company and across businesses within the supply chain, for the purposes of improving the long term performance of the individual companies and the supply chain as a whole" [5].

Supply chain structure defines the way various organizations within the supply chain are arranged and related to each other. The supply chain structure falls into four main types [6]: Convergent: each node in the chain has at least one successor and several predecessors. Divergent: each node has at least one predecessor and several successors. Conjoined: which is a combination of each convergent chain and one divergent chain. Network: which cannot be classified as convergent, divergent or conjoined, and is more complex than the three previous types [7], [8], [9].

Sustainable supply chains are essential to sustain modern business growth and ensure healthy market environment [10]. In contrast to traditional SCM, which typically focuses on economic and financial business performance, sustainable SCM (SSCM) is characterized by explicit integration of environmental and/or social objectives which extend the economic dimension to the TBL [11]. In this context, SSCM focuses on the forward SC only [12] and is complemented by closed-loop SCM (CLSCM) [11], [12] including reverse logistics, remanufacturing, and product recovery [13].

Sustainable SCM is the management of material, information and capital flows as well as cooperation among companies along the supply chain while integrating goals from all three dimensions of sustainable development, i.e., economic, environmental and social, which are derived from customer and stakeholder requirements. In sustainable supply chains, environmental and social criteria need to be fulfilled by the members to remain within the supply chain, while it is expected that competitiveness would be maintained through meeting customer needs and related economic criteria. [4]

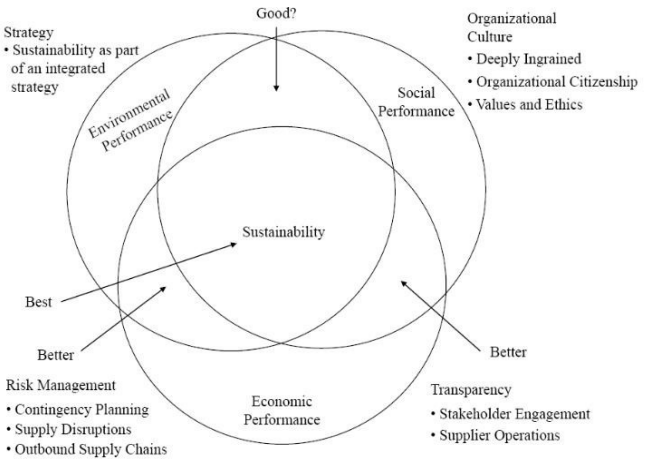


Fig. 1 Sustainable supply chain management [14]

This work is supported by Poznan University of Technology and Budapest University of Technology and Economics

Carter and Rogers define SSCM (Fig. 1) as the strategic, transparent integration and achievement of an organization's social, environmental, and economic goals in the systemic coordination of key interorganizational business processes for improving the long-term economic performance of the individual company and its supply chains [14].

In the article [15] the analysis reveals that six enablers 'Commitment from top management', 'Eco-literacy amongst supply chain partners', 'Corporate social responsibility', 'High level of supply chain integration', 'Waste management' and 'Logistics organisation ensuring goods safety and consumer health' are ranked as Independent enablers as they possess the maximum driver power. This implies that these variables are key barriers in the successful implementation of sustainability in the Supply Chain. The most important among them are 'Eco-literacy amongst supply chain partners', 'Commitment from top management' and 'Corporate social responsibility'.

Conceptualizing sustainability in three dimensions seems to be widely accepted [16, 17]. It allows an easy comprehension of the integration of economic, environmental and social issues. This also offers the justification of applying it in this paper. Most papers spend much more effort on explaining related environmental issues. In many cases, life-cycle assessment data forms the starting point for the analysis. Hence, energy demand and CO2-emissions (e.g. [18, 19, 20]) are among the frequently mentioned topics. Yet, in a number of cases, rather comprehensive lists of environmental impact criteria are taken up, such as referring to all kinds of natural capital (e.g. [21]) or resources, such as water or energy as well as waste (e.g. [22]) [4].

II. Supporting Tools of Collaboration in Sustainable Supply Chains

A. Automatic Identification, Neural Networks

In the last decade artificial intelligence (AI) methods come into prominence. The main reason is that the artificial intelligent methods are the mathematical models of human thinking and natural laws, therefore, a human-made decision support system (DSS) can behave similar way as an intelligent living being. With this ability the commonly used logistics methods can be developed in different fields such as planning and operation. In some cases the human intelligent can be swapped or complemented with artificial intelligence methods [23]. These methods could be inventory, scheduling, shortest route problems. The main purpose of [24] is to develop a time series analysis method in order to increase the demand forecast accuracy. The examined method is the Autoregressive Integrated Moving Average (ARIMA) model [23], [24], which has outstanding forecast accuracy in case of an auspicious identification. The aim is to show that the automatic ARIMA function identification can be accomplished with an artificial neural network (ANN) and with its learning ability the efficiency of identification is growing, see Fig. 2.

The automatic ARIMA (p, d, q) model identification described in [24] is a result of a continuous research and development. It was presented that the current identification methods and tests are hardly usable for "non ideal" time series and they are unable to adapt to the constantly changing characteristics of input data. On the other hand the model identification with a neural network is less sensitive to input errors through its intuitive capability, additionally after a certain number of training steps the algorithm is able to identify time series with unknown characteristics.

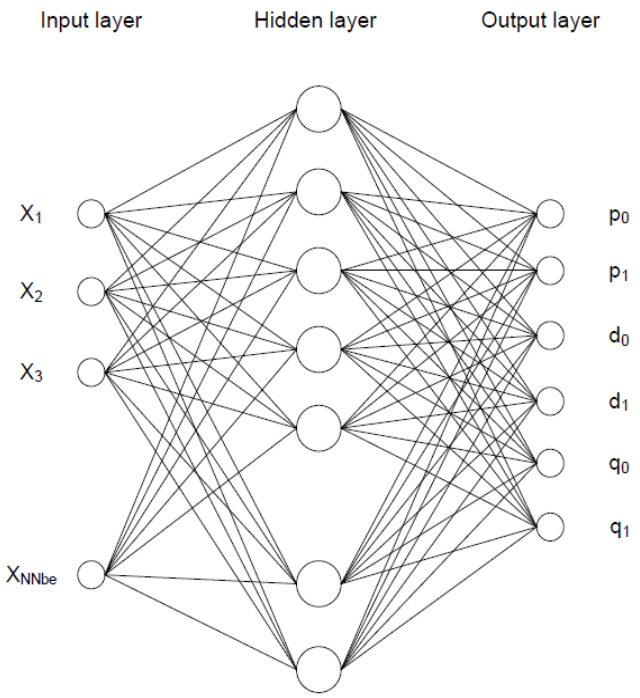


Fig. 2 Neural network [24]

These features make the method capable of integrating it into logistics processes. Of course the presented forecast method [24] is applicable to support other problems. The learning and intuitive capability of the artificial neural network is usable on any fields, where a prompt decision must be done with the support of previously acquired knowledge. It is effectively applicable solving inventory, scheduling, shortest route problems. The accuracy and efficiency of the method highly depends on the previously acquired knowledge, therefore, a great importance must be attached to the planning and operation. If the attention is made then the result is a robust, fast and flexible system wherein the unknown and random events can be treated effectively.

B. MULTI-CRITERIA SCHEDULING OF ORDER PICKING PROCESSES WITH SIMULTAN OPTIMIZATION

The flexibility of labour in a warehouse means that available personnel are redeployed during shifts to activities (storage, order picking, replenishment, etc.) where extra capacity is required. In the case when available labour capacity is not sufficient, temporary staff can be hired from specialized agencies. Order picking – retrieval of products from storage to meet customers' demand – is often the most labour-intense activity in a warehouse. The human hand as a 'handling equipment' is hard to be replaced and economical automation of retrieval of products is seldom possible. Therefore, the costs of order picking may amount to about half of the operational costs in a warehouse, so any improvement in this field may result in significant cost reduction [25, 26, 27, 28].

Both optimization and simulation are tools that support decision making. Optimization uses fixed input data, avoiding uncertainty and details. Optimization models simplify the complexity of the real system and some factors are even not considered. The simulation is not creative like optimization, but can cover uncertainty and complexity of dynamic systems in detail. The combination of optimization and simulation (simulation optimization) can be defined as the process of finding the best set of input variables without evaluating each possibility. The objective of simulation optimization is to minimize the resources spent (i.e. time) while maximizing the quality of information gained in the experiment. The model represented in [27] also uses the benefits of simulation optimization. The designed system supports operative warehouse management personnel in order to pick process scheduling and planning. By evaluating a number of scenarios, the number of the order pickers per shift, and the best sequence of releasing the pick lists to be retrieved from storage are determined [27]. It is the management's responsibility to monitor and control the order picking activities in the warehouse continuously and force the adherence to the schedule. If all order picking activities are realized according to the schedule, then the planning of the replenishment of the order picking places is also possible. The goal of [27] is to further develop the above described planning system and include the scheduling of these activities, too. The connection to the database of the WMS with the simulation model already exists so it is possible to determine when the last products will be picked from each picking place and when replenishment is necessary. By applying advanced search methods – like Genetic Algorithms – the optimal schedule for the replenishment of the picking places can be evaluated. The objective function must reflect the goal of planning the replenishment process so that the order picking processes can be executed continuously and undisturbed – products are available at the picking place and the congestion in the aisles is avoided. It is also the object of further development to improve the optimization algorithm and reach higher speed and accuracy in calculation. In this first version of the model [27], all probabilities to execute each genetic operator were constant. The proposed development (Fig. 3) will operate with variable probabilities for crossover and mutation to inherit the properties of the fittest individuals into the next population, to avoid premature convergence and to close-up the search space [27], [29].

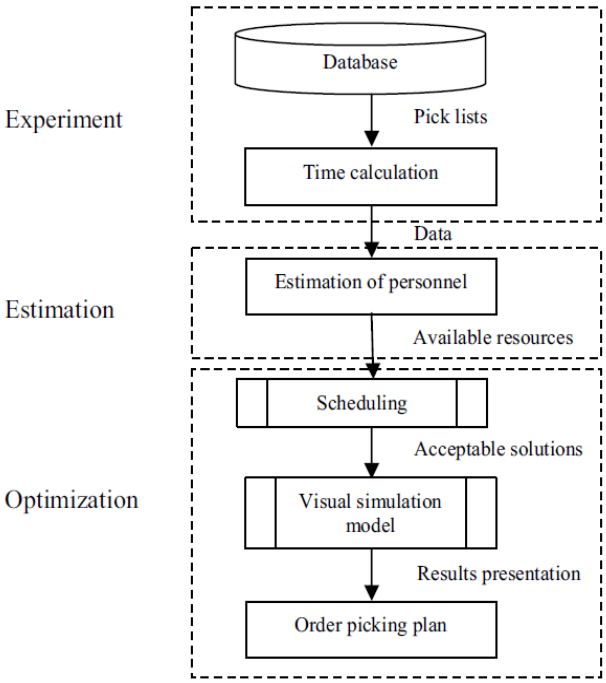


Fig. 3 Scheduling and decision support process for order picking planning [27]

C. OPTIMIZATION OF INVENTORY CONTROL SYSTEMS WITH GENETIC ALGORITHMS

The inventory control systems are responsible for the optimal operation of the inventory processes of the automotive companies. Generally, the optimization of the inventory control system manifests itself in a target conflict representing the implementation of the optimal operation in economic and reliability terms. For the process optimization, the control parameters of the regulation system should be defined. Their actual settings determine the time of placing orders of auto parts and the required quantities for the optimal operation of the processes defined above [30] presents a particular method of exploitation of the opportunities provided by the computer aided simulation and the genetic algorithms for the optimization of inventory control systems applying classical inventory mechanisms (Fig. 4).

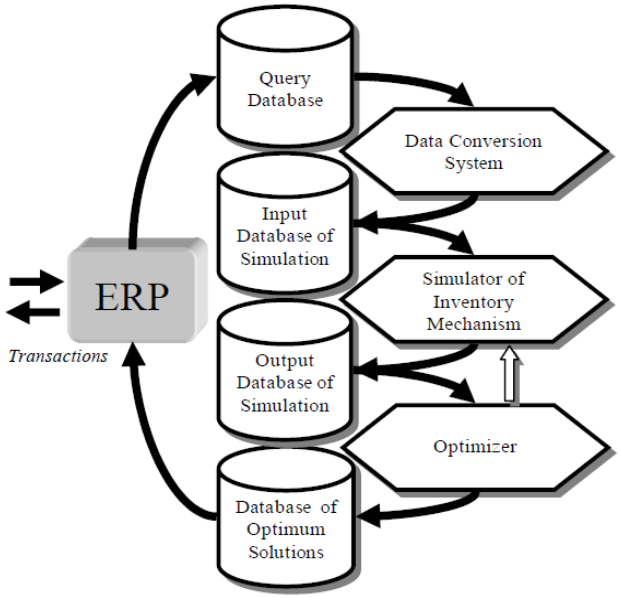


Fig. 4 Process of the dynamic inventory management [30]

The simulation inventory model presented in [30] and the binary genetic optimizing algorithm determining the control parameters showed beneficial properties in managing of stochastic inventory processes. For the establishment of proper applications, it is worthwhile to examine also the services rendered by the genetic algorithms operating with real number representation, as it is possible that this type of algorithm is able to provide the same results in a faster, more accurate way. Experiences show that the inventory processes in the future may constitute a special application field of the simulation supported optimization with genetic algorithms.

D. CITY LOGISTICS: DIFFERENT APPROACH, MODELS

The urban travel and land-use problems are not just urban problems. Their economic, social and environmental impacts extend well beyond the geographic jurisdictions of cities and towns to regions and to countries as a whole. The policies are designed to shape travel and land-use patterns to maximize the benefits of transport while minimizing their negative impacts. Given the broad spectrum of economic sectors and actors potentially impacted by urban travel and land-use activity, a package of complementary policy instruments needs to be developed that provides clear and well-targeted incentives to reduce the impacts of urban travel and land-use activities. This involves better integration of land use and transport planning. It involves finding ways to manage growth in car use and ensuring that alternative modes of travel by car – public transport, walking and cycling – are promoted. Fiscal and pricing instruments, legal and regulatory tools, currently available technology, and public information are some of the main policy tools available. A policy framework that embodies clear long-term objectives for urban travel may provide the essential parameters for implementation of integrated sustainable urban travel policies [31], [33].

By city logistics, we mean the technically, economically, organizationally efficient and environmentally friendly synchronization of goods distribution (and reverse logistics) tasks generated mainly by the secondary and tertiary sectors, and mostly retailers in downtown areas and historical city centers. There are many best practices to be found worldwide, that have already been identified and classified. Different city logistics system solutions affect the goods distribution of a city in various ways and magnitude, thus a model is desirable that helps the decision-making of stakeholders. [34] Is examining the indicators of cities housing city logistics system solutions, and some that are lacking those. The problem with the latter is that the shops, that form the demand scattered throughout the city, are not visited by their suppliers in a coordinated fashion, taking advantage of the common capacity, but rather they compete. Satisfying the demand takes place with presumably sub-optimal logistics-related costs [35]. This comes from the different suppliers transporting same types of goods to the same destinations with different – redundant – infrastructure, which could be avoided, according to [36]. Moreover, the attributes of the supply chains are adjusted to the regulation of the given municipality, but they seldom take advantage of certain possibilities (e.g. river, railways), and usually do not utilize integrated solutions, preferring road to multimodal transportation [32].

In order to assess the possibilities, [34] is developing a model that can compare various scenarios. The model is constantly evolving, but its fundamentals are: it maps an area with a graph, generates variable demand, and compares total costs. The model is basically static in structure: the different alternatives are constituted by nodes, and the transport system between each of these nodes. The demand is stochastic: the destinations and their daily demand (quantity of goods ordered) is a random variable. The total demand has to be satisfied with a – a priori unknown – number of vehicles. The common elements of the solutions are the location of the suppliers (LS), the urban consolidation centres (UCC), the urban relay stations (URS), and the urban loading points (ULP).

The number and location of these varies with each alternative. Further variable elements are the local and regional transport systems, with different vehicles and tracks. Accordingly, the model consists of a network, modelled as a graph. This structure of the model is suitable for the present and the planned systems, so they can be compared (Fig. 7).

The greatest challenge that can help the application of the model is the acquisition of more precise unit costs derived from logistics performance. External costs should later be included next to the existing ones, because a primary goal of a city logistics system solution is the reduction of the air and noise pollution and the augmentation of the standards of living. The fine-tuning of the model should produce precise

enough results that can point out an advantage of a specific alternative. The model, since it was developed generally, can be used extensively and in a wide number of cities and urban areas: the model parameters can be modified so as it can help decision-making at different locations [34].

Element	Type	Vehicle	Function
Location of suppliers (LS)	Node		Source of goods
Long-distance transport paths	Edge	Regional vehicles	Large-scale, homogeneous goods transport
Urban consolidation centres (UCC)	Node		Consolidation
Main urban transport paths	Edge	Local vehicles	Large-scale, heterogeneous goods transport
Urban relay stations (URS)	Node		Fast transhipment
Feeder urban transport paths	Edge	Last mile vehicles	Small-scale, heterogeneous goods transport
Urban loading points (ULP)	Node		Sink, points of sale

Fig. 5 Elements of a city logistics network [34]

As further possibility, in general terms, electronic freight and warehouse exchanges are virtual market places established for the harmonization of freight demand and supply. Due to their characteristics, however, these may be also suitable for the division of capacities (capacity load and storage capacity) of certain freighters and forwarders. Upon these trade directed to cities may be optimized, as groupage transport may be organized to cities, or within these to districts, considering, at the same time, the possibility of acquiring back haul, as well. In addition to this, by the division, optimal exploitation of freight capacities trade directed to the cities may be significantly reduced with the application of a transfer location in the outskirts of the city and a related warehouse exchange [36].

To effectuate a consistent methodology for urban planning – taking into consideration the viewpoints of land use and transportation – we need to approach the subject with considering complex social and economic aspects. To handle both of the mentioned urban planning areas together, we shall develop models, which are able to pay attention to all of their restrictive factors within the temporal properties as well. The efficiency of urban transportation is getting more and more important because of the increasing rate of mobility demand. To plan, control and organize urban transportation in the most efficient way, we also need to consider the aspects of land use. [37].

E. INTERMODAL LOGISTICS PROCESSES

The role of the intermodal logistic processes and related services are continuously changing and developing due to the spreading of transportation processes. One of the most frequent attribute of the service functions is the implementation parameters (for example place and material requirement) which are being changed, so the logistics system must be able to follow its flexibly. Because of the complexity, at any given time and location the implemented service requires cooperation between multiple logistics subsystems which are connected together only with the common management system and the endpoint of materials flow. One of the possible surface to satisfy the ever growing and changing claims if these services are supported by electronic freight and warehouse exchanges to perform the logistics processes.

The modified supply chain by electronic freight and warehouse exchanges is shown in Fig. 6. The main features of the modified supply chain system:

- The wholesalers are responsible for the information processes; they manage the demands of retailers.
- The logistics providers (storage providers, transportation providers, logistics centres) perform the physical freight and storage tasks, whereas they have:
  - suitable stock capacities,
  - suitable freight capacities,
  - logistics know-how.
- Electronic freight and warehouse exchanges perform the supply-demand (capacities-tasks) harmonization; the decision supporting and the optimization.
- In the case of logistics providers and electronic freight and warehouse exchanges the logistics processes are core.
- Due to the above the modified logistics system (e.g. combined transportation system) may be optimal.

Consequently, green logistics systems, e.g. green combined transportation supply chains can be realized (Example can be seen in Fig. 7). In addition, this system is beneficial not only for the individual actors (e.g. retailers, wholesalers, logistics providers, manufacturers, intermodal centres) but also for the national economy. The future plans include further development of algorithms and tests in real supply chains, e.g. supply chains of drink industry or other possible combined or complex city transportation system.

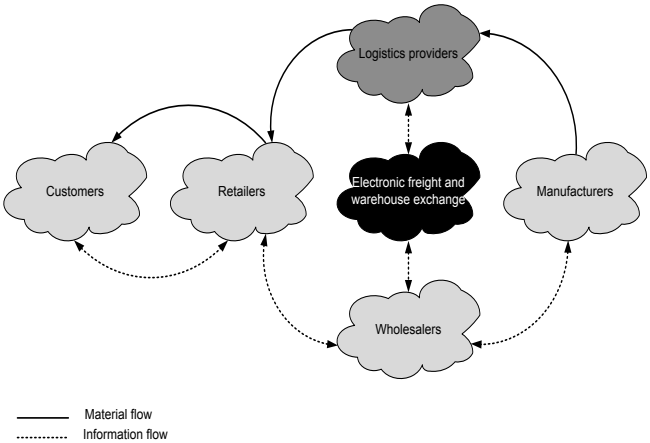


Fig. 6 The simplified system model of the supply chain supported by electronic freight and warehouse exchanges [38]

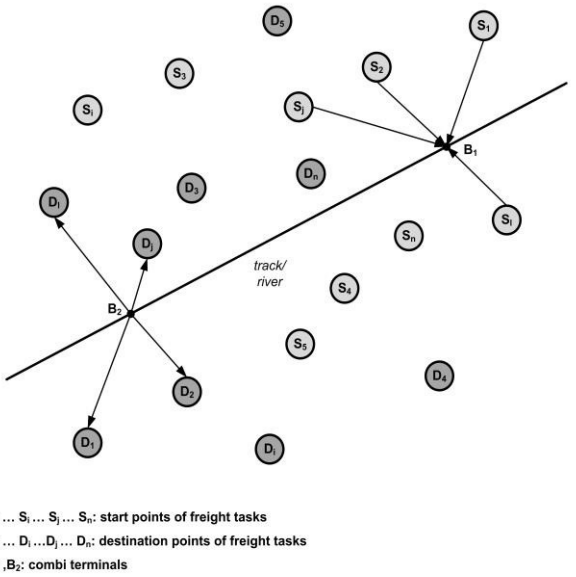


Fig. 7 Multimodal transportation supported by freight and warehouse exchange [38], [39]

F. ROUTING (TSP, VRP) IN ELECTRONIC FREIGHT AND WAREHOUSE EXCHANGES WITH ANT COLONY ALGORITHMS

The basic function of electronic freight and warehouse exchanges is to establish connection between free freight and storage capacities and tasks [40]. In the database of such online fairs there is high number of freight and storage capacity offers and tasks, which provides good optimization opportunity for those with free capacity [42].

The Vehicle Routing Problem (VRP) is used to design an optimal route for a fleet of vehicles to service a set of customers' orders (known in advance), given a set of constraints. The VRP is used in supply chain management in the physical delivery of goods and services. The VRP is of the NP-hard type [41].

In the freight exchange the optimum search task may be formulated on the basis of the following objective function: those having free freight capacity wish to establish routes providing optimal profit from the freight tasks appearing in the freight exchange. Many freight tasks may be included into the route, but a new freight task may be commenced only after the completion of the previous one.

The ACO (ant colony optimization) is an optimizing algorithm, a method developed by Marco Dorigo based on the modelling of the ants' social behaviour. In nature ants search for food by chance, then if they find some, on their way back to the ant-hill they mark the way with pheromone. Other ants – due to the pheromone sign – choose the marked way with higher probability instead of accidental wandering. Shorter ways may be completed quicker, thus on these ways more pheromone will be present than on longer ones. After a while the amount of pheromone drops (evaporation), by this preventing sticking to local optimum [42, 43].

In the electronic freight exchange similar problem emerges as the ants' search for food: the target is the performance of freight tasks offering the higher route level profit departing from the depot of the vehicle, with taking into account the limiting conditions. The problem, therefore, is twofold: on the one hand, the freight tasks to be performed shall be selected, and, on the other hand, their order shall be defined (FB_ACO, Fig. 8).

In the electronic warehouse exchange the task is simpler but is still similar to the food search: storage tasks shall be selected by taking into account storage capacity, possibly with the best possible exploitation of capacity (RB_ACO, Fig. 8).

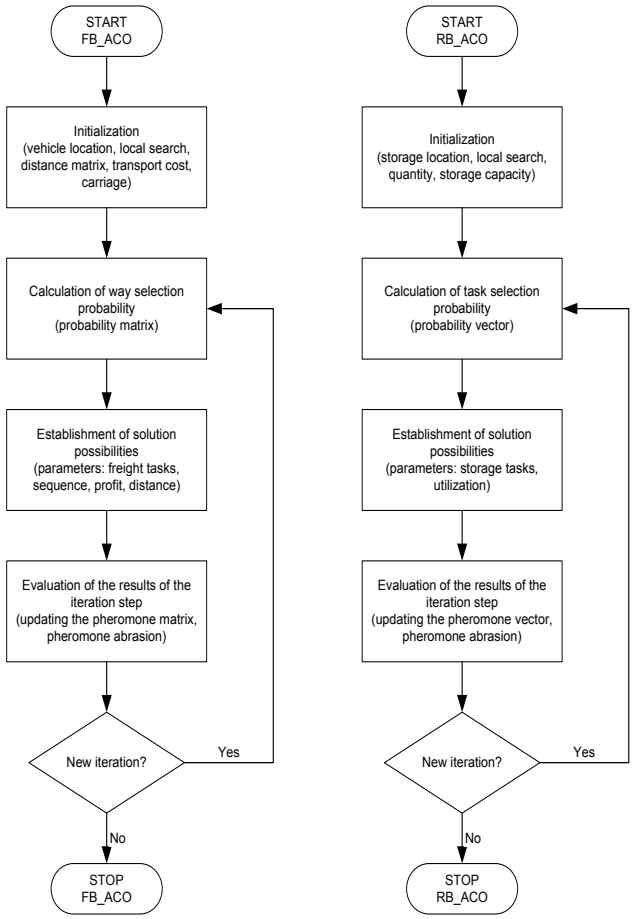


Fig. 8 The developed ant colony algorithms (FB_ACO and RB_ACO) [42]

G. SIMULATION IN EARTHWORK MODEL AND ITS LOGISTICS PROCESSES

Vehicle-based transport systems using automated guided vehicles (AGVs) are commonly used in facilities such as manufacturing plants, warehouses, distribution centers and transshipment terminals [44]. They are referred to as automated guided vehicle systems (AGVSs) [41], [45]. Early AGV systems were developed at universities and research institutes, named commonly as mobile robots. Later the industry realized advantages of autonomous transport vehicles for repeating transport tasks. One main application area for AGVs is the intralogistics or manufacturing

logistics, where the vehicles are mainly used for transporting raw materials, half-ready parts and ready products. These automated systems became moreover integral part of automated manufacturing systems as pointed out in [46]. The other important application is the automated container terminals, where AGVs transport various sized containers between the quay and the stacking area. More detailed system description find in e.g. [47]. Automation comes forward in other application areas as well [48]. Presents an automation example of a mobile excavator. The paper describes a pilot project investigates how autonomous functions could be realized on a mobile wheeled excavator. Rough terrain which characterizes construction sites however makes use of automated vehicles difficult [43]. Is aimed to help automation processes at the construction industry [43]. Concentrate on the logistics aspects, where organization of the material flow is an important task. The main of the focus of [48] to presents an automated soil transport system's simulation model which can be used to prove that intelligent systems may help construction processes as well. For the implementation agent-based approach is used.

[43] is organized as follows. First a short summary is given about the applicability of agent-based approaches. Next it describes which components or agents should be defined in different material flow systems using AGVs. In the following section detailed concept is presented for the implementation of the agent-based model for an earthmoving system (Fig. 9). Finally some remarks are made about the model's implementation in simulation environment.

The paper [43] surveyed applicability of agent-based principles for modeling material flow systems. An agent-based simulation model is also proposed. This model can be used for material flow systems' analysis not only for the case when each machine operates automatically, but due to the complex behavior for cases when the machines are operated through human workers. The proposed model is due to its modular construction can be adapted for different applications easily. Next step of the research is to build in so called information modules, which model information flow of the construction processes.

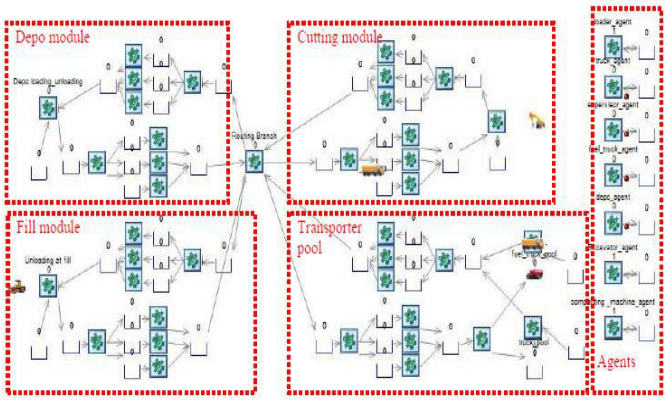


Fig. 9 Simul8 implementation of the proposed earthwork model

III. CONCLUSION

In this survey paper, there are several examples to optimizing the intra and extra logistics processes to. Automatic identification, neural networks, genetic algorithms, ant colony algorithms help find the optimal solution, and thereby to optimize the use of resources (e.g. human, material handling machines, transport machines, transport routes etc.), and to reduce the pollutants. Thus, sustainable supply chains can be realizing, that means economic and environmental efficiency too, whether it be city logistics supply chain, order picking processes or even earthwork.

We think that this topic deserves extensive additional research. Management studies and environmental science need to bridge the disciplinary distance that until now has characterized the two fields. There is a great need and urgency for further progress, and management theory must investigate the complexity of the relationship between organizations in supply chain and nature in order to support firms in the development of effective environmental strategies [49]. In this paper, we have stressed the importance of a holistic perspective in the service of sustainability. We argue that the individual firm is not the right unit of analysis for assessing environmental progress. Companies have many options to reduce their impact at the single organizational level (from clean technologies, to ecosystem restoration). But global ecological problems are not the result of a single firm's action. Ecosystem complexity over spatial and temporal scales requires close involvement and coordination across supply chains and industries as the appropriate unit of analysis for facing environmental problems [50,51].

REFERENCES

[1] P.R. Kleindorfer, K. Singhal, L. N. Van Wassenhove, "Sustainable Operations Management", Production and Operations Management 14 (4), pp. 482–492, 2005. DOI:10.1111/j.1937-5956.2005.tb00235.x

[2] J.D. Linton, R. Klassen,V. Jayaraman, "Sustainable Supply Chains: An Introduction", Journal of Operations Management, 25 (6), pp. 1175–1082, 2007. http://dx.doi.org/10.1016/j.jom.2007.01.012

[3] S.K. Srivastava, "Green Supply-Chain Management: A State-of-the-Art Review". International Journal of Management Reviews, 9 (1), pp. 53–80, 2007. http://dx.doi.org/10.1111/j.1468-2370.2007.-00202.x

[4] S. Seuring, A Review of Modeling Approaches for Sustainable Supply Chain Management, Decision Support Systems, 54 (4), pp. 1513–1520, 2013. http://dx.doi.org/10.1016/j.dss.2012.05.053

[5] A. Awasthi, K. Grzybowska, S. Chauhan, Goyal S K ., "Investigating Organizational Characteristics for Sustainable Supply Chain Planning Under Fuzziness", Kahraman, Cengiz, Öztayşi, Başar (eds.), Supply Chain Management Under Fuzziness, Studies in Fuzziness and Soft Computing 313, Springer- Ferlag Berlin Heidelberg 2014. http://dx.doi.org/10.1007/978-3-642-53939-8_5

[6] B.M. Beamon, V.C.P. Chen, 2001. Performance analysis of conjoined supply chains. International Journal of Production Research 39, 3195–3218. http://dx.doi.org/10.1080/00207540110053156

[7] J. Mula, D. Peidro, M. Díaz-Madroñero, E. Vicens, Mathematical programming models for supply chain production and transport

planning, European Journal of Operational Research, Vol. 204, (3), pp. 377–390, 2010. http://dx.doi.org/10.1016/j.ejor.2009.09.008

[8] P. Sitek, J. Wikarek, Cost optimization of supply chain with multimodal transport, Federated Conference on Computer Science and Information Systems (FedCSIS), 2012, pp. 1111–1118.

[9] Sitek, P., Wikarek, J., A hybrid approach to modeling and optimization for supply chain management with multimodal transport, IEEE Conference: 18th International Conference on Methods and Models in Automation and Robotics (MMAR), 2013, Pages: 777-782.

[10] S. Seuring, M. Müller, "From a literature review to a conceptual framework for sustainable supply chain management". Journal of Cleaner Production, 16 (15), pp. 1699–1710, 2008. http://dx.doi.org/10.1016/j.jclepro.2008.04.020

[11] B. Lebreton, "Strategic Closed-Loop Supply Chain Management". Lecture Notes in Economics and Mathematical Systems 586. Berlin: Springer 2007.

[12] V. D. R. Guide, L. N. van Wassenhove, "The evolution of closed-loop supply chain research". Operations Research, 57 (1), pp. 10-18, 2009. http://dx.doi.org/10.1287/opre.1080.0628

[13] M. Brandenburg, K.. Govindan, J. Sarkis, S. Seuring, "Quantitative models for sustainable supply chain management: developments and directions", European Journal of Operational Research, Vol. 233, Issue 2, pp. 299–312, 2014. http://dx.doi.org/10.1016/j.ejor.-2013.09.032

[14] C.R. Carter, D.S. Rogers, "Sustainable Supply Chain Management: Toward New Theory in Logistics Management," International Journal of Physical Distribution and Logistics Management, 2008, (38:5), pp. 360-387.

[15] K. Grzybowska, "Supply Chain Sustainability – analysing the enablers", Environmental issues in supply chain management - new trends and applications, P. Golinska, C. A.Romano (Eds.), pp. 25-40, Springer, 2010. http://dx.doi.org/10.1007/978-3-642-23562-7_2

[16] T. Dyllick, K. Hockerts, "Beyond the business case for corporate sustainability", Business Strategy and the Environment 11 (2), pp. 130-141, 2002. http://dx.doi.org/10.1002/bse.323

[17] C.R. Carter, P.L. Easton, "Sustainable supply chain management: evolution and future directions", International Journal of Physical Distribution & Logistics Management 41 (1), pp. 46-62, 2011. http://dx.doi.org/10.1108/09600031111101420

[18] S. Cholette, K. Venkat, "The energy and carbon intensity of wine distribution: A study of logistical options for delivering wine to consumers", Journal of Cleaner Production 17 (16), pp. 1-13, 2009. http://dx.doi.org/10.1016/j.jclepro.2009.05.011

[19] J.B. Edwards, A.C. McKinnon, S.L. Cullinane, "Comparative analysis of the carbon footprints of conventional and online retailing: A "last mile" perspective", International Journal of Physical Distribution & Logistics Management 40 (1-2), pp. 103-123, 2010. http://dx.doi.org/10.1108/09600031011018055

[20] R.B.H. Tan, H.H. Khoo, "An LCA study of a primary aluminum supply chain", Journal of Cleaner Production 13 (6), pp. 607-618, 2005. DOI: 10.1016/j.jclepro.2003.12.022

[21] J.F. Neto, J.M. Bloemhof-Ruwaard, J.A.E.E. van Nunen, E. van Heck, "Designing and evaluating sustainable logistics networks", International Journal of Production Economics 111 (2), pp. 195-208, 2008.

[22] N.U. Ukidwe, B.R. Bakshi, "Flow of natural versus economic capital in industrial supply networks and its implications to sustainability", Environmental Science and Technology 39 (24) (2005) 9759-9769.

[23] L.Y.Y. Lu, C.H. Wu, T-C. Kuo, "Environmental principles applicable to green supplier evaluation by using multi-objective decision analysis", International Journal of Production Research 45 (18), pp. 4317-4331, 2007. http://dx.doi.org/10.1108/17410381211196276

[24] G. Box, G. Jenkins, Time series analysis: Forecasting and control, San Francisco: Holden-Day, 1970.

[25] B. Lénárt, "Automatic identification of ARIMA models with neural network", Periodica Polytechnica Transportation Engineering, 39/1, pp. 39-42, 2011. http://dx.doi.org/10.3311/pp.tr.2011-1.07

[26] K. J. Roodbergen, Layout and Routing Methods for Warehouses, Ph.D. thesis, Erasmus University, Rotterdam, 2001.

[27] J. P. Van Den Berg, "A Literature Survey on Planning and Control of Warehousing Systems", IIE Transactions, 31, pp. 751-762, 1999. http://dx.doi.org/10.1023/A:1007606228790

[28] B. Molnár, "Multi-criteria scheduling of order picking processes with simultan optimization", Periodica Polytechnica Transportation Engineering, 33/1-2, pp. 59-68, 2005.

[29] D. Al-Dabass, D. Evans, M. Ren, "Observability in Hybrid Multi Agent Recurrent Nets for Natural Language Processing", IEEE 5th Int. Conference on Hybrid Intelligent Systems 6-9 November 2005, Rio de Janeiro, pp 506-508, 2005.

[30] D. Al-Dabass, A. Cheetham, D. J. Evans, "Simulation of a Multi-Dimensional Pattern Classifier", Int. J. of Computer Mathematics 71/2, pp. 197-233, 1999. http://dx.doi.org/10.1080/002071-69908804803

[31] K. Bóna, "Optimisation of inventory control systems with genetic algorithms", Periodica Polytechnica Transportation Engineering, 33/1-2, pp. 89-102, 2005.

[32] V. K. Banabakova, S. E. Stefanov, "Simulation model of logistic services machines, technologies, materials", Machines, Technologies, materials 7, pp. 28-31. 2013.

[33] Integration and Competition between Transport and Logistics Businesses, Discussion Paper, internationaltransportforum.org

[34] D. Kiss, "Sustainable development in urban transportation", Periodica Polytechnica Transportation Engineering, 29/1-2, pp. 147-157, 2001.

[35] A. Bakos, K. Bóna, Sz. Foltin, "The development of a complex city logistics cost model according to a multiple-stage gateway concept", Periodica Polytechnica Transportation Engineering, 40/1, pp. 17-20, 2012. http://dx.doi.org/10.3311/pp.tr.2012-1.03

[36] A. Bakos, "Modern Freight Distribution Model for Urban Areas", Proceedings of International Conference on Innovative Technologies, pp. 726-727, INTECH Conference, 2011.

[37] G. Kovács, "Possible methods of application of electronic freight and warehouse exchanges in solving the city logistics problems", Periodica Polytechnica Transportation Engineering, 38/1, pp. 25-28, 2010. http://dx.doi.org/10.3311/pp.tr.2010-1.05

[38] K. Tánczos, A. Török, "Introducing decisive development orientations into transport modelling", Transport 23(4), pp. 330-334, 2008. http://dx.doi.org/ 10.3846/1648-4142.2008.23.330-334

[39] K. Grzybowska, G. Kovács, "Developing Agile Supply Chains - system model, algorithms, applications", Agent and Multi-Agent Systems. Technologies and Applications, Lecture Notes in Computer Science, Jezic G. et. al. (eds.), Springer, pp. 576-585, 2012. http://dx.doi.org/10.1007/978-3-642-30947-2_62

[40] G. Bohács, I. Frikker, G. Kovács, "Intermodal logistics processes supported by electronic freight and warehouse exchanges", Transport and telecommunication, 14/3, pp. 206-213, 2013. http://dx.doi.org/DOI: 10.2478/ttj-2013-0017

[41] G. Kovács, "The structure, modules, services, and operational process of modern electronic freight and warehouse exchanges", Periodica Polytechnica Transportation Engineering, 37/1-2, pp. 33-38, 2009. http://dx.doi.org/10.3311/pp.tr.2009-1-2.06

[42] P. Sitek, A Hybrid Approach to the Two-Echelon Capacitated Vehicle Routing Problem (2E-CVRP), Recent Advances in Automation, Robotics and Measuring Techniques, Advances in Intelligent Systems and Computing Volume 267, 2014, pp 251-263. DOI: 10.1007/978-3-319-05353-0_25

[43] G. Kovács, "The ant colony algorithm supported optimum search in the electronic freight and warehouse exchanges", Periodica Polytechnica Transportation Engineering, 39/1, pp. 17-21, 2012. http://dx.doi.org/10.3311/pp.tr.2011-1.04

[44] G. Kovács, "Freight and warehouse exchanges: modern logistic information systems", Research in Logistics & Production 2(1) pp. 43-54, 2012.

[45] A. Rinkács, A. Gyimesi, G. Bohács, "Adaptive Simulation of Automated Guided Vehicle Systems Using Multi Agent Based Approach for Supplying Materials", Applied Mechanics and materials, 474/79, pp. 79-84, 2014. http://dx.doi.org/ 10.4028/www.scientific.net/AMM.474.79

[46] T. Le-Anh, M. B. M. De Koster, "A review of design and control of automated guided vehicle systems", European Journal of Operational Research, 171, pp. 1-23, 2006.

[47] R. Holubek, M. Vlasek, P. Kostal, "General Process Control for Intelligent Systems", World Academy of Science Engineering and Technology, 77, 2013.

[48] Z. Jianyang, H. Wen-Jing, “Conflict-free container routing in mesh yard layouts”, Robotics and Autonomous Systems, 56, pp. 451-460, 2007.
[49] M. Luck, P. McBurney, O. Shehory, S. Willmott, “Agent Technology: Computing as Interaction (A Roadmap for Agent Based Computing)”, AgentLink, 2005.
[50] S. Pogutz, M. Winn, “Organizational Ecosystem Embeddedness and its Implication for Sustainable Fit Strategies”, Academy of Management Annual Meeting Green Management Matters, Chicago, August, pp. 7-11, 2009.
[51] S. Pogutz, V. Micale, M. Winn, Corporate Environmental Sustainability Beyond Organizational Boundaries: Market Growth, Ecosystems Complexity and Supply Chain Structure as Co-Determinants of Environmental Impact, Journal of Environmental Sustainability, 1(1), 2011, http://dx.doi.org/10.14448/jes.01.0004

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1331–1336
DOI: 10.15439/2014F357
ACSIS, Vol. 2

Adaptive scheduling in dynamic environments

Hanno Hildmann, Miquel Martin
NEC Laboratories Europe
Kurfürsten-Anlage 36
D-69115 Heidelberg
Email: {hanno.hildmann, miquel.martin}@neclab.eu

***Abstract*—We present a method for fair agent scheduling in transportation scenarios. The approach is designed to first ensure the scheduling of all required task locations and, once this is achieved, focus on a balancing the workload across the population of transportation units. This, while almost certainly sub-optimal in the context of efficiency, facilitates the speedy allocation of new geographically located tasks due to the distribution of the remaining capacity across the agent population.**

We discuss our method, present results from simulations and discuss the advantages and disadvantages of the approach.

I. Introduction

LARGE scale transportation scheduling is a well known high complexity problem [1]. Arguably, the best known instance is the class of *Travelling Salesman Problems*, of which there are many variations which have been extensively discussed in the literature. Generally speaking, the problem is known to be NP-complete, meaning that while it is straight forward to check whether a given solution is correct, it is extremely difficult to construct such a solution [2].

In the classic *Travelling Salesman Problems* (TSP) one is concerned with finding the shortest path that connects a finite set of locations; in the multiple travelling salesman problem (MTSP) this is extended to a finite set of disjoint / mutually exclusive routes. The method presented in this paper solves a problem similar to the MTSP with the difference that we are not interested in the shortest path but in a collection of paths that do not exceed a certain length or agent capacity.

A. Motivation

The motivation for this is the idea that in the considered use cases we are in charge of a fleet of mobile agents which have a certain capacity (fuel, battery level, working time, etc) which we deem an acceptable investment.

While it is of course of interest to reduce the aggregated amount of the consumed resources we focus instead on the quick adaptation we can offer in a dynamic environment. Specifically, the empirical results we present are generated using a simulation that adds tasks to the problem after the initial solutions have been created. The motivation for this is that we consider scheduling scenarios where locations may be added throughout the active period of the fleet.

B. Aim of the paper

The aim of this paper is to present the method used and prove its performance through empirical results. The evaluation is intentionally kept generic and the focus is on providing the reader with all the information required to apply the method to specific problem instances. To this end, a variety of parameters are reported without any claim regarding their optimal settings (since we present results on a generic simulation it is of little interest to report on tuning results).

C. Application scenarios

The initial use case for the approach is the dispatch of service personnel to a number of locations, as major telecom operators or internet providers would routinely do. This can be extended to maintenance personnel as well as corporate security services on large estates. It should be noted that neither time windows, nor ordering locations according to some priority are currently supported.

We are at the moment considering the approach for allocating computational tasks to processing resources (e.g. servers in a server farm), also, the approach has potential in large scale disaster relief or humanitarian aid scenarios.

II. Method

We present a novel method [3] for adaptive scheduling in dynamic environments. At the core of the approach is a self determined paradigm shift [4] which enables a population of agents to switch their priorities on the basis of their (locally) available information: agents are assumed to have a limited view on the tasks, meaning that they are only aware of tasks within a certain distance from their position or their path. Their decisions are based on this limited partial visibility [5].

By *stance* we understand a predisposition to act one way or the other [6]. Specifically, agents are considered to be *maximizing* (which we will abbreviate to *max*) if they are aware of un-scheduled tasks, and *balanced* (*bal*) otherwise.

During each iteration all agents are triggered to locate a task within their reach (i.e. within their remaining capacity) and to attempt to assimilate this task into their own schedule. They will make this decision stochastically and depending on their and the other agent's remaining capacity. The other agent here is the agent to whom the task is currently scheduled, if there is no such agent the decision is foregone and the agent will simply schedule the task (cf. Figure 1). If the task is already scheduled to an agent then a value is calculated for both agents and from these a probability for re-allocating the task is derived.

978-83-60810-58-3/$25.00

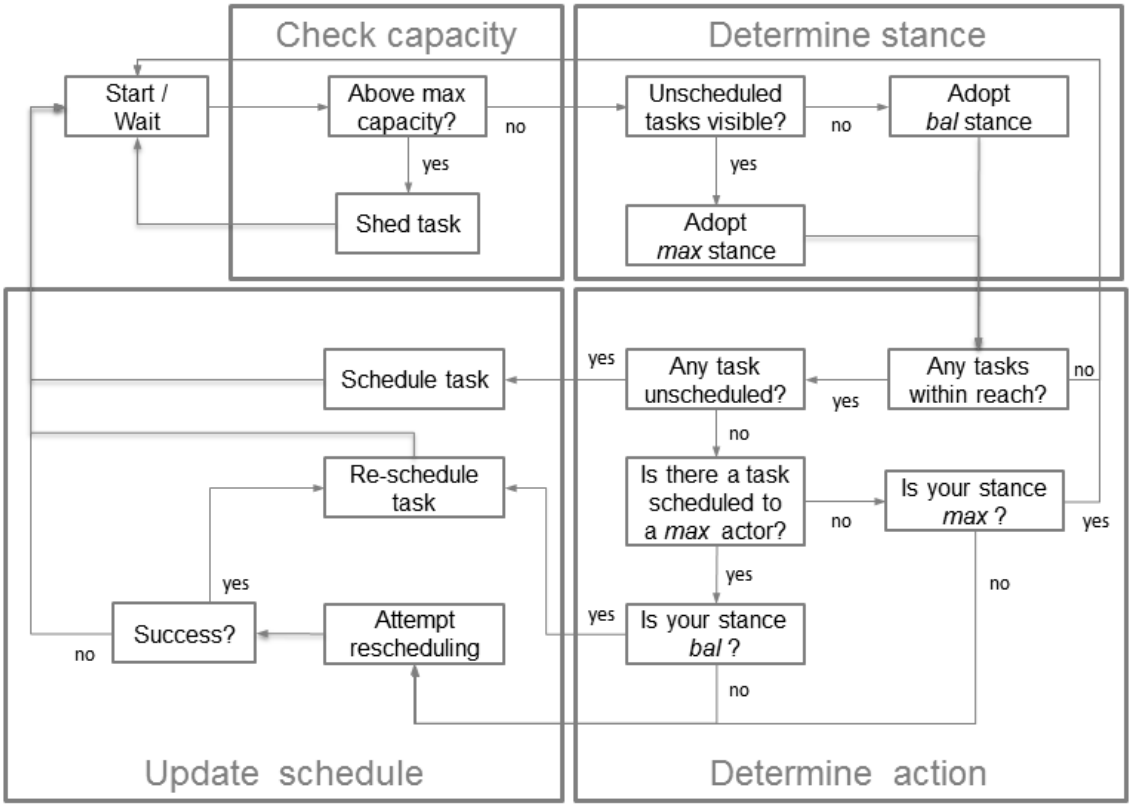


Fig. 1. Flow diagram of the approach

By reversing this value (the probability of successfully assimilating a task) we enable the agents to shift between two polar opposite stances: *rich gets richer* and *rich gets poorer*, where richness is associated with load (i.e. the higher the load the *richer* the agent). While the *max* stance means that an initially homogenous population will split into two groups: one that is maximised its load and one that is minimizing it, the *bal* stance results in agents with a higher load constantly losing tasks to agents with a lower load, effectively balancing their loads between themselves and their neighbours.

We require the approach to be non-deterministic (through the use of random elements like. e.g. the order in which agents are triggered, the choice of which task to consider for re-allocation and the stochastic decision whether they will win or lose the bid for re-allocating a task), as well as to be applied many times over. This will not result in an immediate convergence towards the best possible load distribution, but instead will slowly converge towards a *fair* state, i.e. the load distribution that results in roughly even loads for all agents.

Given that the choice of which task to consider is non-deterministic (and thus not related to the stance of either agent) we have 4 possible situations to consider (2 stances for each of 2 agents) when defining the interaction between any two agents. For the remainder of the paper we will call the initiating agent *active* and consequently call the other agent (the agent to whom the task in question is scheduled) *passive*. These 4 combinations and the resulting interactions are:

- **bal-bal** (*active* agent *balanced* - *passive* agent *balanced*): Agents that are not aware of unallocated tasks pursue their long term objective (load balancing). The agents will follow the *rich gets poorer* paradigm when calculating whether to reschedule the task.
- **max-max** (*active* agent *max* - *passive* agent *max*): If both agents are aware of un-allocated tasks they will focus on their short term objective, which is to quickly assimilate new/unscheduled tasks. They will follow the *rich gets richer* paradigm which will result in some agents depleting their capacity almost entirely while others are freed up, enabling them to assimilate the new tasks.
- **max-bal** (*active* agent *max* - *passive* agent *balanced*): If an agent is aware of unallocated tasks it is not interested to receive any tasks from an agent that is not. In this case, the interaction is aborted (this is simply prevented by restricting the choice of tasks, and thus agents).
- **bal-max** (*active* agent *balanced* - *passive* agent *max*): Balanced agents that happen to come across a task currently scheduled to an agent that is aware of unallocated tasks will always take the task in question. The reasoning is that the overall aim is to remove as many tasks as possible from the agents that are struggling to accommodate unallocated tasks.

III. Model

A. Parameters

The following are the parameters used for the model:

- number of agents
- number of tasks
- capacity (the same for all agents per simulation run)
- visibility range (considerably smaller than capacity)
- map size (this was 100x100 for all simulations)
- single depot versus individual depots

In line with our intention to enable the reader to implement and evaluate the approach independently, we also briefly discuss the only tuning parameter which we have included in this paper: α (used in the formulae in §III-B). Changes in α will affect the speed with which the approach converges to a somewhat stable solution, while on the other hand determining the degree of change in the environment which the algorithm can handle while still performing well.

B. Formulae

The decision whether to re-allocate a task from the active agent to the passive agent is stochastic. The probability of a *max* agent stealing as task from another *max* agent is:

$$P_A^{max} = \frac{(rem.capacity_B)^\alpha}{(rem.capacity_A)^\alpha + (rem.capacity_B)^\alpha} \quad (1)$$

with P_A^{bal} the probability of agent *A* stealing a task from *B*, and $rem.capacity_X$ the remaining capacity of an agent *X*.

The corresponding P_A^{bal} is simply the opposite of P_A^{max}:

$$P_A^{bal} = 1 - P_A^{max} \quad (2)$$

In other words, $P_A^{bal} = P_B^{max}$, which preserves the symmetry of the probabilities in the opposing stances, i.e. if *A* were likely to win for *rich gets richer*, it should be equally likely to loose for *rich gets poorer*.

C. Scenario

a) Agents: have a capacity, a visibility range, a starting depot and a route. The capacity is static, while the *remaining capacity* is the capacity minus the cost of the current route.

b) Tasks: have a location, expressed by x-y coordinates.

c) Depots: mark the beginning and the end of each route. Each agent is assigned a depot, expressed by x-y coordinates.

d) Routes: start and end with the depot of the respective agent. Routes are assigned a cost, which is the sum of the travel distances between the individual entries in the route. The distance is calculated as the Euclidean Distance.

e) World: has a map, a list of agents and a list of routes (schedules). The act of re-allocating a task from one route to another is assumed to be instantaneous and cost free.

IV. Simulations

A. Scenarios

Two scenarios were used for the simulations:

1) 40 tasks were added to the map once (at iteration 25): This is used for smaller scenarios, where the 40 tasks constitute a substantial percentage of the existing tasks, and where a small number of agents attempt to efficiently schedule the new tasks. This scenario is used to see whether, and how fast, the new tasks can be allocated and how long it takes for the average of the individual schedules to converge to a somewhat stable value.
   All tasks were randomly given coordinates in the range ($[50, 70]$, $[-50, 50]$); since the board is initialized to $[-50, 50]$ for both x and y coordinates these new tasks appear across the range of the y axis and further on the x than any other tasks.
2) 100 tasks are added every 25 iterations:
   This scenario is used to investigate the robustness of the approach and to see how a simulated population holds up in the face of a task load that is getting closer and closer to the total capacity of the population. Contrary to the above these tasks were placed all over the map ($[-50, 50]$, $[-50, 50]$).

B. Route calculations

We are interested in the least cost for a schedule, that is, the shortest path thorough all tasks in a list. This is the well known *Travelling Salesman Problem*, which is known to be NP-complete. Since the algorithm will need to calculate this for every task in the vicinity of an agent when computing the list of visible or reachable tasks, this is a calculation that is performed many times (millions of times, to be more precise, for any of the presented simulations).

To reduce the computation times two steps are taken: First of all, for routes of short length ($\leq$8) the best route is calculated via brute force, as this has proven to be the most effective approach for us. Secondly, for longer routes, we use simulated annealing with very soft convergence criteria. Because paths are recalculated and built upon at every iteration, repeatedly accepting a sub-optimal path still tends to progressively improve the solutions, yielding good results.

C. Extremely large scenarios

We investigated the performance of the approach for very large scenarios to verify that the approach is a) scalable and b) does indeed perform better with larger agent populations. To this end we simulated a scenario with 10,000 initial tasks and 1000 agents. We then added 100 tasks every 25 iterations and let the simulation run for 1000 iterations. This was done 4 times with different seeds. The generation of the results took 4-5 weeks because of the massive number of route calculations that had to be done.

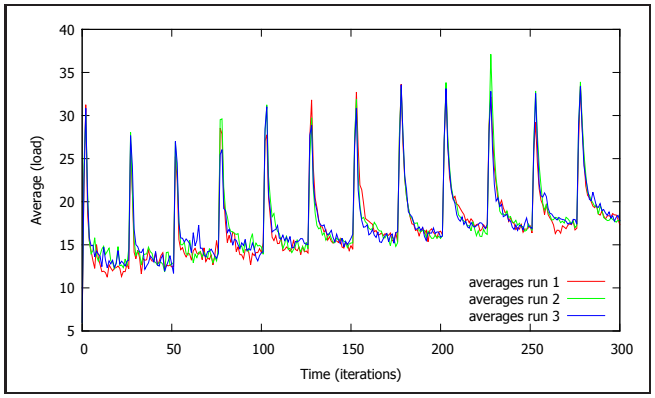


Fig. 2. Three separate runs of the simulation (x-axis: iterations; y-axis: load). Simulation setting: Scenario 2, agents = 50, tasks = 100, capacity = 75, 300 iterations, visibility = 40, map = 100x100. Δ: 100 new tasks every 25 steps

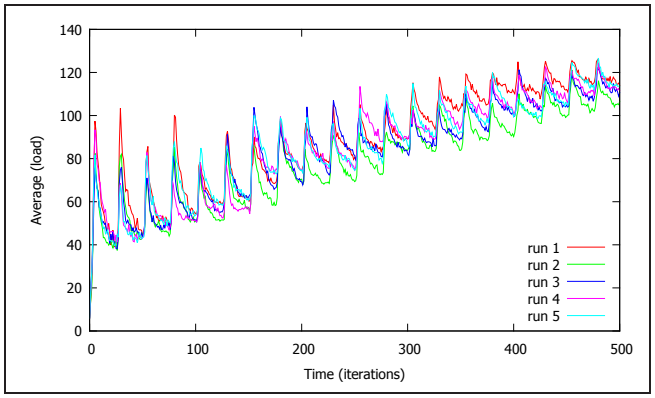


Fig. 3. Five separate runs of the simulation (x-axis: iterations; y-axis: load). Simulation setting: Scenario 2, agents = 50, tasks = 100, capacity = 150, 500 iterations, visibility = 40, map = 100x100. Δ: 100 new tasks every 25 steps

V. Results and Discussion

A. Parameter space exploration

We ran a number of identical simulations (for each agent population size, simulations were ran multiple times with different seeds, the seeds used for the different population sizes were the same so as to ensure comparability of the results) where we increased the number of agents to investigate the relevance of the size of the agent population. To this end, Scenario 1 was used and separate batches were run for small populations (20-100 agents, increasing in steps of 10) and large populations (200-1000 agents, increasing in steps of 100). The adaptation of the new tasks worked equally well across all population sizes, with tasks being assimilated into schedules as fast as possible (considering that each agent can only add one task per iteration; larger populations could assimilate tasks faster). The same holds for the aggregated load as well as the individual agents' loads. Other than the obvious differences in the performance, the algorithm performed well, even for very small numbers of agents (e.g. 20 agents for 500 tasks).

B. Performance evaluation

1) Adaptation of new tasks: To investigate the adaption of new tasks into schedules we looked at the average load of the *max* agents (as their the presence indicates the existence of unallocated tasks). For very small numbers of agents (fewer than the number of new tasks that were added) there is a noticeable difference, but as soon as the number of agents equals the number of new tasks there is no more difference in the number of turns it takes for all new tasks to be assimilated.

2) Optimization of the aggregated load: Addition of new tasks accounts for a sudden increase in the averages. For all but the smallest agent populations, however, we can see a steady and fast decrease in the reported averages. As expected, smaller populations struggle more with the process of decreasing the aggregated workload. This can be explained by the fact that for smaller populations the exchange of tasks is much more likely to have a more dramatic impact on their route length. This was also reflected in the standard deviations.

3) Optimization of the individual agents' loads: Population size has an impact on the standard deviation itself as well as on the rate of its decline. Smaller populations seem to quickly decrease the standard deviation between time step 120 and 140. This was, however, not reflected in the averages themselves. For larger populations we witnessed a much smoother return to small standard deviations, with the largest populations reaching a plateau very quickly.

4) Adaptivity and resilience: In order to evaluate the resilience of the approach and to investigate how the approach performs when it subjected to repetitive heavy strain, we ran a number of concurrent simulations for 300, 500 and (due to time constraints) one single simulation for 1000 time steps. All simulations started with 100 initial tasks which were supplemented by another 100 every 25 iterations thereafter (Scenario 2). The first set of simulations increased the number of tasks to 1200, the second set to 2000 and the final simulation ended with 4000 tasks all-together. While the first of these 3 batches was run with a capacity of 75, the latter two used a capacity of 150 (so as to provide the agents with the capacity required to accommodate all the new tasks). As before, we investigated the performance with regard to the aggregated load of all agents as well as the load of the individual agents:

a) Aggregated load of the population: Figures 2, 3 and 4 show the average loads of the agents over the course of the simulations. Even for large (and very large, as depicted in Figure 7) numbers of added tasks, the rate of convergence towards a good average remains stable. The overall increase of the load is not very large, and is decreasing with time as new tasks are added. This is understandable since the agents are using the shortest path through all tasks to calculate their load, and for increasing task numbers new tasks are more likely to be very close to already scheduled tasks.

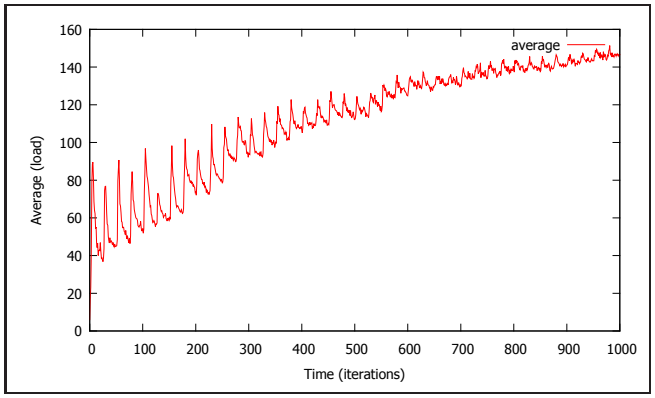


Fig. 4. The average load (y-axis) for one long simulation (x-axis: iterations). Simulation setting: Scenario 2, agents = 50, tasks = 100, capacity = 150, 1000 iterations, visibility=40, map=100x100. Δ: 100 new tasks every 25 steps

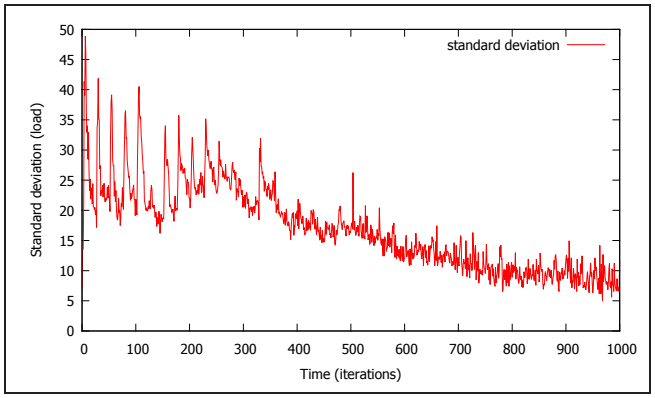


Fig. 6. The standard deviation (y-axis) from the average load for agents over 1000 iterations (x-axis). Setting: Scenario 2, agents = 50, tasks = 100, capacity = 150, visibility = 40, map = 100x100. Δ: 100 new tasks every 25 steps

b) Performance of max agents: Regarding the performance of the *max* agents, we can see in Figure 5 that each time new tasks are added there is a quick burst of activity before the averages drop again sharply. This corresponds to the number of *max* agents dropping to zero and the balanced agents optimizing their schedules. Besides showing us that the time used for the assimilation of the new tasks does not increase with larger task counts, it also demonstrates the effectiveness of the paradigm shift. These claims are best discussed in combination with the next paragraph.

c) Load balancing between individual agents: Regarding the performance of the individual agents, the graphs presented in Figure 6 show the standard deviation from the averages (already discussed above). We can see that the deviation from the average more than doubles briefly as the balanced agents take on tasks from the *max* agents. However, in all simulations it then drops almost immediately to the previous values. There seems to be a turning point around time step 200 (or when the task count has reached 900) after which the deviation is slowly decreasing. This is to be expected: with an increasingly denser task distribution we can realistically assume that the individual agents can balance their load more minutely.

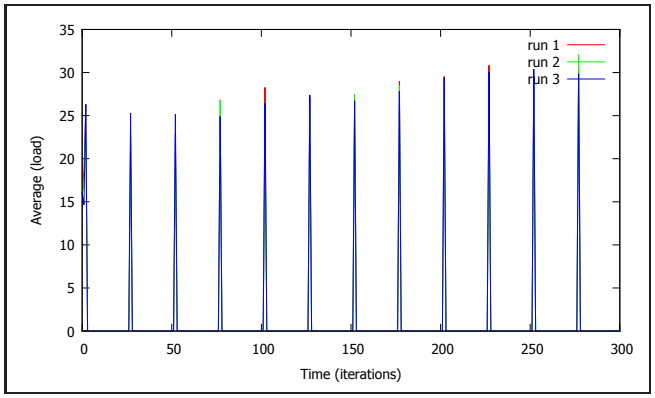


Fig. 5. The averages (y-axis) of the *max* agents for 3 separate runs (x-axis: iterations). Setting: Scenario 2, agents = 50, tasks = 100, capacity = 75, 300 iterations, visibility = 40, map = 100x100. Δ: 100 new tasks every 25 steps

d) Performance of balanced agents: We observed that the averages of the balanced agents jumped each time a new batch of tasks was added to the map but quickly dropped again, with the standard deviation following shortly thereafter.

Figures 4 and 6 report on results from small agent populations while Figures 7 and 8 show the performance of very large populations, both running for 1000 iterations, but with widely different task counts.

In Figure 4, the averages of the balanced agents increased slowly for smaller agent populations (capacity = 150) but their averages converged towards a plateau at around 120. The standard deviation, converging towards 25 (which is about the remaining capacity left for the agents) explains this plateau.

Regarding the corresponding standard deviation (Figure 6), we noticed a decrease in the deviation with increasing task counts. This matches the increased averages (decreased remaining capacity) but can also be explained by the fact that in larger task count new tasks are placed in closer proximity of existing (and already scheduled) tasks. In addition, the load balancing improves with the number of tasks to exchange.

As for the much larger simulation reported upon in Figures 7 and 8, we argue that the results show that the approach is scalable. We furthermore suggest that the increases in average load and standard deviation are stable because, unlike the other simulation discussed above, the agents had sizable remaining capacity and the sheer number of the agents in the simulation resulted in a much higher chance of tasks being placed in close proximity of agents.

The patterns observed in the other simulations are repeated here: after tasks are added a brief period of ensues within which the agents change paradigm, assimilate the new tasks and increase their load. This is followed immediately by a steady optimization which reduces the average load to (almost) the values from before the adding of the tasks. As far as the standard deviation is concerned, while the above graph (Figure 6) shows a convergence towards somewhat stable values, the much larger simulation (Figure 8) does not. This is because in the larger scenario agents can still assimilate far away tasks.

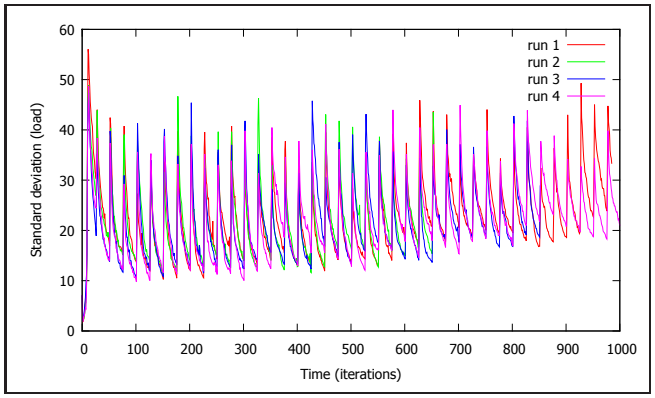


Fig. 8. The standard deviation (y-axis) from the average load for 4 runs (x-axis: iterations) of a very large simulation (Scenario 2, with 1000 agents) starting with 10,000 tasks and increasing to 14,000 tasks. Setting: see Fig. 7

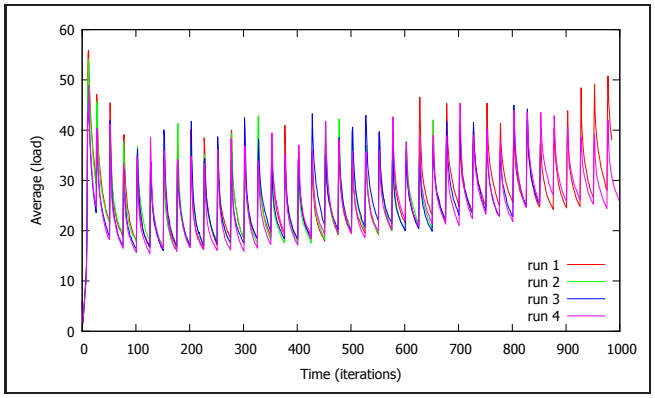


Fig. 7. The average load (y-axis) of the agents for 4 runs (x-axis: iterations) of a very large simulation (Scenario 2, with 1000 agents) that started with 10,000 tasks and over the course of 1000 iterations increased to 14,000 tasks. Setting: capacity = 400, visibility = 40, map = 100x100. Cf. also Fig. 8

C. Restrictions to the approach

The proposed method relies on a number of agents working in proximity, such that the schedule assigned to a specific agent can be partly absorbed into another agent's schedule. It is expected that there is a critical mass which is required in order for the method to outperform other approaches. We tested the approach against small numbers (as few as 20) of agents and it performed well; however the critical mass depends on the size of the map, the visibility and the capacity of the agents.

Furthermore, there is an upper limit to the degree of change over time in which the method can be expected to perform well. If the changes are too rapid or dramatic, recalculating the entire solution will be the better approach. This is due to the iterative nature of the approach, which will quickly adapt to changes and follow moving centers of gravity in the problem space. If such centers appear and disappear seemingly at random it becomes impossible to follow them, and thus the method loses its edge over other approaches.

D. Closing remarks

The combination of probabilistic decision making as well as the balancing and maximizing paradigm results in the minimization of agents to which a subset of the tasks (i.e. tasks from some region of the problem space) is scheduled to. This means that under the right conditions some agent might have very few or no tasks scheduled, which enables them to take on tasks from other regions of the problem space. This is expected to result in the ability to cross over large distances and effectively to reallocate agents within a problem space. Local optimization techniques normally consider only local alterations to a solution and recalculating the complete set of schedules for the whole problem space may be expensive in terms of computational effort.

REFERENCES

[1] R. Xu and I. Wunsch, D., "Survey of clustering algorithms," *Neural Networks, IEEE Transactions on*, vol. 16, no. 3, pp. 645 –678, May 2005.

[2] H. Hassan and A. Al-Hamadi, "On comparative evaluation of Thorndike's psycho-learning experimental work versus an optimal swarm intelligent system," in *Computational Intelligence for Modelling Control Automation, 2008 International Conference on*, Dec. 2008, pp. 1083 –1088.

[3] Patent, "Ditributed task scheduling using multiple agent paradigms," US Patent No: 14/250,470 (filed), H. Hildmann and M. Martin (inventors), Mar. 2014.

[4] S. Camazine, J.-L. Deneubourg, N. R. Franks, J. Sneyd, G. Theraulaz, and E. Bonabeau, *Self-Organization in Biological Systems*. Princeton Univ Press., 2001.

[5] E. Bonabeau, M. Dorigo, and G. Theraulaz, "Inspiration for optimization from social insect behaviour," *Nature*, July 2000. [Online]. Available: http://dx.doi.org/10.1038/35017500

[6] D. M. Gordon, "The organization of work in social insect colonies," *Nature*, vol. 380, pp. 121–124, Mar. 1996. [Online]. Available: http://dx.doi.org/10.1038/380121a0

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1337–1343
DOI: 10.15439/2014F364
ACSIS, Vol. 2

Information System Framework Architecture for Organization Agnostic Logistics Utilizing Standardized IoT Technologies

Dimitris Karadimas
Industrial Systems Institute/RC Athena, Platani Patras, Greece
Email: karadimas@ieee.org

Elias Polytarchos
Athens University of Economics and Business, Athens, Greece
Email: e.polytarchos@gmail.com

Kyriakos Stefanidis
Library & Information Center, University of Patras, Greece
Email: stefanidis@ece.upatras.gr

John Gialelis
Industrial Systems Institute/RC Athena, Platani Patras, Greece
Email: gialelis@isi.gr

Abstract— Logistics or supply-chain services provide enterprises and organizations with the necessary level of flexibility and efficiency in order to retain competitiveness under the increasingly turbulent e-business area. Web-Services are utilized by organizations in order to integrate high and low level applications, thus providing a collaborative environment without affecting inter- and intra-enterprise processes. Nevertheless, the above context should be enhanced in order to comply with the Web-of-Things concept. This paper describes a sustainable approach towards the above requirement by employing ONS based services able to provide targeted information regarding RFID-enabled physical objects that are handled in an organization agnostic collaborative environment.

I. Introduction

RADIO Frequency Identification (RFID) technology has already delivered revolutionary aspects in various areas such as logistics (supply chains), e-health management and materials identification / traceability. RFID technology itself allows an object's identification with effectiveness and efficiency. However traceability of an object calls for a robust and reliable system operating seamlessly over its entire lifecycle. Such a traceability system has to be implemented so as: a) its data model allows unique object identification and scalable, often big-data, databases, b) its underlying framework supports interoperability and c) its mechanism is capable to achieve end-to-end tracing providing full history information.

Despite RFID technology's nature in tracking, there are several challenges that need to be addressed. Since an RFID tag can be read from a quite long distance without requiring line-of-sight, it is possible that collisions may occur whereas also multiple tags could be read simultaneously. Therefore, there is no guarantee that a single tag will be consecutively detected on consecutive scans. Moreover, the use of RFID may constitute a serious threat for the information privacy, as it could be easily facilitated to espionage or unauthorized requests.

Based on the previously described situation there is a real need for an underlying framework able not only to support and complement the tracing functionality offered by the RFID technology but also to take into consideration the relevant tracking information of a physical object through its entire lifecycle in a secure and effectively protected way.

The work presented in this paper focuses in the challenging concept of an RFID-enabled, organization unaware, logistics management, by introducing an architecture that utilizes components of the Electronic Product Code (EPC) global network, such as the Object Naming Services (ONS) and the EPC Information Services (EPCIS), in order to support the Internet of Things concept. Our implementation is capable of enhancing the architecture of e-business frameworks, thus introducing an innovative collaborative business model which seamlessly integrates the inter-enterprise (public) with the intra-enterprise (private) processes.

The implemented architecture is demonstrated by the presentation of a case study in documents (books, papers, etc.) tracking and management in an academic library environment. Despite the fact that such an environment has quite lot variations from an e-business logistics environment, it has the potential to illustrate (and, even, simulate) the key-points of the presented architecture when not a standalone but a whole network of academic libraries are taken into account.

II. Current Status

Traceability is defined as the ability to trace the history, application or location of an entity, by means of recorded identifications. It also may be defined in general as the ability to trace and follow any product through all stages of production, processing and distribution. Traceability itself can be divided into three types:

- Back traceability (supplier traceability)

This work was financially supported by the General Secretariat for Research and Technology (GSRT) [16] of the Hellenic Ministry of Development in the framework of project SELIDA – contract #09SYN-72-646 which runs under the "Cooperation", 2009 Call.

- Internal traceability (process traceability)
- Forward traceability (end-user traceability)

Having the end-to-end traceability encompasses all three types of traceability and since traceability is defined over every stages of a value-chain, several researchers have pointed out various elements that should be taken into account.

Traceability systems store information and show the path of a particular object of interest along the whole value-chain from the supplier/producer to the retailer/distributor and eventually to the consumer/end-user. Throughout this process, secure, reliable and automatic object identification is crucial to provide effective and efficient tracing.

Barcode technology, in the past, has been used for the identification of items. However, in order to meet the traceability requirements imposed by the governments, a new technology that allows automated recording of information was needed. This need has been partially fulfilled by the revolutionary developments regarding the RFID technology.

Many logistic services have already integrated RFID identification technology into their services and products but these solutions most often implement a custom or proprietary communication flow. This means that it is quite difficult to come up with a generic approach against these solutions.

On the other hand, providing traceability services apart from trading logistics services i.e. physical documents interchange is even more demanding since existing services (e.g. Xerox Docushare, Papyrus, etc.) refer only to digitized documents management. Nowadays, document interchange between organizations, authorities and citizens is realized via the well-known courier shipping services (i.e. FedEx, etc.), but these services are almost always built into proprietary protocols while a gap often occurs when different services, even of the same type, in inter-continental transactions are involved.

Especially for book tracking (i.e. lending libraries, etc.) libraries and Inter-Librarian Loan (ILL) services in general employ standard interchange formats, exploited via web services, in order to share repositories and establish collaboration among them. However, a global standard elaborating libraries worldwide does not exist.

III. Architecture of the Deployed Information System

In the context of this section we are going to concisely present the architecture of the Document Tracking System (DoTS), which has been designed in order to tackle the issues mentioned in the previous section.

Section A concisely describes the technologies employed whereas section B presents a brief description of the entire architecture and its basic components.

A. Technologies and specifications employed

The following technologies and specifications have been utilized in the context of the system: **RFID**, used to uniquely identify physical objects and **EPC**, providing the underlying framework that the system takes advantage of in order to offer standardized tracking services.

a) RFID

The RFID (Radio Frequency IDentification) is a well-documented [1]-[6] and widely adopted [7], [8] technology, that provides the ability to uniquely identify objects tagged with RFID tags using special readers [9]. The main goal of the architecture is to be able to track documents on a potentially global scale. RFID provides significant advantages over other automatic ID technologies (specifically the widely applied bar and QR codes), as RFID tags:

- can be detected in bulk
- don't need to be aligned with the reader (line of sight) in order to be read
- can be detected from a greater distance
- have a larger data capacity
- are less susceptible to damage

These outweigh the benefits of the bar and QR code tags (which are less expensive and quite ubiquitous compared to RFID tags) for the application on important documents and can enable the introduction of innovative services, such as real time traceability and theft prevention. Additionally, RFID is intrinsic in the framework of global standards published by the GS1 that concern the EPC, which have been exploited in order to provide a method to globally provide tracking information services.

b) EPCglobal

The GS1 EPC global is a suite of standards and specifications that leverage the RFID technology in order to globally enable visibility and collaboration on an item level. These standards comprise the framework depicted in Fig. 1.

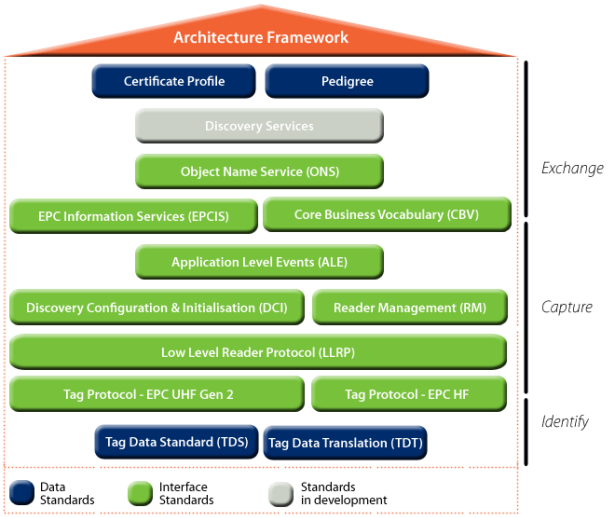


Fig. 1 GS1, EPCglobal framework standards, 2014 [6]

Information about these standards can be found in the GS1 website [5]. In the context of the proposed system, most of the EPC related GS1 standards have been utilized.

c) Object Name Service Specification

The discovery and tracking service of physical documents that has been implemented exploits the Object Name Service (ONS) 2.0.1 [5] and the EPC Information Services (EPCIS) 1.0.1 [2], in order to enable the mapping of EPC tagged documents to addresses of arbitrary, but with a standardized interface, object management services (OMS).

B. Framework Architecture

The proposed framework aims to support as many of the EPC global standards as possible; in order to provide the ability to map single physical objects to URIs and to track related information regarding the entire lifecycle of the representation for all involved organizations in the value chain, together with the realization of ONS-based web services available in the cloud.

The proposed architecture is value chain agnostic pertaining:

- common logistics value-chain (i.e. manufacturer, logistics service, retail and end-user/customer)
- physical documents inter-change value-chain between public authorities or organizations of the public sector and citizens or companies of the private sector
- objects inter-change value-chain in demanding cases such as insurance organizations, shipping companies, courier companies, etc.

Fig. 2 illustrates the incorporation of the aforementioned technologies in the proposed framework architecture.

The proposed ONS service layer consists of a collection of ONS-based web services that are able to provide information from the organization's internal hierarchy model breakdown, using the global EPC notation, without affecting the existing processes of the value chain. Each organization's ONS service layer is responsible for providing per object, both public and private information, using the global EPC notation.

The private web services satisfy per sector- needs for real-time, synchronous physical objects tracking. These needs have various orientations, depending on the corresponding node of the chain. The public web services address needs of common operations, regarding single object's lifecycle information, such as location, history etc.

The integration of the RFID subsystem, into the architecture framework raises two main issues. The first regards the capturing of tags' information as well as the identification of the tags themselves. The RFID reader scaling, range and reading angle are of major importance since it is incorporated in a versatile environment (many different types of organizations performing totally different internal processes). The second issue regards security and privacy issues that are raised when the identified object's data should be classified. Security and privacy issues are presented in detail in next paragraph.

a) RFID Middleware

The RFID middleware is responsible for receiving, analyzing processing and propagating the data collected by

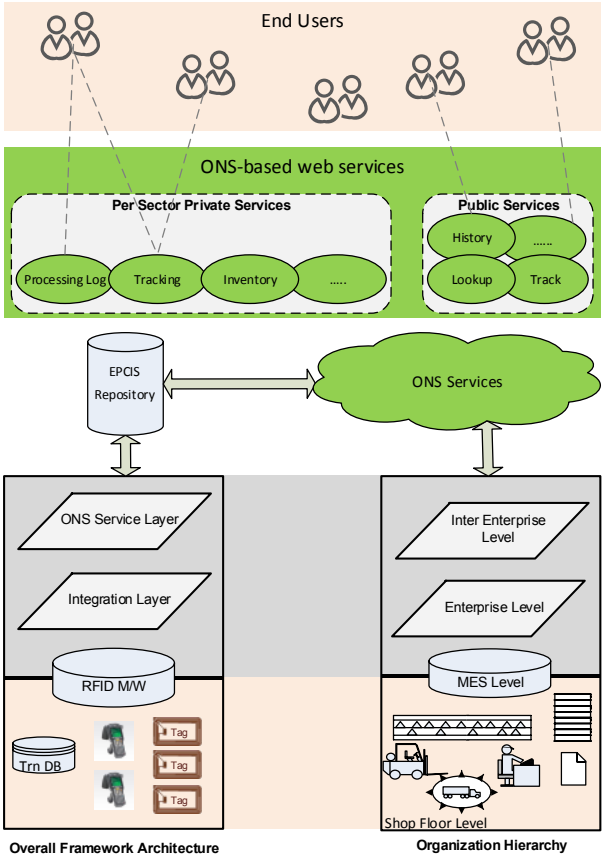


Fig. 2. Proposed framework architecture.

the RFID readers to the Information System that supports the business processes. The middleware hides the complexity of the actual RFID infrastructure and only provides business events. On the other hand, the middleware is oblivious of how the data it provides gets handled afterwards.

Specifically, the RFID middleware provides facilities for:

1. EPC Allocation
2. Device Management and Monitoring
3. Data Collection and Integration
4. Data Structure and Data Association
5. Data Filtering and Data Routing
6. Line Coordination and Process Control
7. Legend and Graphics Creation
8. Visibility and Reporting
9. Track and Trace Applications

b) ONS Resolver

The purpose of the ONS Resolver is to provide secure access to the ONS infrastructure, so that its clients would not only be able to query for the OMSs related to EPCs (which is the de facto use case of the ONS), but also introduce new or delete any existing OMSs for the objects. This has been accomplished by creating a SOAP web service layer that functions on top of the ONS, which provides authenticated and authorized users with the capability to query the whole

ONS infrastructure and discover the OMSs for the given EPCs, to add or delete OMSs or to add or delete users of the ONS Resolver, depending on their permissions.

In addition to the secure access to the ONS infrastructure, the ONS Resolver acts as an authorization server [10] for the relevant OMSs. This way, whenever a user uses the ONS Resolver to get the address of an OMS and authenticates successfully, an access token will be returned along with the result of the query, which, if used in the subsequent interaction between the user and the OMS, it can provide privileged access to the service. The authorization procedure that has been developed is depicted in Fig. 3.

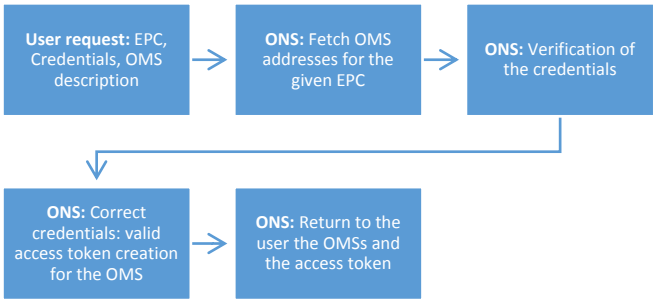


Fig. 3. ONS authorization procedure.

c) Object Management Services

The object management services (OMS) provide management, tracking and other value added services for the EPC tagged objects. The ONS Resolver maps the object management services to the objects according to their owner and type (for example the OMS for objects tagged using an SGTIN EPC is going to be determined by their GTIN; i.e. the OMS for the EPCs with the same GTIN is common [5]) and they should be implemented according to the EPCIS specification [2].

The most common case of the aforementioned context is the presentation of all historical data relevant to a single physical object. The resulting historical data can contain information regarding object transactions within locations controlled by the owner of the item or other organizations, as long as all the implicated parties implement the described framework; this was rendered possible through the utilization of the ONS Service Layer. The data of the objects stored by the different parties adhere to the same API and, as a result, one can select the exact kind of these data that are to be presented. For example historical data could describe an object's transaction only with dates on interaction with an RFID M/W, or with geographic interpreted location of the interaction, or even with more meta-data of the object like its purpose of transportation, original owner, final destination and whether it is classified or not. In an expanded version of multiple organizations running the presented framework, any tagged object (document, package, suitcase, etc.) could be easily and reliable recognized in the entire framework's context.

Finally, the arbitrary nature of the OMS themselves, even though they are implemented with standardized formation, enables each organization to level-up objects' related information according to the specific organization needs and requirements.

d) Security and Privacy aspects

Since the proposed architecture is based on web services, our first goal is to identify and classify those services in regards to their security requirements. Even before that, we can safely assume that all web services can and should implement TLS as a standard form of encrypting the data channel in use (usually the Internet).

Based on the aforementioned architecture diagram, we can easily identify that there are public and non-public facing web services. In regards to the non-public facing web services, the ones that reside within the internals of the proposed architecture stack, we can exploit the useful fact that both the clients and the servers of this part of the architecture are known and can be controlled in regards to their implementation of the security mechanisms. Therefore it is safe to assume that the use of client certificates is a feasible security mechanism. Client certificates are a very robust way of handling secure and, in conjunction with TLS, encrypted authentication and authorization and the issues with scalability and deployment that are usually encountered in more general scenarios are not applicable to our proposed architecture. As for the public facing web services, we follow the industry standard of API keys due to the fact that although we name those services "public facing", in reality those services will be accessed not by casual end users but by the information systems of the organizations that will employ that services of our proposed architecture. A case study of such a deployment is described in the next chapter.

Moreover, all the web services should follow a standard of secure design that, although already a common practice in popular web services around the web, we will briefly describe below.

As mentioned above, all services should be authenticated over an encrypted communication channel. Messages should be digitally signed, as well as encrypted, to provide privacy and tamper-proofing when the message travels through intermediary nodes route to the final destination even within different organizations that implement the same proposed architectural stack. The usage of the access token (a unique ID or nonce, a cryptographically unique value) generated by the ONS Resolver within every request, will, obviously, provide protection against unauthorized usage and it will also aid to the detection message replay and man in the middle attacks. HTTP methods should be valid for each API key and associated resource collection and method by white-listing allowable methods. Any request for exposed resources should be protected against CSRF and insecure direct object references should be avoided.

IV. Case Study on Physical Documents Tracking in a Library Environment

In order to evaluate the applicability of the proposed architecture on a real environment we deployed it on the existing Integrated Library System (ILS) that is being used in the central university library of Patras, Greece. The ILS that is being used by the institution is the well-known open source ILS named KOHA [11].

As with all ILS, KOHA supports a variety of workflows and services to accommodate the needs of the institution. Our proposed scheme focuses on a handful of those services and augments them with additional features. This is generally done by adding, in a transparent way, the additional UI elements and background processes that are needed for our scheme to work.

The specifics on how this is done will be presented in the following section, while first we will discuss briefly on the exact services of the KOHA ILS that our scheme aims to augment.

A. Supported services

In its initial design, our scheme aims to provide additional functionality on the core services of an ILS. Those services are:

- Check Out
- Check In
- New Record
- Delete Record

There are also a number of tracking services that our scheme aims for and those are:

- History
- Location
- Search

To elaborate, when the check-out or check-in services are called in KOHA, an additional call is made on the SELIDA (the name of the developed framework from the related project) middleware that updates the status of the affected documents on the SELIDA database. The details on how those calls are made will be described in the next section. Similar functionality can be seen on the entire core and tracking services that are applicable in our scheme as described in the previous chapters.

B. Integration layer

In order to provide the added functionality to the existing KOHA services we designed and implemented an integration layer that seamlessly handles all the extra work along with the usual service workflow.

The primary reason to provide a seamless layer instead of changing the actual services (i.e. the source code) is the fact that every integrated system that is actually in a production environment needs the benefits of a continuous and stable update process that is offered by the systems development team along and implemented by the organization's administrator. Adding extra functionality in the form of changes to the system's code would require continuous maintenance of this part of the system on par with the normal updates of the KOHA ILS.

To overcome this obstacle, and given to the fact that KOHA is a web based application as most of the modern ILS, we designed the integration layer so that it is injected upon page load as a JavaScript file on the pages that we are interested in. On our specific case study, we used the "mod_substitute" directive on the Apache web server that was serving the KOHA web pages. Each time a module/page of interest is requested by the server (i.e. check-out), the web server adds a <script> tag that loads all the additional functionality in the form of a JavaScript file.

This layer, in the form of a JavaScript application module, adds the required UI elements for our proposed scheme to work and handles all the web service requests that are necessary.

As an example, we will showcase the check-out process. When the user navigates to the check-out KOHA module by requesting the module's URL, the apache server injects the selida.js file which is the integration layer code (so called SELIDA module). SELIDA module starts executing upon page load and adds the button "Scan" next to the button "Checkout". When the user presses the button "Scan", a web service request is launched from the SELIDA module which starts up the RFID reader via the SELIDA middleware services. The results (per example the barcodes and titles of the documents that the reader identified) are communicated back to the SELIDA module which in turn shows a pop-up window to the user indicating those results. After this procedure, the normal check-out workflow resumes by sending the required POST requests to the KOHA web server (just like when the user presses the "Checkout" button after adding the barcode). When the check-out process ends and the web server responds with the next web page as per KOHA workflow, the SELIDA module sends a second web service request to the middleware indicating that the check-out is complete so that the rest of SELIDA architecture continues with its own check-out workflow presented in the previous chapters.

As we can see on this example, the extra steps that are needed for our scheme to work are completely transparent from the ILS itself. KOHA as an application retains its original functionality and workflow while the user benefits from the added services that our SELIDA module provides.

Fig. 5 illustrates the framework's activities during the first object identification phase, previously described.

C. Object Management Services

Since RESTful [12] web services are becoming the most common developers' choice for implementing API's and data retrieval services SELIDA's services have been designed following the RESTful architectural style and JavaScript Object Notation (JSON) [13] for data exchanging. Although these technologies do not comply with a specific standard, as other technologies do, i.e. SOAP, XML web services, when combined together they constitute a lightweight data-interchange mechanism that is easy for humans to read and

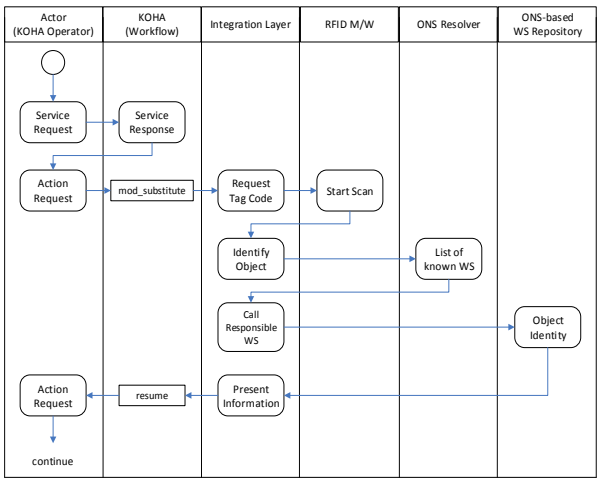
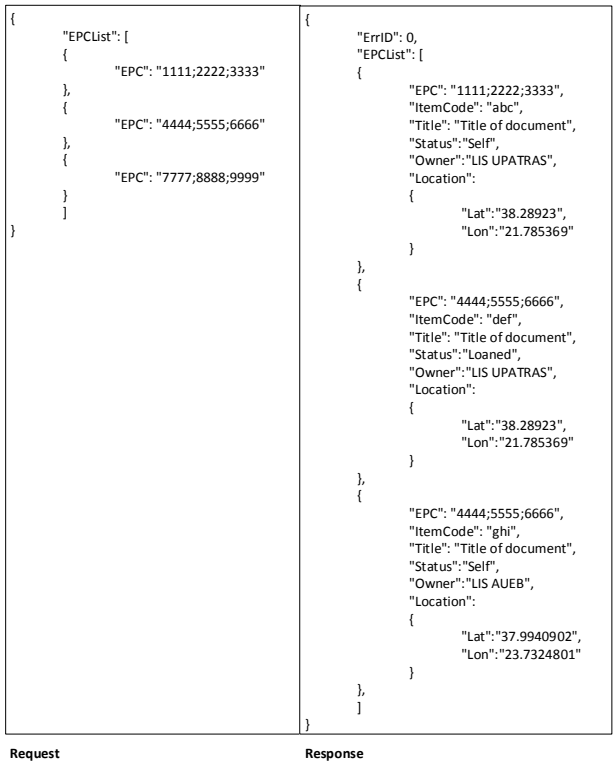


Fig. 5. Framework's activities during an object identification phase.

write and also easy for machines to parse and generate, thus appearing to be the easiest and most comprehensive messaging way among web services. Moreover, nowadays numerous open standards-like specifications (i.e. JSONDoc [14], Swagger [15], etc.) have been presented towards describing, producing, consuming, and visualizing RESTful web services so as RESTful web services can become a complete framework.

Currently, based on the deployment of the proposed architecture on the existing Integrated Library System (ILS) that is being used in the central university library of Patras, Greece, three types of management services (history, location and search) have been implemented, as described in paragraph A.



Request

```
{
        "EPCList": [
        {
                "EPC": "1111;2222;3333"
        },
        {
                "EPC": "4444;5555;6666"
        },
        {
                "EPC": "7777;8888;9999"
        }
        ]
}
```

Response

```
{
        "ErrID": 0,
        "EPCList": [
        {
                "EPC": "1111;2222;3333",
                "ItemCode": "abc",
                "Title": "Title of document",
                "Status":"Self",
                "Owner":"LIS UPATRAS",
                "Location":
                {
                        "Lat":"38.28923",
                        "Lon":"21.785369"
                }
        },
        {
                "EPC": "4444;5555;6666",
                "ItemCode": "def",
                "Title": "Title of document",
                "Status":"Loaned",
                "Owner":"LIS UPATRAS",
                "Location":
                {
                        "Lat":"38.28923",
                        "Lon":"21.785369"
                }
        },
        {
                "EPC": "4444;5555;6666",
                "ItemCode": "ghi",
                "Title": "Title of document",
                "Status":"Self",
                "Owner":"LIS AUEB",
                "Location":
                {
                        "Lat":"37.9940902",
                        "Lon":"23.7324801"
                }
        },
        ]
}
```

Fig. 4. An example of JSON request and response data structure for status OMS.

As for an example, the JSON schemes for the request and response of the search service is presented in Fig. 4.

As depicted in Fig. 4 request to the status OMS may contain multiple EPC tags at a time, while the implemented system response contains an indicative error code (0 in case of no error) and relevant data for recognized books. In the illustrated example the first two EPCs belong to books owned by the central university library of Patras, Greece (LIS UPATRAS) while the third belongs to a book owned by the university library of Athens University of Economics and Business, Athens, Greece (LIS AUEB).

V. Benefits and Scalability

Although the described framework has been deployed in an ILL (Inter-Librarian Loan) environment its initial inspiration and design has been originated in general logistics or warehouse inventory stocktaking environments. Based on this origin the whole framework has been designed and deployed so as to offer organization agnostic information, supported by ONS-based web-services that are integrated at the top of each organization's hierarchy stack, as illustrated in Fig. 2.

The proposed architecture offers a set of versatile characteristics due to the combination of reliability and uniqueness, induced by RFID technology along with the ONS perspective implementation throughout the architecture that derives interoperable information exchange. These characteristics could constitute the basis for the employment of such a framework so as to derive into a generic Internet-Of-Things service platform able to handle information and processes of any value-chain type, including and not limited at food supply chain, luggage handling, physical document interchange, etc.

This prospect could be further substantiated by the nature of the ONS, which is actually a mechanism that leverages Domain Name System (DNS) to discover information about an object and its related services from the EPC code. Conclusively, the presented architecture could be scaled in the same way as internet does, since internet is based on DNS and the presented architecture is based on ONS, inheriting DNS scalability capabilities.

VI. Conclusion

In spite of the promising nature of RFID technology, numerous applications in the actual logistics field area have not been reported. Only a few pilot studies as well as experimental tests have proved that RFID would be a successful tool to enable supply chain traceability. The reasons why companies are yet reluctant to have confidence in adopting the technology to gain their product visibility may be attributed to the several challenges such as lack of standards, immaturity of RFID, and privacy issues.

This paper presents a ubiquitous approach towards a collaborative logistics environment and concludes with a case study implementation involving physical documents tracking

in an academic library. The main focus of the proposed methodology and implemented architecture addresses the issue of empowering the whole framework with a standard specification for objects tracking services, thus the integration of the ONS-perspective in the architecture. The integration itself, which, within the EPC global framework is mainly achieved by utilizing the Object Naming Service, enables involved organizations to act agnostically of their entities and provides them with the ability to resolve EPC tagged objects to arbitrary services, but with a standardized manner.

However, security and privacy issues should be further investigated as future work, apart from the issues already covered though the implementations of the framework's web services, so as the presented framework will be promising enough for evolutionizing the way currently exploited for tracking items in the supply chain.

REFERENCES

[1] ISO. (2010). ISO RFID Standards: A Complete List. Retrieved from http://rfid.net/basics/186-iso-rfid-standards-a-complete-list
[2] GS1. (2007). EPC Information Services (EPCIS) Version 1.0.1 Specification. Retrieved from http://www.gs1.org/gsmp/kc/epcglobal/epcis/epcis_1_0_1-standard-20070921.pdf
[3] GS1. (2010). EPCglobal: Specifications. Retrieved from http://www.gs1.org/epcglobal/standards/specs
[4] GS1. (2010). EPCglobal: Standards. Retrieved from http://www.gs1.org/gsmp/kc/epcglobal
[5] GS1. (2013). Object Name Service (ONS). Retrieved from http://www.gs1.org/gsmp/kc/epcglobal/ons/ons_2_0_1-standard-20130131.pdf
[6] GS1. (2014). EPCglobal framework standards. Retrieved from http://www.gs1.org/gsmp/kc/epcglobal
[7] Hossain, M. (2012). A Comparison of Voluntary and Mandatory Adoption of Radio Frequency Identification (RFID) Technology in Organizations. Retrieved from http://www.wbiconpro.com/421-Hossain.pdf
[8] RFIDjournal. (2013). Survey Shows Half of U.S. Retailers Have Already Adopted Item-Level RFID. Retrieved from http://www.rfidjournal.com/articles/view?9168
[9] Wikipedia. (2007). Radio-frequency identification. Retrieved from http://en.wikipedia.org/wiki/Radio-frequency_identification
[10] IETF. (2012). RFC 6749: The OAuth 2.0 Authorization Framework. Retrieved from http://tools.ietf.org/html/rfc6749
[11] Official Website of Koha Library Software. Retrieved from http://koha-community.org/
[12] Wikipedia (2014). Representational state transfer. Retrieved from http://en.wikipedia.org/wiki/Representational_state_transfer
[13] JSON (JavaScript Object Notation). Retrieved from http://json.org
[14] JSONDoc. Retrieved from http://jsondoc.org
[15] Swagger. Retrieved from https://helloreverb.com/developers/swagger
[16] General Secretariat for Research and Technology, http://www.gsrt.gr

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1345–1352
DOI: 10.15439/2014F89
ACSIS, Vol. 2

A hybrid CP/MP approach to supply chain modelling, optimization and analysis

Paweł Sitek
Kielce University of Technology Al. 1000-lecia PP 7, 25-314 Kielce,
Poland Institute of Management Control Systems
e-mail:sitek@tu.kielce.pl

***Abstract*—The paper presents a concept and implementation of a novel hybrid approach to the modelling, optimization and analysis of the supply chain problems. Two environments, mathematical programming (MP) and constraint programming (CP), in which constraints are treated in different ways and different methods are implemented, were combined to use the strengths of both.**

This integration and hybridization, complemented with an adequate transformation of the problem, facilitates a significant reduction of the combinatorial problem. The whole process takes place at the implementation layer, which makes it possible to use the structure of the problem being solved, implementation environments and the very data. The superiority of the proposed approach over the classical scheme is proved by considerably shorter search time and example-illustrated wide-ranging possibility of expanding the decision and/or optimization models through the introduction of new logical constraints, frequently encountered in practice. The proposed approach is particularly important for the decision models with an objective function and many discrete decision variables added up in multiple constraints.

The presented approach will be compared with classical mathematical programming on the same data sets.

I. Introduction

The supply chain is commonly seen as a collection of various types of companies (raw materials, production, trade, logistics, transport, etc.) working together to improve the flow of products, information and finance. As the words in the term indicate, the supply chain is a combination of its individual links in the process of supplying products (material/products and services) to the market.

Huang [1] studied the shared information of supply chain production. This consists in the information shared between each network node determined by the model, which enables production, distribution and transport planning dependent on the purpose. The shared information process is vital for effective supply chain production, distribution and transport planning. In terms of centralized planning, the information flows from each node of the network where the decisions are made. Shared information includes the following groups of parameters: resources, inventory, production, transport, demand, etc. Minimization of total costs is the main purpose of the models presented in the literature, while maximization of revenues or sales is considered to a smaller scale [12].

The vast majority of the works reviewed [2]–[7], [9],[10],[12] have formulated their models as linear programming (LP), integer programming (IP) and mixed integer linear programming (MILP) problems and solved them using the Operations Research methods. Nonlinear programming, multi-objective programming, fuzzy programming with stochastic programming are used much less frequently [12] [25].

Problems related to the design, integration and management of the supply chain affect many aspects of production, distribution, warehouse management, supply chain structure, transport modes etc. Those problems are usually closely related to each other, some may influence one another to a greater or lesser extent. Because of the interconnectedness and a very large number of different constraints: resource, time, technological, and financial, the constraint–based environments are suitable for producing "natural" solutions for highly combinatorial problems. In the literature, references to modeling and optimizing supply chain problems using constraint–based environments are relatively few in number [11], [12].

This paper deals with a problem of supply chain modelling, optimization and analysis. An important contribution of the presented hybrid CP/MP approach is to propose a hybrid implementation platform that supports the modelling, optimization and analysis of decision problems in the supply chain. In this platform two environments, mathematical programming (MP) and constraint programming (CP), in which constraints are treated in different ways and different methods are implemented, were combined to use the strengths of both in the presented platform.

The rest of the paper is organized as follows: Section II describes our motivation and analyses the state of the art in this domain. Section III gives the concept of the novel hybrid CP/MP approach and implementation platform. The optimization model as an illustrative example is described in Section IV. Computational examples and tests of the implemented model are presented in Section V. The discussion on possible extensions of the proposed approach and conclusions is included in Section VI.

978-83-60810-58-3/$25.00

II. MOTIVATION

We strongly believe that the constraint-based environment [13], [14], [16], [19] offers a very good framework for representing the knowledge and information needed for the decision support. The central issue for a constraint-based environment is a constraint satisfaction problem (CSP) [13]. Constraint satisfaction problem is the mathematical problem defined as a set of elements whose state must satisfy a number of constraints. Constraint satisfaction problems (CSPs) on finite domains are typically solved using a form of search. The most widely used techniques include variants of backtracking, constraint propagation, and local search. Constraint propagation embeds any reasoning that consists in explicitly forbidding values or combinations of values for some variables of a problem because a given subset of its constraints cannot be satisfied otherwise [16]. CSPs are frequently used in constraint programming. Constraint programming is the use of constraints as a programming language to encode and solve problems. Constraint logic programming (CLP) is a form of constraint programming (CP), in which logic programming is extended to include concepts from constraint satisfaction. A constraint logic program is a logic program that contains constraints in the body of clauses. Constraints can also be present in the goal. These environments are declarative. The declarative approach and the use of logic programming provide incomparably greater possibilities for decision problems modelling than the pervasive approach based on mathematical programming. Unfortunately, discrete optimization is not a strong suit of these environments.

Based on [8], [15], [16] and previous work [14], [17], [18], we observed some advantages and disadvantages of these environments. An integrated approach of constraint programming (CP) and mathematical programming (MP) can help to solve optimization problems that are intractable with either of the two methods alone [20]–[23]. Although mathematical programming and constraint programming have different roots, the links between the two environments have grown stronger in recent years.

Both MP and finite domain CP/CLP involve variables and constraints. However, the types of the variables and constraints that are used, and the way the constraints are solved, are different in the two approaches [23].

In both MILP and CP/CLP, there is a group of constraints that can be solved with ease and a group of constraints that are difficult to solve. The easily solved constraints in MILP are linear equations and inequalities over rational numbers.

Integrity constraints are difficult to solve using mathematical programming methods and often the real problems of MILP make them NP-hard.

In CP/CLP, domain constraints with integers are easy to solve. The system of such constraints can be solved over integer variables in polynomial time. The inequalities between more than two variables, general linear constraints and symbolic constraints are difficult to solve, which makes real problems in CP/CLP NP-hard. This type of constraints reduces the strength of constraint propagation. As a result, CP/CLP is incapable of finding even the first feasible solution. This is the greatest weakness of this approach.

As mentioned earlier, the vast majority of decision-making models for the problems of production, logistics, supply chain are formulated in the form of mathematical programming (MIP, MILP, IP).

Due to the structure of these models (adding together discrete decision variables in the constraints and the objective function) and a large number of discrete decision variables (integer and binary), they can only be applied to small problems. Another weakness is that only linear constraints can be used. In practice, the issues related to the production, distribution and supply chain constraints are often logical, nonlinear, etc. For these reasons the problem was formulated in a new way

The motivation and contribution behind this work was to create a hybrid method for supply chain decision problems modelling and optimization instead of using mathematical programming or constraint programming separately. It follows from the above that what is difficult to solve in one environment can be easy to solve in the other. Furthermore, such a hybrid CP/MP approach allows the use of all layers of the problem to solve it (Fig. 1). And finally, the transformation of the problem to a form that can fully exploit the strengths of the constraint propagation.

The hybrid method is not inferior to its component elements applied separately. This is due to the fact that the number of decision variables and the search area are reduced. The extent of the reduction directly affects the effectiveness of the method.

III. THE CONCEPT OF THE CP/MP HYBRID APPROACH

Due to the structure of the decision models for supply chain problems (summing of discrete decision variables in the constraints and the objective function) and a large number of discrete decision variables (integer and/or binary) they can only be applied to small problems. Another disadvantage is that only linear constraints can be used. In practice, the issues related to the production, distribution and su25]5ply chain constraints are often logical, nonlinear, etc. For these reasons the problem was formulated in a new way.

In our approach to modeling and optimization these problems we proposed the implementation platform, where:

- knowledge related to the supply chain can be expressed as linear and logical constraints (implementing all types of constraints of the previous MILP models [17], and introducing new types of constraints (logical, nonlinear, symbolic etc.));
- the decision models solved using the proposed platform can be formulated as a pure model of MILP or of CP/CLP, or it can also be a hybrid model with logical and other types of constraints;
- the problem is modelled in CP/CLP, which is far more flexible than MILP;

- the possibility to transform the problem of using the flexibility of declarative environment (CP/CLP) is introduced;
- the novel method of constraint propagation is proposed (obtained by transforming the decision model to explore its structure and properties);
- constrained domains of decision variables, new constraints and values for some variables are transferred from CP-based environment into MP-based environment;
- the efficiency of finding solutions to larger size problems is increased.

The concept of the proposed implementation platform is presented in Fig. 1. In the first stage, a formal model is implemented in the form of predicates in CLP and the data in the form of facts. In the next step constraint propagation is performed. Constraint propagation is one of the basic methods of CLP. As a result, the variable domains are narrowed, and in some cases, the values of variables are set, or even the solution can be found. In order to increase the efficiency of the constraint propagation transformation of the problem and its representation can be made. The transformation uses the structure and properties of the problem. The most common effect is a change in the representation of the problem by reducing the number of decision variables, and the introduction of additional constraints and variables, changing the nature of the variables, etc.

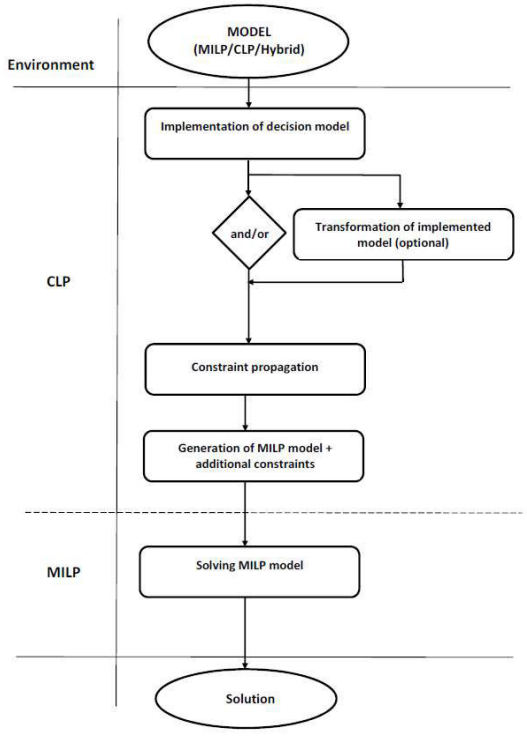


Fig. 1 Concept scheme of the CP/MP hybrid approach

The next step is the generation of the MILP model using predicates in CLP. All the information obtained in previous stages are used during the generation of the model. The final step is to solve the model by the solver MILP.

The implementation details of the CP/MP hybrid approach have been discussed in [24]. The motivation was to offer the most effective tools for model–specific constraints and solution efficiency.

IV. EXAMPLES OF SUPPLY CHAIN OPTIMIZATION

The proposed approach was used and tested on two supply chain optimization models.

First model was formulated as a mixed linear integer programming (MILP) problem [18] under constraints (2) .. (23) in order to test the proposed approach (Fig. 1) against the classical integer programming approach [17]. Then the hybrid model (1) .. (26) was implemented and solved. Indices, parameters and decision variables used in the models together with their descriptions are summarized in Table I. The simplified structure of the supply chain network for this model, composed of producers, distributors and customers is presented in Fig. 2.

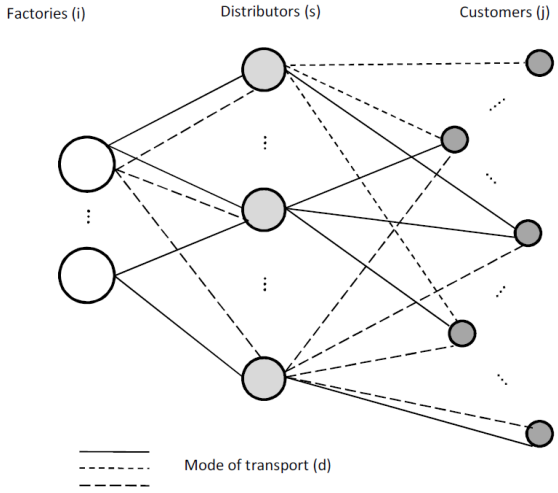


Fig. 2 The simplified structure of the supply chain network

Both models are the cost models that take into account three other types of parameters, i.e., the spatial parameters (area/volume occupied by the product, distributor capacity and capacity of transport unit), time (duration of delivery and service by distributor, etc.) and the transport mode.

The main assumptions made in the construction of these models were as follows:

- the shared information process in the supply chain consists of resources (capacity, versatility, costs), inventory (capacity, versatility, costs, time), production (capacity, versatility, costs), product (volume), transport (cost, mode, time), demand, etc;
- part of the supply chain has a structure as in Fig. 2.;
- transport is multimodal (several modes of transport, a limited number of means of transport for each mode);
- the environmental aspects of use of transport modes are taken into account;
- different products are combined in one batch of transport;
- the cost of supplies is presented in the form of a function (in this approach, linear function of fixed and variable costs);

- the models have linear or linear and logical constraints.

TABLE I.
INDICES, PARAMETERS AND DECISION VARIABLES

Symbol	Description
Indices	
k	product type (k=1..O)
j	delivery point/customer/city (j=1..M)
i	manufacturer/factory (i=1..N)
s	distributor /distribution center (s=1..E)
d	mode of transport (d=1..L)
N	number of manufacturers/factories
M	number of delivery points/customers
E	number of distributors
O	number of product types
L	number of mode of transport
Input parameters	
F_s	the fixed cost of distributor/distribution center *s*
P_k	the area/volume occupied by product *k*
V_s	distributor *s* maximum capacity/volume
$W_{i,k}$	production capacity at factory *i* for product *k*
$C_{i,k}$	the cost of product *k* at factory *i*
$R_{s,k}$	if distributor *s* can deliver product *k* then $R_{s,k}$=1, otherwise $R_{s,k}$=0
$Tp_{s,k}$	the time needed for distributor *s* to prepare the shipment of product *k*
$Tc_{j,k}$	the cut-off time of delivery to the delivery point/customer *j* of product *k*
$Z_{j,k}$	customer demand/order *j* for product *k*
Zt_d	the number of transport units using mode of transport *d*
Pt_d	the capacity of transport unit using mode of transport *d*
$Tf_{i,s,d}$	the time of delivery from manufacturer *i* to distributor *s* using mode of transport *d*
$K1_{i,s,k,d}$	the variable cost of delivery of product *k* from manufacturer *i* to distributor *s* using mode of transport *d*
$R1_{i,s,d}$	if manufacturer *i* can deliver to distributor *s* using mode of transport *d* then $R1_{i,s,d}$=1, otherwise $R1_{i,s,d}$=0
$A_{i,s,d}$	the fixed cost of delivery from manufacturer *i* to distributor *s* using mode of transport *d*
$Koa_{s,j,d}$	the total cost of delivery from distributor *s* to customer *j* using mode of transport *d*
$Tm_{s,j,d}$	the time of delivery from distributor *s* to customer *j* using mode of transport *d*
$K2_{s,j,k,d}$	the variable cost of delivery of product *k* from distributor *s* to customer *j* using mode of transport *d*
$R2_{s,j,d}$	if distributor *s* can deliver to customer *j* using mode of transport *d* then $R2_{s,j,d}$=1, otherwise $R2_{s,j,d}$=0
$G_{s,j,d}$	the fixed cost of delivery from distributor *s* to customer *j* using mode of transport *d*
$Kog_{s,j,d}$	the total cost of delivery from distributor *s* to customer *j* using mode of transport *d*
Od_d	the environmental cost of using mode of transport *d*
Decision variables	
$X_{i,s,k,d}$	delivery quantity of product *k* from manufacturer *i* to distributor *s* using mode of transport *d*
$Xa_{i,s,d}$	if delivery is from manufacturer *i* to distributor *s* using mode of transport *d* then $Xa_{i,s,d}$=1, otherwise $Xa_{i,s,d}$=0
$Xb_{i,s,d}$	the number of courses from manufacturer *i* to distributor *s* using mode of transport *d*
$Y_{s,j,k,d}$	delivery quantity of product *k* from distributor *s* to customer *j* using mode of transport *d*
$Ya_{s,j,d}$	if delivery is from distributor *s* to customer *j* using mode of transport *d* then $Ya_{s,j,d}$ =1, otherwise $Ya_{s,j,d}$ =0
$Yb_{s,j,d}$	the number of courses from distributor *s* to customer *j* using mode of transport *d*
Tc_s	if distributor *s* participates in deliveries, then Tc_s=1, otherwise Tc_s=0
CW	Arbitrarily large constant

A. Objective function

The objective function (1) defines the aggregate costs of the entire chain and consists of five elements. The first element comprises the fixed costs associated with the operation of the distributor involved in the delivery (e.g. distribution centre, warehouse, etc.). The second element corresponds to environmental costs of using various means of transport. Those costs are dependent on the number of courses of the given means of transport, and on the other hand, on the environmental levy, which in turn may depend on the use of fossil fuels and carbon-dioxide emissions [26],[27].

The third component determines the cost of the delivery from the manufacturer to the distributor. Another component is responsible for the costs of the delivery from the distributor to the end user (the store, the individual client, etc.). The last component of the objective function determines the cost of manufacturing the product by the given manufacturer.

Formulating the objective function in this manner allows comprehensive cost optimization of various aspects of supply chain management. Each subset of the objective function with the same constrains provides a subset of the optimization area and makes it much easier to search for a solution.

$$\sum_{s=1}^{E} F_s \cdot Tc_s + \sum_{d=1}^{L} Od_d \left(\sum_{i=1}^{N}\sum_{s=1}^{E} Xb_{i,s,d} + \sum_{s=1}^{E}\sum_{j=1}^{M} Yb_{j,s,d}\right) + \sum_{i=1}^{N}\sum_{s=1}^{E}\sum_{d=1}^{L} Koa_{i,s,d} + \sum_{s=1}^{E}\sum_{j=1}^{M}\sum_{d=1}^{L} Kog_{s,j,d} + \sum_{i=1}^{N}\sum_{k=1}^{O}\left(C_{ik} \cdot \sum_{s=1}^{E}\sum_{d=1}^{L} X_{i,s,k,d}\right) \quad (1)$$

B. Constraints

The model was based on constraints (2) .. (26) Constraint (2) specifies that all deliveries of product *k* produced by the manufacturer *i* and delivered to all distributors *s* using mode of transport *d* do not exceed the manufacturer's production capacity.

Constraint (3) covers all customer *j* demands for product *k* ($Z_{j,k}$) through the implementation of delivery by distributors *s* (the values of decision variables $Y_{i,s,k,d}$). The flow balance of each distributor *s* corresponds to constraint (4). The possibility of delivery is dependent on the distributor's technical capabilities – constraint (5). Time constraint (6) ensures the terms of delivery are met. Constraints (7a), (7b), (8) guarantee deliveries with available transport taken into account. Constraints (9), (10), (11) set values of decision variables based on binary variables Tc_s, $Xa_{i,s,d}$, $Ya_{s,j,d}$. Dependencies (12) and (13) represent the relationship based on which total costs are calculated. In general, these may be any linear functions. The remaining constraints (14)..(23) arise from the nature of the model (MILP).

Constraint (24) allows the distribution of exclusively one of the two selected products in the distribution center *s*. Similarly, constraint (25) allows the production of exclusively one of the two selected products in the factory *i*.

Constraint (26) allows the transport of exclusively one of the two selected products in the same route and transport unit.

Those constraints result from technological, marketing, sales or safety reasons. Therefore, some products cannot be distributed and/or produced and/or transported together. The constraint can be re-used for different pairs of product *k* and for some of or all distribution centers *s* and factories *i.* A logical constraint like this cannot be easily implemented in a mathematical programming model. Only declarative application environments based on constraint satisfaction problem (CSP) make it possible to easy implement constraints such as (24), (25), (26).

The addition of constraints of that type changes the model class. It is a hybrid model (1)..(26).

$$\sum_{s=1}^{E}\sum_{d=1}^{L} X_{i,s,k,d} \cdot R_{s,k} \le W_{i,k} \text{ for } i=1..N, k=1..O \quad (2)$$

$$\sum_{s=1}^{E}\sum_{d=1}^{L} (Y_{s,j,k,d} \cdot R_{s,k}) \ge Z_{j,k} \text{ for } j=1..M, k=1..O \quad (3)$$

$$\sum_{i=1}^{N}\sum_{d=1}^{L} X_{i,s,k,d} = \sum_{j=1}^{M}\sum_{d=1}^{L} Y_{s,j,k,d} \text{ for } s=1..E, k=1..O \quad (4)$$

$$\sum_{k=1}^{O} (P_k \cdot \sum_{i=1}^{N}\sum_{d=1}^{L} X_{i,s,k,d}) \le Tc_s \cdot V_s \text{ for } s=1..E \quad (5)$$

$$Xa_{i,s,d} \cdot Tf_{i,s,d} + Xa_{i,s,d} \cdot Tp_{s,k} + Ya_{s,j,d} \cdot Tm_{s,j,d} \le Tc_{j,k} \text{ for } i=1..N,\ s=1..E,\ j=1..M,\ k=1..O, d=1..L \quad (6)$$

$$R1_{i,s,d} \cdot Xb_{i,s,d} \cdot Pt_d \ge X_{i,s,k,d} \cdot P_k \text{ for } i=1..N, s=1..E, k=1..O, d=1..L \quad (7a)$$

$$R2_{s,j,d} \cdot Yb_{s,j,d} \cdot Pt_d \ge Y_{s,j,k,d} \cdot P_k \text{ for } s=1..E,\ j=1..M,\ k=1..O, d=1..L \quad (7b)$$

$$\sum_{i=1}^{N}\sum_{s=1}^{E} Xb_{i,s,d} + \sum_{j=1}^{M}\sum_{s=1}^{E} Yb_{j,s,d} \le Zt_d \text{ for } d=1..L \quad (8)$$

$$\sum_{i=1}^{N}\sum_{d=1}^{L} Xb_{i,s,d} \le CW \cdot Tc_s \text{ for } s=1..E \quad (9)$$

$$Xb_{i,s,d} \le CW \cdot Xa_{i,s,d} \text{ for } i=1..N, s=1..E, d=1..L \quad (10)$$

$$Yb_{s,j,d} \le CW \cdot Ya_{s,j,d} \text{ for } s=1..E, j=1..M, d=1..L \quad (11)$$

$$Koa_{i,s,d} = A_{i,s,d} \cdot Xb_{i,s,d} + \sum_{k=1}^{O} K1_{i,s,k,d} \cdot X_{i,s,k,d} \text{ for } i=1..N, s=1..E, d=1..L \quad (12)$$

$$Kog_{s,j,d} = G_{s,j,d} \cdot Yb_{j,s,d} + \sum_{k=1}^{O} K2_{s,j,k,d} \cdot Y_{s,j,k,d} \text{ for } s=1..E, j=1..M, d=1..L \quad (13)$$

$$X_{i,s,k,d} \ge 0 \text{ for } i=1..N, s=1..E, k=1..0,\ d=1..L \quad (14)$$

$$Xb_{i,s,d} \ge 0 \text{ for } i=1..N,\ s=1..E,\ d=1..L, \quad (15)$$

$$Yb_{s,j,d} \ge 0 \text{ for } s=1..E,\ j=2..M,\ d=1..L, \quad (16)$$

$$X_{i,s,k,d} \in C \text{ for } i=1..N, s=1..E, k=1..0, d=1..L, \quad (17)$$

$$Xb_{i,s,d} \in C \text{ for } i=1..N,\ s=1..E,\ d=1..L \quad (18)$$

$$Y_{s,j,k,d} \in C \text{ for } s=1..E, j=1..M,\ k=1..0, d=1..L \quad (19)$$

$$Yb_{s,j,d} \in C \text{ for } s=1..E,\ j=1..M,\ d=1..L, \quad (20)$$

$$Xa_{i,s,d} \in \{0,1\} \text{ for } i=1..N,\ s=1..E,\ d=1..L, \quad (21)$$

$$Ya_{s,j,d} \in \{0,1\} \text{ for } s=1..E,\ j=2..M,\ d=1..L, \quad (22)$$

$$Tc_s \in \{0,1\} \text{ for } s=1..E \quad (23)$$

$$ExclusionD(k_1,k_2,s) \text{ for } k_1,k_2 \in 1..O,\ s \in 1..E,\ k_1 \ne k_2 \quad (24)$$

$$ExclusionP(k_1,k_2,i) \text{ for } k_1,k_2 \in 1..O,\ i \in 1..N,\ k_1 \ne k_2 \quad (25)$$

$$ExclusionT(k_1,k_2) \text{ for } k_1 \in 1..O,\ k_2 \in 1..O,\ k_1 \ne k_2 \quad (26)$$

C. Model transformation

Due to the nature of the decision problem (adding up decision variables and constraints involving a lot of variables), the constraint propagation efficiency decreases dramatically. Constraint propagation is one of the most important methods in CLP affecting the efficiency and effectiveness of the CLP and novel hybrid implementation platform (Fig. 1). For that reason, research into more efficient and more effective methods of constraint propagation was conducted. The results included different representation of the problem and the manner of its implementation. The classical problem modeling in the CLP environment consists in building a set of predicates with parameters. Each CLP predicate has a corresponding multi-dimensional vector representation. While modeling both problems, (1) .. (23) and (1) .. (26), quantities *i, s, k, d* and decision variable $X_{i,s,k,d}$ were vector parameters. The process of finding the solution may consist in using the constraints propagation methods, labeling of variables and the backtracking mechanism [13]. The quality of constraints propagation and the number of backtrackings are affected to a high extent by the number of parameters that must be specified/labeled in the given predicate/vector. In both models presented above, the classical problem representation included five parameters: *i, s, k, d* and $X_{i,s,k,d}$. Considering the domain size of each parameter, the process is complex and time-consuming. Our idea was to transform the problem by changing its representation without changing the very problem. All permissible routes were first generated based on the fixed data and a set of orders, then the specific values of parameters *i, s, k, d* were assigned to each of the routes. In this way, only decision variables $X_{i,s,k,d}$ (deliveries) had to be specified. This transformation allows only one parameter search instead of five. This is possible due to the flexibility and features of the CLP environment.

This transformation fundamentally improved the efficiency of the constraint propagation and reduced the number of backtracks. A route model is a name adopted for the models that underwent the transformation (MILP-R).

D. Decision-making support

The proposed models in this platform can support decision-making in the following areas:

- the optimization of total cost of the supply chain (Table II);
- the selection of the transport fleet number, capacity and modes for specific total costs;
- the sizing of distributor warehouses and the study of their impact on the overall costs (Table III, Fig. 3.);
- the selection of transport routes for optimal total cost (Fig. 4.).

Detailed studies of these topics are being conducted and will be described in our future articles.

V. NUMERICAL EXPERIMENTS

In order to verify and evaluate the proposed approach, many numerical experiments were performed. All the examples relate to the supply chain with seven manufacturers *(i=1..7),* three distributors *(s=1..3)*, ten customers *(j=1..10),* three modes of transport *(d=1..3),* and twenty types of products *(k=1..20).* Experiments began with nine examples of P1 .. P9 for the optimization MILP model (1) .. (23). The examples differ in terms of capacity available to the distributors *s* *(V_s),* the number of transport units using the mode of transport *d* *(Zt_d)* and the number of orders (*No*). The first series of experiments was designed to show the benefits and advantages of the hybrid approach. For this purpose the model (1) .. (23) was implemented in both the hybrid and integer programming environments. The experiments that follow were conducted to optimize examples P1..P9 for the optimization HP model (1) .. (26) in the hybrid approach.

Numeric data of input parameters for examples P1.. P9 are shown in Appendix A1. The results in the form of the objective function and the computation time are shown in Table II.

TABLE II
THE RESULTS OF NUMERICAL EXAMPLES FOR BOTH APPROACHES

P(No)	Hybrid Approach		Integer Programming		Hybrid Approach	
	MILP-R		MILP		HM	
	Fc	*T*	*Fc*	*T*	*Fc*	*T*
P1(100)	10791	416	15459	900**	10891	402
P2(90)	9263	323	9636	900**	9377	452
P3(80)	8388	522	8854	900**	8522	438
P4(60)	6330	345	6330	900**	6444	383
P5(40)	4473	203	4473	743	4708	223
P6(30)	3488	83	3488	503	3664	181
P7(20)	2877	23	2877	383	2894	31
P8(15)	2266	7	2266	503	2282	13
P9(10)	1756	2	1756	355	1756	2
Fc	the value of the objective function					
T	time of finding solution (in seconds)					
*	the feasible value of the objective function after the time T					
**	calculation was stopped after 900 s					
MILP	MILP model implementation in the IP environment.					
MILPT	MILP model after transformation-implementation in the hybrid implementation platform.					
HM	Hybrid model after transformation-implementation in the hybrid implementation platform.					

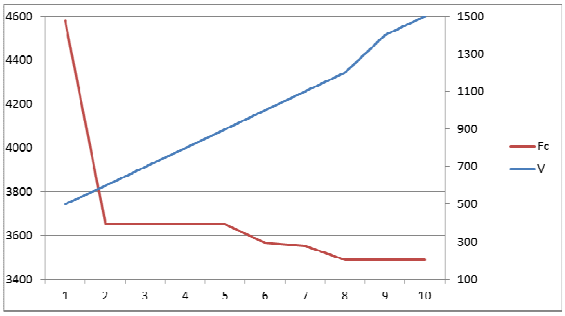


Fig. 3 The value of the objective function depending on the parameter V (Example P6)

For each example the solution for the MILP-R implementation was found faster than that for the MILP implementation. Moreover, for examples P1 .. P4, the traditional approach based on integer programming gives only feasible solution (calculation was stopped after 900 s) despite using highly efficient MILP solvers. It is obvious that the solution of the hybrid model (HM) was, due to its nature, only possible using the hybrid platform. Also, the proposed environment brought the expected results. The results were obtained in only a slightly longer period of time than that necessary for MILP-R (examples P1 .. P9).

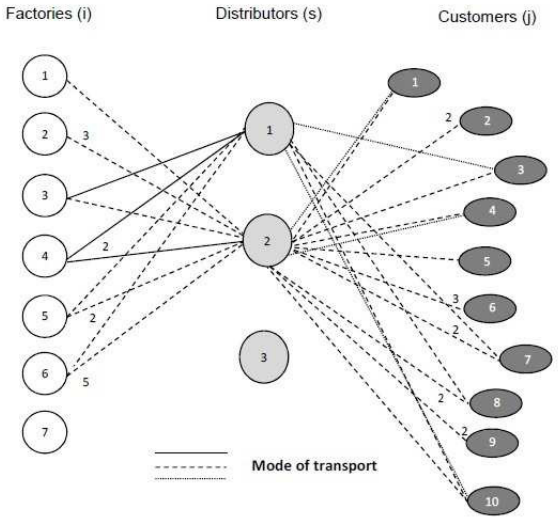


Fig. 4 Optimal transport network for the optimal solution (Fc^{opt}=8388) for P3 (no number means one)

TABLE III
ANALYSIS OF THE IMPACT PARAMETER V_S FOR FC (EXAMPLE P6)

V=V_1= V_2= V_3	Fc^{opt}	Distributor capacity (Vs) utilization		
		V_1	V_2	V_3
500	4581	470	336	350
600	3653	570	586	0
700	3653	570	586	0
800	3653	570	586	0
900	3653	256	900	0
1000	3567	188	968	0
1100	3554	130	1026	0
1200	3488	0	1156	0
1400	3488	0	1156	0

VI. CONCLUSION

The efficiency of the proposed approach is based on the reduction of the combinatorial problem. This means that using the hybrid approach practically for all models of this class, the same or better solutions are found faster (the optimal instead of the feasible solutions). Another element contributing to the high efficiency of the method is a possibility to determine the values or ranges of values for some of the decision variables (phase P3). All effective LINGO solvers can be used in the hybrid method.

It should be emphasized that with the used approach it is possible not only to solve optimization problems faster, but also to solve much larger problems than in the [17].

Therefore, the proposed solution is highly recommended for all types of decision problems in supply chain or for other problems with similar structure. This structure is characterized by the constraints of many discrete decision

variables and their summation. Furthermore, this method can model and solve problems with logical constraints.

In addition to the undoubted effectiveness of the proposed hybrid approach, should underline the possibility of modeling decision problems. The proposed approach can be created a new class of decision problems - hybrid problems that are not only familiar with the constraints from mathematical programming models but also new types of constraints such as logical constraints.

Further work will focus on running the optimization models with non-linear and other logical constraints, multi-objective, uncertainty etc. in the hybrid optimization framework. It is planned also apply this method to various types of scheduling problems [14],[28],[29].

APPENDIX A1

TABLE IV
DATA FOR COMPUTATIONAL EXAMPLES P1,P2,P3,P4,P5,P6,P7,P8,P9

k	V_k	k	V_k
01	1	11	8
02	2	12	4
03	5	13	5
04	9	14	5
05	3	15	7
06	4	16	8
07	5	17	9
08	6	18	1
09	7	19	4
10	8	20	6

j
01
02
03
04
05
06
07
08
09
10

d	Pt_s	Zt_s	Od_s
S1	400	25	240
S2	200	40	180
S3	100	80	100

i
01
02
03
04
05
06
07

k	k
06	07
09	10

k	i	k
01	01	02
01	02	01

s	V_s	F_s
C1	4000	100
C2	4000	100
C3	4000	400

i	s	d	$K_{i,s,d}$	$T_{i,s,d}$
03	01	01	35	4
04	01	01	44	5
07	01	01	17	2
01	01	02	5	1
02	01	02	15	2
03	01	02	18	3
04	01	02	22	4
05	01	02	16	3
06	01	02	18	3
07	01	01	8	1
03	02	01	46	5
04	02	01	38	4
07	02	01	35	4
01	02	02	18	3
02	02	02	14	2
03	02	02	24	4
04	02	02	17	3
05	02	02	8	1
06	02	02	5	1
07	02	02	17	3
03	03	01	5	1
04	03	01	34	4
07	03	01	48	5
01	03	02	15	3
02	03	02	30	5
03	03	02	5	1
04	03	02	15	3
05	03	02	15	3
06	03	02	20	4
07	03	02	22	3

i	k	$W_{i,k}$	$C_{i,k}$
01	01	500	0
01	02	500	0
01	03	500	50
01	04	500	50
02	01	500	50
02	02	500	50
02	03	500	0
02	04	500	0
03	05	500	0
03	06	500	0
03	07	500	0
03	08	500	0
04	08	500	30
04	09	500	0
04	10	500	0
04	11	500	0
07	05	500	40
07	06	500	30
07	07	500	40
07	08	500	50
07	09	500	40
07	10	500	60
07	11	500	10
05	12	500	0
05	13	500	0
05	14	500	10
05	18	500	0
05	19	500	0
05	20	500	20
06	14	500	0
06	15	500	0
06	16	500	0
06	17	500	0
06	18	500	20
06	19	500	20
06	17	500	0
06	20	500	0

s	j	d	$K_{s,j,d}$	$T_{s,j,d}$	s	j	d	$K_{s,j,d}$	$T_{s,j,d}$
01	01	02	10	1	02	05	02	16	2
01	02	02	29	3	02	06	02	5	1
01	03	02	34	3	02	07	02	35	4
01	04	02	44	4	02	08	02	36	4
01	05	02	31	3	02	09	02	28	3
01	06	02	35	3	02	10	02	41	5
01	07	02	17	2	02	01	03	26	3
01	08	02	45	5	02	02	03	16	2
01	09	02	57	6	02	03	03	36	3
01	10	02	17	2	02	04	03	6	3
01	01	03	5	1	02	05	03	5	2
01	02	03	19	3	02	06	03	25	1
01	03	03	24	3	02	07	03	26	3
01	04	03	34	4	02	08	03	26	3
01	05	03	21	3	02	09	03	18	2
01	06	03	25	3	02	10	03	31	4
01	07	03	7	2	03	01	02	33	4
01	08	03	35	5	03	02	02	59	6
01	09	03	40	6	03	03	02	5	1
01	10	03	7	2	03	04	02	34	4
02	01	02	36	4	03	05	02	30	4
02	02	02	26	3	03	06	02	45	5
02	03	02	46	4	03	07	02	48	5
02	04	02	38	4	03	08	02	69	7
03	05	03	20	3	03	09	02	10	2
03	06	03	30	4	03	10	02	52	6
03	07	03	32	4	03	01	03	23	3
03	08	03	50	6	03	02	03	40	4
03	09	03	8	1	03	03	03	5	1
03	10	03	40	6	03	04	03	24	3

s	k	$T_{s,k}$	s	k	$T_{s,k}$	s	k	$T_{s,k}$
01	01	1	02	01	1	03	01	2
01	02	1	02	02	1	03	02	2
01	03	1	02	03	1	03	03	2
01	04	1	02	04	1	03	04	2
01	05	1	02	05	1	03	05	2
01	06	1	02	06	1	03	06	2
01	07	1	02	07	1	03	07	2
01	08	1	02	08	1	03	08	2
01	09	1	02	09	1	03	09	2
01	10	1	02	10	1	03	10	2
01	11	1	02	11	1	03	11	2
01	12	1	02	12	1	03	12	2
01	13	1	02	13	1	03	13	2
01	14	1	02	14	1	03	14	2
01	15	1	02	15	1	03	15	2
01	16	1	02	16	1	03	16	2
01	17	1	02	17	1	03	17	2
01	18	1	02	18	1	03	18	2
01	19	1	02	19	1	03	19	2
01	20	1	02	20	1	03	20	2

Name	k	j	$T_{j,k}$	$Z_{j,k}$	Name	k	j	$T_{j,k}$	$Z_{j,k}$
z0101	01	01	25	8	z0105	05	01	30	10
z0116	16	01	50	7	z0106	06	01	20	10
z0201	01	02	10	10	z0219	19	02	20	10
z0202	02	02	40	8	z0220	20	02	10	10
z0301	01	03	20	10	z0302	02	03	10	10
z1019	19	10	15	8	z0303	03	03	10	8

z1020	20	10	35	8	z0419	19	04	30	10
z0901	01	09	30	10	z0420	20	04	25	10
z0401	01	04	40	10	z0501	01	05	15	10
z0505	05	05	60	8	z0502	02	05	10	10
z1013	13	10	15	8	z1015	15	10	10	8
z0911	11	09	20	10	z1016	16	10	20	10
z0912	12	09	25	10	z1017	17	10	20	10
z0806	06	08	50	8	z1018	18	10	30	10
z0807	07	08	60	10	z0917	17	09	30	10
z0705	05	07	60	10	z0918	18	09	30	10
z0706	06	07	20	8	z0919	19	09	40	10
z0604	04	06	30	10	z0920	20	09	40	8
z0605	05	06	35	10	z0809	09	08	55	10
z0606	06	06	50	8	z0810	10	08	30	8
z0103	03	01	10	8	z0811	11	08	30	10
z0209	09	02	20	8	z0812	12	08	20	10
z0309	09	03	30	10	z0708	08	07	30	8
z0410	10	04	40	10	z0709	09	07	60	10
z0514	14	05	30	8	z0710	10	07	30	10
z0614	14	06	20	10	z0711	11	07	10	10
z0719	19	07	10	8	z0609	09	06	10	8
z0720	20	07	30	10	z0610	10	06	30	10
z0818	18	08	25	8	z0611	11	06	30	10
z0819	19	08	25	10	z0612	12	06	30	10
z0102	02	01	15	10	z0107	07	01	30	10
z0104	04	01	15	10	z0108	08	01	45	10
z0203	03	02	50	10	z0109	09	01	30	10
z0204	04	02	20	10	z0110	10	01	10	10
z0304	04	03	10	8	z0210	10	02	20	10
z0305	05	03	30	10	z0211	11	02	30	10
z0406	06	04	40	8	z0217	17	02	20	10
z0407	07	04	50	10	z0218	18	02	10	10
z0512	12	05	30	8	z0308	08	03	30	10
z0513	13	05	20	10	z0312	12	03	30	10
z0615	15	06	10	10	z0315	15	03	20	10
….									

References

[1] Huang, G.Q., Lau, J.S.K., Mak, K.L., *The impacts of sharing production information on supply chain dynamics: a review of the literature.* International Journal of Production Research 41, 2003, pp.1483–1517.

[2] Kanyalkar, A.P., Adil, G.K., *An integrated aggregate and detailed planning in a multi-site production environment using linear programming*. International Journal of Production Research 43, 2005, pp. 4431–4454.

[3] Perea-lopez, E., Ydstie, B.E., Grossmann, I.E., *A model predictive control strategy for supply chain optimization*. Computers and Chemical Engineering 27, 2003, pp. 1201–1218.

[4] Park, Y.B.,. *An integrated approach for production and distribution planning in supply chain management.* International Journal of Production Research 43, 2005, pp. 1205–1224.

[5] Jung, H., Jeong, B., Lee, C.G., *An order quantity negotiation model for distributor-driven supply chains.* International Journal of Production Economics 111, 2008, pp. 147–158.

[6] Rizk, N., Martel, A., D'amours, S., *Multi-item dynamic production–distribution planning in process industries with divergent finishing stages.* Computers and Operations Research 33, 2006, pp. 3600–3623.

[7] Selim, H., Am, C., Ozkarahan, I., *Collaborative production–distribution planning in supply chain: a fuzzy goal programming approach.* Transportation Research Part E-Logistics and Transportation Review 44, 2008, pp. 396–419.

[8] Relich, M., Jakabova, M. *A decision support tool for project portfolio management with imprecise data*. In Proceedings of the 10th International Conference on Strategic Management and its Support by Information Systems, Valasske Mezirici 2013, Czech Republic, pp. 164-172.

[9] Chern, C.C., Hsieh, J.S., *A heuristic algorithm for master planning that satisfies multiple objectives*. Computers and Operations Research 34, 2007, pp. 3491–3513.

[10] Jang, Y.J., Jang, S.Y., Chang, B.M., Park, J., *A combined model of network design and production/distribution planning for a supply network.* Computers and Industrial Engineering 43, 2002, pp. 263–281.

[11] Timpe, C.H., Kallrath, J., *Optimal planning in large multi-site production networks*. European Journal of Operational Research 126, 2000, pp. 422–435.

[12] Mula J., Peidro D., Diaz-Madronero M., Vicens E., *Mathematical programming models for supply chain production and transport planning*, European Journal of Operational Research 204*,* 2010, pp. 377–390.

[13] Apt K., Wallace M., *Constraint Logic Programming using Eclipse*, Cambridge University Press, 2006.

[14] Sitek, *P. A hybrid approach to the two-echelon capacitated vehicle routing problem (2E-CVRP).* In Recent Advances in Automation, Robotics and Measuring Techniques, Advances in Intelligent Systems and Computing 267, 2014; pp. 251–263.

[15] Bocewicz G., Banaszak Z. Declarative approach to cyclic steady states space refinement: periodic processes scheduling. *In: International Journal of Advanced Manufacturing Technology*, Vol. 67, Issue 1-4, 2013, 137-155.

[16] Rossi F., Van Beek P., Walsh T., *Handbook of Constraint Programming (Foundations of Artificial Intelligence),* Elsevier Science Inc. New York, NY, USA, 2006.

[17] Sitek P., Wikarek J., *Cost optimization of supply chain with multimodal transport*, Federated Conference on Computer Science and Information Systems (FedCSIS), 2012, pp. 1111-1118.

[18] Sitek P., Wikarek J., *Supply chain optimization based on a MILP model from the perspective of a logistics provider*, Management and Production Engineering Review, 2012, pp. 49-61.

[19] Relich, M. *A declarative approach to new product development in the automotive industry.* In Environmental Issues in Automotive Industry, Springer Berlin Heidelberg 2014, pp. 23-45.

[20] Jain V., Grossmann I.E., *Algorithms for hybrid MILP/CP models for a class of optimization problems*, INFORMS Journal on Computing 13(4), 2001, pp. 258–276.

[21] Milano M., Wallace M., *Integrating Operations Research in Constraint Programming,* Annals of Operations Research vol. 175 issue 1, 2010, pp. 37 – 76.

[22] Achterberg T., Berthold T., Koch T., Wolter K., *Constraint Integer Programming. A New Approach to Integrate CP and MIP*, Lecture Notes in Computer Science Volume 5015, 2008, pp. 6-20.

[23] Bockmayr A., Kasper T., *Branch-and-Infer, A Framework for Combining CP and IP*, Constraint and Integer Programming Operations Research/Computer Science Interfaces Series, Volume 27, 2004, pp. 59-87.

[24] Sitek, P., Wikarek, J*., A hybrid approach to modeling and optimization for supply chain management with multimodal transport*, IEEE Conference: 18th International Conference on Methods and Models in Automation and Robotics (MMAR), 2013, Pages: 777-782.

[25] Awasthi A., Grzybowska K., Chauhan S., S K Goyal., *Investigating Organizational Characteristics for Sustainable Supply Chain Planning Under Fuzziness*, Supply Chain Management Under Fuzziness, Studies in Fuzziness and Soft Computing 313, 2014,pp.81-100.

[26] Grzybowska K. *Supply Chain Sustainability – analysing the enablers Springer Environmental issues in supply chain management* - new trends and applications, P. Golinska, C. A.Romano (Eds.) 2012,pp. 25-40.

[27] Grzybowska K., Kovács G. *Developing Agile Supply Chains - system model, algorithms, applications* Springer Agent and Multi-Agent Systems. Technologies and Applications, Lecture Notes in Computer Science, Jezic G. et. al. (eds.) 2012 pp. 576-585 .

[28] S.Deniziak, "*Cost-efficient synthesis of multiprocessor heterogeneous systems*", Control and Cybernetics, Vol.33, No.2, 2004, pp.341-355.

[29] Dang, Q.V., Nielsen, I. and Steger-Jensen, K., 2013, *Scheduling a single mobile robot incorporated into production environment*. In: P. Golinska, ed. EcoProduction & Logistics – emerging trends and business practices, part 2, Springer-Verlag Berlin Heidelberg, ISBN 978-3-642-23552-8, ISBN 978-3-642-23553-5 (eBook), pp. 185-201.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1353–1359
DOI: 10.15439/2014F201
ACSIS, Vol. 2

ROAD VEHICLES IDENTIFICATION AND POSITIONING SYSTEM

Cemil SUNGUR
Selcuk University, School of Technical Sciences.42250 Campus-KONYA
Email: csungur@selcuk.edu.tr

Hacı Bekir GÖKGÜNDÜZ
Selcuk University, School of Technical Sciences.42250 Campus-KONYA
Email: gokgunduz@selcuk.edu.tr

Adem Alpaslan ALTUN
Selcuk University, Faculty of Technology.42250 Campus - KONYA
Email: altun@selcuk.edu.tr

Abstract— Radio frequency identification (RFID) is untouched automatic identification technologies in which information can be transmitted by radio frequency. In this study, an RFID system which can determine the position and place of any desired vehicle is designed and developed. Microprocessor transmitter devices, on which the vehicle information is loaded, are placed on the vehicles. Receiver circuits are located at certain points in order to receive the RF signals sent by these transmitter devices. The communication of all the points with each other is maintained by connecting the receiver circuits via the internet. The functionality of the system is tested by performing receiver-transmitter experiments based on the various speeds and locations of the vehicles. According to the results obtained in the experiments, it is seen that the system designed in the study could be used in place of the GPS system for determining the place and position of vehicles, since it is more economical when compared to the GPS system.

I. Introduction

Nowadays, private organizations, individuals, and particularly public institutions have a rapidly increasing need for the positional information of various objects. The increase in the demand for information and in variety requires the collection of the positional information of individuals, institutions and regions in a faster and more economical way. The desire to maintain continuity and speed in the currency of information has resulted in rapid developments in Global Positioning Systems (GPS). The need for information can be evaluated on three basic scales, as local, regional and global.

The use of GPS systems is expensive and limited, and such systems have strategic importance. For this reason, it is necessary to create alternative systems. Different from the systems that are established and used by developed countries and the use of which by other countries is restricted, in the present study, a system whose use is not dependent on the permission of other countries, which does not require the use of satellites and which is considerably economical when compared to other systems is designed. Vehicle recognition, tracking and positioning system is implemented in order to determine the location and position of vehicles by using RF signals. In this way, although not point-wise, it is possible to determine the position of a vehicle in a narrow area.

Various results have been obtained in previously conducted similar studies [1]. A vehicle management system based on UHF band RFID technology is proposed in [2]. This system is applied to vehicles entering/leaving at road gates. The system consists of tag-on-car, reader antenna, reader controller, and monitoring and commanding software. Schneider proposes radio frequency identification technology for potential applications in the commercial construction industry [3]. It is defined that RFID can increase the service and performance of the construction industry with applications in materials management, tracking of tools and equipment, automated equipment control, jobsite security, maintenance and service, document control, failure prevention, quality control, field operations, and construction safety. Tu et al. examine radio frequency identification (RFID) tags, which are increasingly being used in pervasive healthcare applications [4]. Specifically, they study the dynamics of locating and identifying the presence of a tag in such systems. In [5], web servers build the agents by themselves and an agent-based interaction works with the support of Web services. Thus, they supposed to build an agent-based structure for transportation control that is similar to the structure of the Internet. Agent-based transportation management is a possible contribution to make transportation management more effective in regard to saving energy (fuel) and protecting our environment by stopping the increase of existing traffic ways. Liu et all. study on a platform for moving vehicle detection and tracking [6]. A novel Marr wavelet, kernel-based background modeling method and a background subtraction method based on binary discrete wavelet transforms (BDWT) are introduced. In addition, an automatic particle filtering algorithm is used to track the vehicle after detection and obtaining the center of the object. Böse et al. investigate an innovative approach to autonomous control in automobile logistics, considering as example the logistic order processing of an idealized automobile terminal of the company [7]. They present the results of an executed case study concerning the implementation and test of an

This work was not supported by any organization

autonomously controlled, radio frequency identification (RFID) based storage management system. In [8], an in-vehicle signing system is built and assessed that uses general-purpose RFID tags as digital traffic signs, and a field test is conducted using tags installed on a road to verify whether the system works effectively. In [9], a radio frequency identification (RFID) system for defining railway vehicles and an antenna is designed for the receiver/transmitter circuit of the RFID system is proposed. An intelligent traffic management system by using RFID technology is proposed in [10]. Using the system, important traffic and control data is obtained in a practical way, and also illegal vehicles, such as those which have been stolen are tracked.

II. GLOBAL POSITIONING SYSTEM

Global Positioning System (GPS) is a satellite network which conveys coded information in a regular manner. This system is composed of 24 satellites which continuously move in orbit and broadcast very low-power radio signals. The GPS receiver on the Earth's surface makes it possible to determine our definite position on the Earth by receiving those radio signals and measuring our distance from the satellites [11]. The functioning structure of the global positioning system is presented in Figure 1.

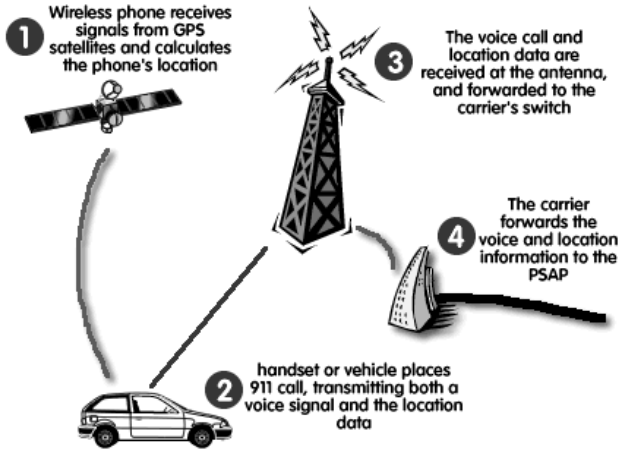


Fig. 1. The structure of the GPS system

The structure of the GPS system is composed of three major parts: the space segment, the control segment and the user segment. The space segment consists of at least 21 active satellites and 3 on-orbit spares. Each of the satellites transmits radio signals on two different frequencies. Civilian GPS satellite signals are transmitted at a frequency of 1575.42 MHz in the UHF band, while military GPS signals are transmitted at a frequency of 1227.6 MHz and are used by the US defense unit. Each satellite broadcasts two special pseudo-random codes which enable the receivers on the Earth to recognize the signals. These signals include the orbit, time, general system and ionospheric delay information of the satellite. The control segment obtains the exact orbit and time information by continuously tracking the satellites. There are several control stations on the surface of the Earth [12]. The user segment is comprised of all of the GPS receivers on the Earth.

The GPS tracking system was originally intended for military applications, but in the 1980s, the system was made available for civilian use upon the increasing demand. The proportion of the civilian use of GPS systems has reached a high rate of 90% compared to its use for military purposes [13].

III. RF COMMUNICATION

Communication technology, which has gained a new dimension and speed in recent years, has led to the development of techniques that enable the use of the air as a transmission environment and the sending of information to the desired point by coding the signal on a high frequency carrier wave, which is also named as modulation.

An electromagnetic wave is formed when an electric field combines with a magnetic field. These waves, which are combined and sent by the RF transmitter, are separated by the RF receiver and transformed into electric and magnetic waves [14].

Devices which are used for modulating the amplitude, frequency or the phase of the information signal and the high frequency carrier signal and then transmitting the information through the air or a conductive environment are called RF transmitters. RF receivers are devices by which the signals coming from the RF transmitter are received by an antenna and amplified by the power amplifier, then mixed with the local oscillator signal. Then, the RF signal is transformed to information by being demodulated at the output of the mixer. The operating principles of RF transmitter and receiver circuits are presented in Figure 2 [14].

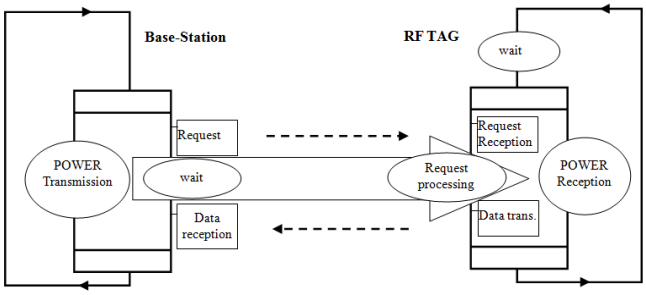


Fig. 2. RF Receiver and Transmitter Circuit Block Diagram

As shown in Figure 2, RFID system is composed of tags, which carry the data in suitable transponders, and an RFID reader, which retrieves the data from the tags, and software such as a driver and middleware. The main function of the RFID system is to retrieve information (ID) from a tag (also known as a transponder). Tags are usually affixed to objects such as goods or vehicles so that it becomes possible to locate where the goods and vehicles are without line-of-

sight. A tag can include additional information other than the ID, which opens up opportunities to new application areas. In these systems, the RF area is broken down into a number of grouped frequencies that are called bands. The frequencies used in RFID systems are analyzed below:

• Less than 135 kHz: This is the low frequency (LF) which allows the detection of RFID tags in a short range. The data reading speed is low and in this frequency the electromagnetic waves penetrate water but not metal. This frequency is used for animal identification, inventory control and car immobilizer [15].

• 13.56MHz: This frequency, which is called as high frequency (HF), allows the detection of RFID tags for a distance of up to 1.5m. The data transfer rate for this specific frequency is approximately 25 Kbits per second. In this frequency, the electromagnetic waves can penetrate water but have poor performance around metal. This frequency is used for applications related to access, security payment and various item levels tracking such as books, luggage, etc. [16].

• 433MHz, 850–956MHz: The frequencies which belong to this range are characterized as ultra-high frequencies (UHF). This frequency allows high reading speed as approximately 100 Kbits per second. The frequencies at this range are used for applications in vehicle tracking, railway vehicle monitoring [16].

• 2.45-5.8GHz: This frequency enables an RFID reader to detect a tag from a distance of ten meters. This specific frequency is characterized as microwave frequency. Data transfer rates are fast but the reading range is similar to UHF. The specified frequency is used for applications related to industrial automation [16].

In our vehicle tracking system, an ATX-34 circuit is used as the RF transmitter and an ARX-34 circuit is used as the RF receiver. ATX-34 and ARX-34 circuits are preferred for their high stability and wide area of use. These circuits comply with the radio standard EN 300 220 at 433.920 MHz UHF band, have high frequency stability and are ideal for battery applications with their low current consumption [17].

In this study, the PIC16F877 microcontroller manufactured by Microchip Technology Inc is used for entering the license plate information of the vehicles, and programming and storing the information coming from the transmitter [18]. This microcontroller is programmed using the JAL programming language [19]. The antenna used in the study has an impedance of 50 Ω and is approximately 17.3 cm.

The transmitter circuit is programmed together with the license plate information. The PIC16F877 integrated circuit can be locked against rewriting and reading during the programming process. In this way, the access of unauthorized persons to the license plate information of vehicles can be restricted. The license plate information is not transferred by using a standard on the transmitter circuit so that it cannot be easily acquired. The circuit transmits the license plate information through a completely specific protocol. Figure 3 shows the fundamental steps of the RFID system.

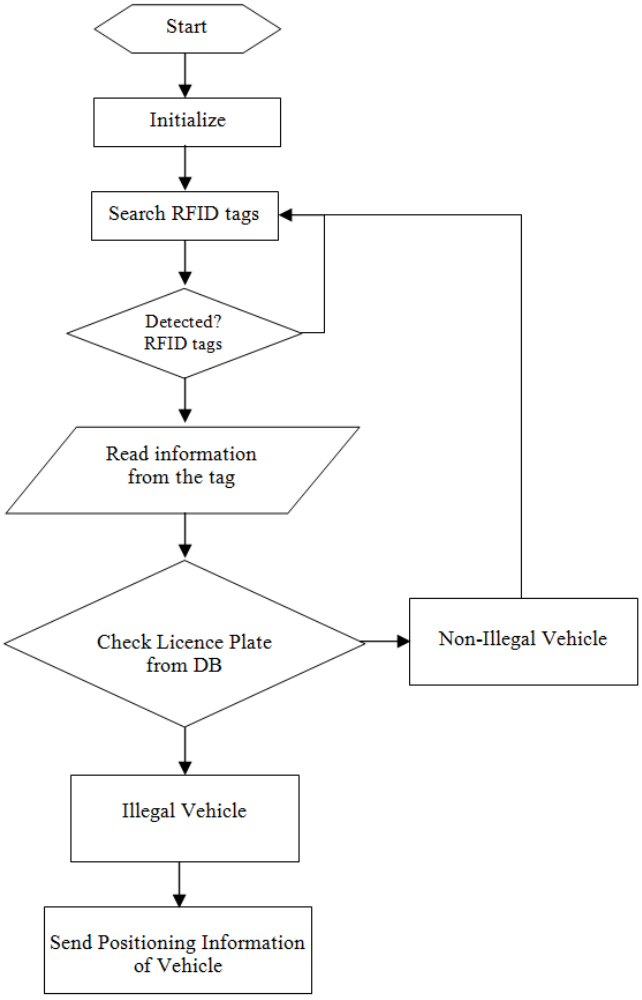


Fig. 3. Diagram of the RFID system for vehicle identification

In each transmission, the transmitter circuit sends a data package. The eight-character license plate information is transmitted in this data package as shown in Figure 4. The license plate information is added to the beginning of the preamble data package. The radio preamble is a section of data at the head of a packet that contains information the access point and the client devices need when sending and receiving packets. The preamble data is a leading signal which is composed of 6 bits and each period of which is 1ms. This signal is used in order to prevent the faulty reception of the initial data by a receiver. The faulty reception of the initial data is prevented by sending a leading preamble data. After this, the synchronous signal and the 8-byte data are transmitted. The permanent duty of the transmitter circuit is to continuously broadcast the recorded license plate information.

4	2	B	V	0	2	1	3

Fig. 4. 8-bit license plate information data

The general structure and functioning of the system is presented in the block diagram of the circuit shown in Figure 5.

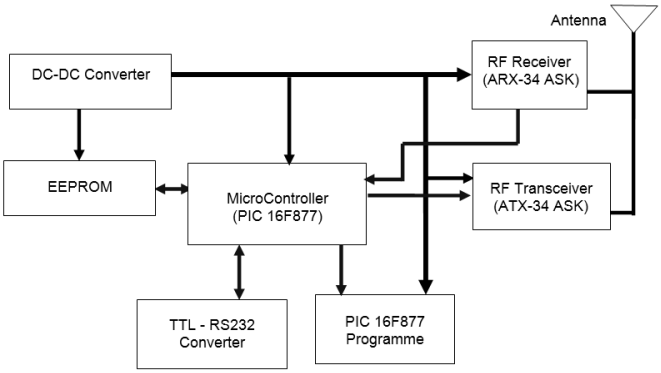


Fig. 5. The circuit block diagram of the designed RFID system

When the receiver circuit receives RF information from the transmitter circuit that complies with its protocol, it transmits this information to the computer through the RS232 port. The data is stored on the computer and the desired information can be selected. The searched license plates are reported to the program and the program searches for the wanted plates among the received data. When one of the searched license plates enters into the coverage area of the tracking system, the program records the entry with date and time, and displays an alert on the screen to warn the user. The circuit transmits the data which is only in a compatible format at a speed of 115.200 bps.

The signals received from the receivers are processed by the interface program. There is a transmitter circuit which has a specific code on each vehicle. As it can be seen in Figure 6, the interface consists of three main parts. The first of these is the Passing Vehicles Part, the second is the part where we enter the license plate information of the vehicle we want to find, and the third is the part where we determine at what time a certain vehicle passes a certain point.

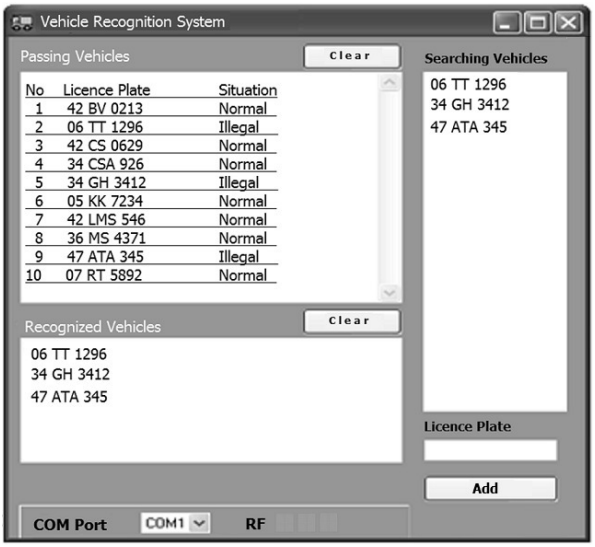


Fig. 6. The interface program developed for the vehicle tracking system

When the program is started, the information obtained in the three parts begins to be stored in the database. In this way, access is provided to the information concerning a desired time, day and hour.

IV. EXPERIMENTAL RESULTS

Two transmitter and one receiver circuits are used in the system designed for this study. The receiver circuit is located at a point on the route determined for the study. The two transmitters are placed on two different vehicles. The vehicle information of each vehicle passing along the route is read by the RF receiver. In this way, it is known which city the vehicle travels to, where it stops over and from which direction the vehicle enters the city. An area on which the license plate information of the searched vehicles can be entered is added to the computer software used for recording and monitoring vehicle information. In this way, when the sought vehicle enters the coverage area of the receiver, the vehicle information is delivered to the system user through the computer.

The license plate information of the sought vehicle is entered to the system by the traffic data processing unit of the city and it can be seen by the traffic information centers of all the cities with the help of a computer network. When the vehicle is detected within the borders of a city, the location of the vehicle is transmitted to the system users of all the cities via the computer network.

In this study, it is planned to place the system at the entrance and the exit of the city. However, the system can also be used at traffic points, intercity traffic stations, intersections and highway entrance and exit booths.

In the present study, a total of five experiments are conducted on; the distance required for the recognition of the vehicles equipped with a transmitter by the receiver, the speed of the vehicle which is effective on recognition,

weather conditions, the effect of the position of the receiver and the transmitter on the recognition of the vehicle.

A. Determination of the Recognition Distance

The receiver is placed 1 m above the ground, and the approach and deviation distances are found to be 30 m at the 10 passes made at a speed of 50 Km/h by the transmitter placed on the vehicle, as shown in Figure 7.

When the receiver is placed at a height of 2.20 m, it is observed that the approach and deviation distances are found to be still equal at the 10 passes made at a speed of 50 Km/h, but the recognition distance of the vehicle has increased to 50 m.

When the receiver is placed at a height of 4.60 m, it is observed that the approach and deviation distances are found to be still equal at the 10 passes made at a speed of 50 Km/h, but the recognition distance of the vehicle has increased to 75 m.

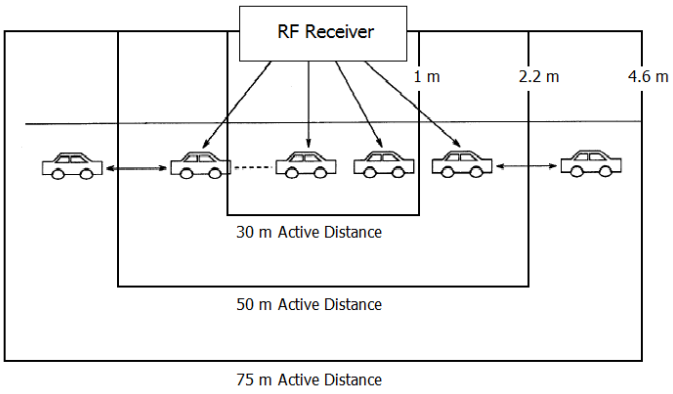


Fig. 7. The distances at which the transmitter is detected by the receiver

B. Determination of the Speed that Prevents the Recognition of the Vehicle

It is necessary that the vehicles passing the areas where the receivers are located be sensed by the receiver. The receiver may not sense the vehicle due to several factors, such as the passing of more than one vehicle simultaneously and the speed of the passing vehicle. The height of the receivers and the speed of the vehicles are important for sensing the passing vehicles in a reliable manner. For this reason, the receiver is placed at a height of 2.20 m and the approach and deviation distances of the vehicle are found to be the same at the passes made at speeds of 20 Km/h, 50 Km/h, 80 Km/h, 100 Km/h and 120 Km/h, and the recognition distance of the transmitter by the receiver is found to be 50 m.

C. Recognition Number of the Vehicle

It is necessary that the receiver senses the vehicles passing the active region and obtains reliable data. For this reason, it is important how many times the receiver communicates with the transmitter in the active region. In this study, it is determined how many times the receiver recognizes the transmitter in the active region depending on the speed of the vehicle and the obtained number is compared with the calculations performed based on the frequency of the transmitter. When the period required for the wavelength of the transmitter broadcasting at a frequency of 433.927 MHz is calculated.

$$T = 1 / f \quad (1)$$

where; T is the period (s), f is the frequency (Hz). Thus period is obtained as 2.3 ns.

When the distance covered by a vehicle travelling at a speed of 100 Km/h is calculated.

$$S = V / t \quad (2)$$

where; S is the distance traveled in 1 s; V is the speed of the vehicle (m/s); t is the total time (s).

In Eq. 2, S is calculated as 100.000 / 3.600 = 27.7 m. Therefore, the vehicle would pass the 50m active region in 2 s considering the length of the vehicle (supposing that the vehicle is a passenger car).

A number of 2 s / 2.3 ns = 0,869 x 109 signals are sent to the receiver from the transmitter. The vehicle license plate consists of 8 characters. Each character requires successive continuous data transfer at periods of 1 ms. For this reason, information of a vehicle is sent over to the receiver every 8 ms. Moreover, when we add the starting and ending bits, this period is extended to 10 ms. That is, a vehicle information reaches the receiver every 10 ms, therefore the receiver can read the information from the transmitter 2 s / 10 ms = 200 times within a passing period of 2 s. Each piece of information is read 4 times consecutively in the interface program to ensure that the vehicle is completely recognized. Thus, the vehicle information is read 200 / 4 = 50 times. It is expected that this number of readings will be adequate for the receiver to read the transmitter without missing the receiver even if there is more than one vehicle in the active reading area.

The receiver is placed at a height of 2.20 m and the experiment is repeated by passing the transmitter by the receiver point 5 times at each speed of 20 Km/h, 50 Km/h, 80 Km/h, 100 Km/h and 120 Km/h.

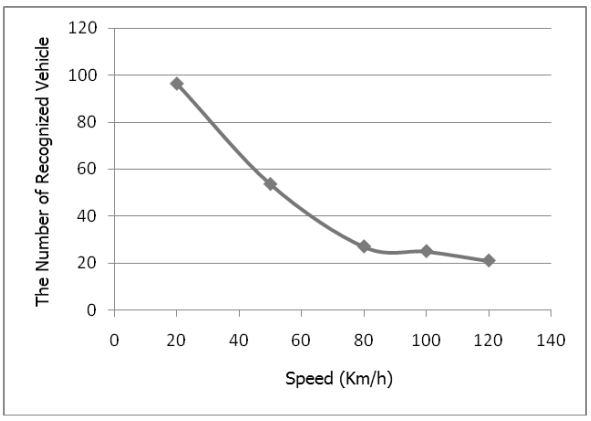


Fig. 8. Speed – Number of Detection Graph

The data obtained as the result of this experiment is shown in Figure 8. As it can be seen in the graph, as the speed of the vehicles increase their number of detection decreases. The maximum speed limit is 120 Km/h in our country. For this reason, our experiment is conducted up to this limit.

When more than one vehicle enters the recognition distance of the receiver at the same time, the number of detection of each vehicle is measured separately. As it can be seen in Figures 9 and 10, the entry of two vehicles of the same type into the active area does not affect the detection of the vehicles. However, although the distance of detection remains the same, the number of detection decreases to half, even to 40-45% when compared to the experiment conducted with one vehicle.

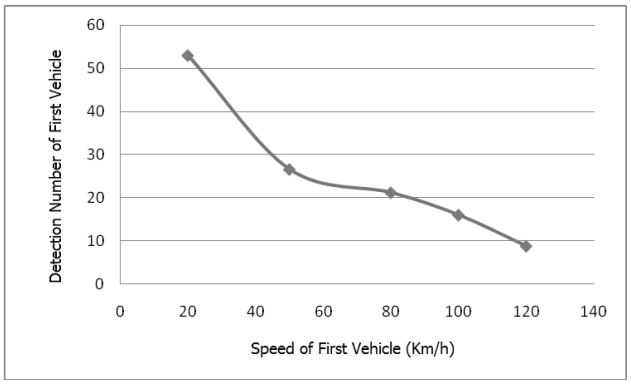


Fig. 9. 1st Vehicle Speed – Number of Detection Graph

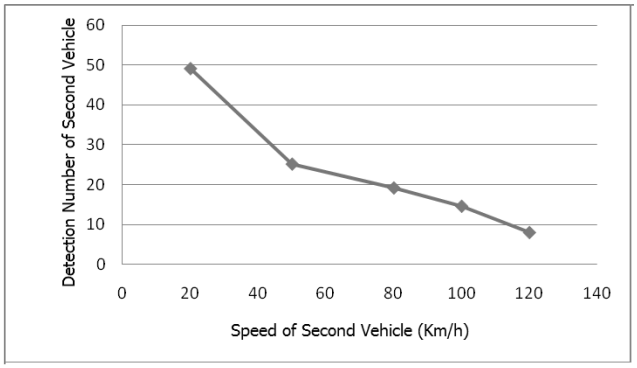


Fig. 10. 2nd Vehicle Speed – Number of Detection Graph

In a further experiment, the transmitter vehicle is passed by a heavy vehicle at different speeds as shown in Figure 11, and the values presented in Table 1 are obtained as the result of the experiment.

As it can be seen in Table 1, the number of detection of the transmitter vehicle which is behind the heavy vehicle significantly decreased. This could be considered as a disadvantage of our vehicle tracking system.

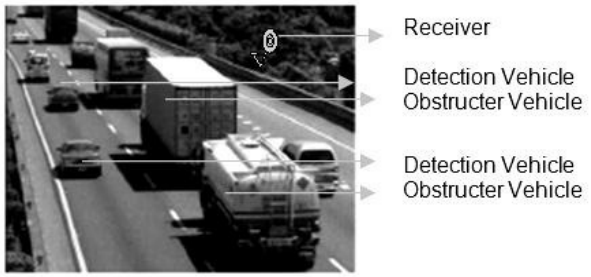


Fig. 11. Moment of passing of the undetected vehicles

TABLE I.
NUMBER OF DETECTION OF A VEHICLE BEHIND HEAVY VEHICLES

Speed (Km/h)	Number of Detection					Mean
	1st Pass	2nd Pass	3rd Pass	4th Pass	5th Pass	
20	45	15	14	14	15	20.6
50	6	5	8	7	8	6.8
80	3	2	0	0	2	1.4
100	2	1	0	0	1	0.8
120	0	0	1	0	0	0.2

In this study, an RFID system for tracking and determining the position of vehicles is designed. In this system, all the vehicles on which transmitters are placed are detected and recognized by the system when they pass by the points where receivers are located. Experiments are conducted in order to determine the speed that prevents the recognition of the vehicle, to recognize each vehicle when more than one vehicle enters the recognition distance of the receiver at the same time and to determine the physical obstacles that might prevent the recognition of the vehicle. In the designed system, the license plate information of the vehicles can be viewed in the interface program. This information is recorded on the computer via the interface program. When the license plate information of the sought vehicle is entered into the interface program, the exact time when the vehicle passes a previously determined point is displayed as day, hour, minute and second. Furthermore, all the vehicles detected by the receiver are listed in the interface program. The process of determining the place and position of a vehicle, which can be carried out using the GPS system whose use is restricted, is easily performed. In this way, a moving vehicle can be tracked by using the system developed in this study.

It is observed that the performance of the system decreases when the experiments described above are performed under maximum conditions, that is, under very heavy traffic and unfavorable weather conditions. Under such conditions, the performance of the system can be enhanced by placing the receiver in a higher position, adjusting its direction and increasing the signal level of the transmitter.

The system designed in the present study can be adjusted in a manner that enables access to insurance, vehicle inspection, vehicle license and driving license information

through certain changes and additional software. Access to the mentioned information can be rendered exclusive to certain institutions, organizations and individuals. In this way, it can be possible to receive information regarding when and from where a vehicle passes and also to access vehicle license and driving license information.

D. Conclusion

This research has investigated the practical issues of deploying the RFID technology for the vehicle identification started from hardware procurement process. The paper has researched several technical aspects of RF-ID implementation. Microprocessor transmitter devices, on which the vehicle information is loaded, are placed on the vehicles. Receiver circuits are located at certain points in order to receive the RF signals sent by these transmitter devices. The communication of all the points with each other is maintained by connecting the receiver circuits via the internet. According to the results obtained in the experiments, it is seen that the system designed in the study could be used in place of the GPS system for determining the place and position of vehicles, since it is more economical when compared to the GPS system.

The results of the experiments indicate that even implementing the system in a small scale (prototype system), there are a number of challenges to be solved in the middleware development phase, and extraordinary challenges in the configuration and tuning phases. Therefore to gain a cost effective system from deploying RFID for vehicle tracking in practice, extra attention is definitely required, and further experimentation on a real-life scale (e.g., real vehicle and gate) and/or using different sets of equipment will help confirm the results from this research.

References

[1] Young CP, Chang BR, Tsai HF, Fang RY. and Lin JJ (2009) Vehicle Collision Avoidance System using Embedded Hybrid Intelligent Prediction based on Vision/Gps Sensing, International Journal of Innovative Computing, Information and Control, vol. 5 (12(A)), pp. 4453-4468

[2] Ning Y., Zhong-qin W., Malekian R., Ru-chuan W. and Abdullah A.H. (2013) Design of Accurate Vehicle Location System Using RFID, Elektronika Ir Elektrotechnika, Vol. 19, No. 8, pp. 105-110.

[3] Schneider M(2003) Radio Frequency Identification RFID Technology and its Applications in the Commercial Construction Industry, Master Thesis, University of Kentucky

[4] Tu YJ, Zhou W and Piramuthu S (2009) Identifying RFID-embedded objects in pervasive healtcare applications, Decision Support Systems, vol. 46, pp. 586-593

[5] Franke H and Dangelmaier W (2004) A web-based multi-agent system for transportation management to protect our natural environment, Cybernetics and Systems, vol. 35:7, pp. 627-638

[6] Liu ZG, Yue SH, Mei JQ and Zhang J (2009)Traffic video-based moving vehicle detection and tracking in the complex environment, Cybernetics and Systems, vol. 40: 7, pp. 569-588

[7] Böse F, Piotrowski J and Scholz-Reiter B (2008) Autonomously controlled storage management in vehicle logistics–applications of RFID and mobile computing systems, International Journal of RF Technologies: Research and Applications, pp. 1-20,

[8] Zoghi H, Tolouei M, Siamardi K and Araghi P (2008)Usage of ITS in the In-vehicle Signing System with RFID tags and Vehicle Routing and Road Traffic Simulation, 11th International IEEE Conference on Intelligent Trasportation Systems, pp. 408-413

[9] Jinghui Q, Bo S and Qidi Y (2006) Study on RFID Antenna for Railway Vehicle Identification, 6th International Conference on ITS Telecommunications, pp. 237-240

[10] Antikainen H, Colpaert A, Jaako N, Rusanen J, Bendas D, Myllyaho M, Oivo M, Kuvaja P, Similä J, Marjoniemi K, Laine K and Saari E (2004) Mobile Environmental Information Systems, Cybernetics and Systems, vol. 35: 7, pp. 737-751

[11] Sun Z. and (Jeff) Ban X. (2013) Vehicle classification using GPS data, Transportation Research Part C: Emerging Technologies, Volume 37, pp. 102-117.

[12] El-Rabbany (2006) A Introduction to GPS: The Global Positioning System 2nd Ed., Artech House

[13] Psiaki, M. L.; O'Hanlon, B. W.; Bhatti, J. A.; Shepard, D. P.; Humphreys, T. E. (2013) GPS Spoofing Detection via Dual-Receiver Correlation of Military Signals, Aerospace and Electronic Systems, IEEE Transactions on , vol.49, no.4, pp.2250-2267.

[14] Park S.; Lee H. (2013) Self-Recognition of Vehicle Position Using UHF Passive RFID Tags, Industrial Electronics, IEEE Transactions on, vol.60, no.1, pp.226-234.

[15] Liu, S., Wang, X., Shen, J., Wang, B., Ye, T., & Huang, R. (2014). A novel low-noise high-linearity CMOS transmitter for mobile UHF RFID reader. Science China Information Sciences, 1-8.

[16] Ward M and van Kranenburg R (2006) RFID:Frequency, Standards, Adoption and Innovation, JISC Technology and Standards Watch

[17] Kucukkomurler A (2009) Thermoelectric Powered High Temperature Wireless Sensing, Journal of Thermal Science and Technology, vol. 4, no. 1, pp. 63-73

[18] Ragul, M., & Venkatesh, V. (2013). Autonomous vehicle transportation using wireless technology. International Journal of Engineering and Technology (IJET), 5(2).

[19] Apaydin, S. F., Çavuşoğlu, A., & Lami, K. A. Y. A. (2013). The Experiments And Performance Analyzes Of The Tracker Robots In Different Lighting Environments. International Journal, 2(1), 2305-1493.

Collaborative Human-Machine Intelligence 20th Conference on Knowledge Acquisition and Management and 2nd Workshop on Artificial Intelligence for Knowledge Management

KNOWLEDGE management is a large multidisciplinary field having its roots in Management and Artificial Intelligence. Activity of an extended organization should be supported by an organized and optimized flow of knowledge to effectively help all participants in their work.

We have the pleasure to invite you to contribute to and to participate in the conference "Knowledge Acquisition and Management" and 2nd Workshop on "Artificial Intelligence for Knowledge Management" (KAM&AI4KM'14). The predecessor of the KAM conference has been organized for the first time in 1992, as a venue for scientists and practitioners to address different aspects of usage of advanced information technologies in management, with focus on intelligent techniques and knowledge management. In 2003 the conference changed somewhat its focus and was organized for the first under its current name. Furthermore, the KAM conference became an international event, with participants from around the world. In 2012 we've joined to Federated Conference on Computer Science and Systems becoming one of the oldest event.

The "Artificial Intelligence for Knowledge Management" Workshop was initiated by IFIP Group TC12.6 in 2012 as the separate event during European Conference on Artificial Intelligence in Montpellier (ECAI'2012). From the beginning the workshop aims in bringing together researchers and practitioners involved in Knowledge Management using the methods and techniques of AI for building and improving all aspects of KM and of knowledge flow, among others, improvement of the innovation process. This year both teams KAM and AI4KM, have decided to join efforts in the FedCSIS framework with common challenge: "Collaborative Human-Machine Intelligence" under IFIP supporting.

The aim of this common event is to create possibility of presenting and discussing approaches, techniques and tools in the knowledge acquisition and other knowledge management areas with focus on contribution of artificial intelligence for improvement of human-machine intelligence and face the challenges of this century. We expect that the conference&workshop will enable exchange of information and experiences, and delve into current trends of methodological, technological and implementation aspects of knowledge management processes.

Topics

The following group topics, concerning both theory and applications, will be included (unavoidably incomplete):

- Knowledge discovery from databases and data warehouses
- Methods and tools for knowledge acquisition
- New emerging technologies for management
- Organizing the knowledge centers and knowledge distribution
- Knowledge creation and validation
- Knowledge dynamics and machine learning
- Distance learning and knowledge sharing
- Knowledge representation models
- Management of enterprise knowledge versus personal knowledge
- Knowledge managers and workers
- Knowledge coaching and diffusion
- Knowledge engineering and software engineering
- Managerial knowledge evolution with focus on managing of best practice and cooperative activities
- Knowledge grid and social networks
- Knowledge management for design, innovation and eco-innovation process
- Business Intelligence environment for supporting knowledge management
- Knowledge management in virtual advisors and training
- Management of the innovation and eco-innovation process
- Human-machine interfaces and knowledge visualization

Event Chairs

Hauke, Krzysztof, Wroclaw University of Economics, Poland

Nycz, Małgorzata, Wrocław University of Economics, Poland

Owoc, Mieczyslaw, Wroclaw University of Economics, Poland

Pondel, Maciej, Wroclaw University of Economics, Poland

Program Committee

Abramowicz, Witold, Poznan University of Economics, Poland

Andres, Frederic, National Institute of Informatics, Tokyo, Japan

Chmielarz, Witold, Warsaw University, Poland

Christozov, Dimitar, American University in Bulgaria, Bulgaria

Goluchowski, Jerzy, University of Economics in Katowice, Poland

Helfert, Markus, Dublin City University, Ireland

Jelonek, Dorota, Faculty og Management of Czestochowa University of Technology

Ligęza, Antoni, AGH University of Science and Technology, Poland

Mach-Król, Maria, University of Economics in Katowice, Poland

Matouk, Kamal, Wroclaw University of Economics

Mercier-Laurent, Eunika, IAE Lyon3, France

Sobińska, Małgorzata, Wroclaw University of Economics

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1363–1369
DOI: 10.15439/2014F277
ACSIS, Vol. 2

CKD: a Cooperative Knowledge Discovery Model for Design Project

Xinghang Dai
University of Technology of Troyes, Tech-cico, 12 Rue Marie Curie, BP 2060, 10010 Troyes, France
Email: xinghang.dai@utt.fr

Nada Matta
University of Technology of Troyes, Tech-cico, 12 Rue Marie Curie, BP 2060, 10010 Troyes, France
Email: nada.matta@utt.fr

Guillaume Ducellier
University of Technology of Troyes, LASMIS, 12 Rue Marie Curie, BP 2060, 10010 Troyes, France
Email: guillaume.ducellier@utt.fr

***Abstract*—Knowledge management has become a vital strategy to conserve company knowledge and to reuse them. Research in this area has always been consenting on domain knowledge, domain ontology, expert systems etc. have been developed to manage professional domain knowledge, but less effort has been done on cooperative knowledge. In this paper, a cooperative knowledge discovery method DKD is proposed, we elaborated this method in design project knowledge management area.**

I. Introduction

COOPERATIVE activity is defined as an activity of several actors having a given goal [1], communication, coordination and collaboration are the three dimensions of cooperative activity [2]. Workflow, Groupware tools [3], design-rationale approaches [4] have been developed for CSCW issues. Design is a highly cooperative activity, in which people from different background, different organizations and with different skills work together to reach a given goal. Knowledge are produced in design activities. As design project team is a short-lived organization, in the end of a project team members will be engaged into another project under another project organization, the challenges for design project knowledge management is to enhance learning in an organization from experiences [5]. As in this paper, we will focus our knowledge management on cooperative knowledge produced in design projects.

Recent knowledge management research has proposed community of practices and story telling to enhance knowledge sharing in an organization. Experience shows that the success of these techniques depend on the dynamic of animation in these communities. Our work is based on knowledge engineering approaches. We believe that learning from experience requires two fundamental elements: reasoning strategies (also called behavior laws) [6] and production context of these strategies [7]. "The learning content is context specific, and it implies discovery of what is to be done when and how according to the specific organisations routines"[8]. These two elements are especially important for cooperative knowledge representation.

This paper will propose a cooperative knowledge discovery model CKD. It will be elaborated in design project knowledge management area. Our ambition is to define a cooperative ontology and a classification framework for cooperative knowledge management.

II. Cooperative Knowledge

Cooperative knowledge is defined as knowledge produced in cooperative activities [7]. As we mentioned above, three dimensions have to be considered in cooperative activities: communication, coordination and cooperative decision-making. In order to define cooperative knowledge in a formalized manner. We are going to propose a cooperative activity ontology. Ontology is a description of shared concepts. It consists of term, definitions, axioms, and taxonomy. It facilitates knowledge comprehension and knowledge sharing by setting the standard knowledge structure [9][10]. Traditional ontology consists of a hierarchy of concepts. However, in cooperative activity, concept can only have a sense when it is put in a specific context, in other words, interactions between concepts instead of concepts themselves are considered essential in cooperative activity. Hence, we come up with a cooperative activity ontology consisted of types of actions.

III. Cooperative Knowledge in Design Project

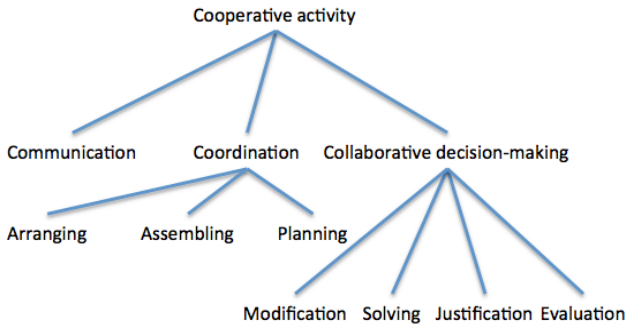


Fig 1. Cooperative activity ontology

Design activities have gone through some major changes with the use of IT tools in design projects. At the same time, developing pace of high technology pushes each day design product to be more and more complex. These changes

require design project to be multi-organizational and multi-disciplinary [11] [12]. Moreover, linear project management can no longer be applied on long-period multi-organizational and multi-disciplinary projects, a new concurrent working mode has emerged to allow people to work in a parallel manner, which necessitate more cooperative activities.

Both domain knowledge and cooperative knowledge are produced in design project. Past researches have progressed a lot on design domain knowledge management, but cooperative knowledge produced in design projects is different from design domain knowledge:

- The nature of knowledge is different: The domain knowledge is related to a field and contains routines and strategies developed individually from experiences, which involve a number of experiments. The cooperative knowledge is related to several fields, i.e. several teams (of several companies) and in several disciplines collaborates to carry out a project. So there is a collective and organizational dimension to consider in cooperative knowledge. Representing domain knowledge consists in representing the problem solving (concepts and strategies) [13]. On the contrary, emphasizing knowledge in cooperative activity aims at showing organization, negotiation and cooperative decision-making [14]. Otherwise, knowledge observed in a corporative constitutes examples to be structured in order to extract strategies.
- Capturing of knowledge is different: The realization of a project in a company implies several actors, if not also other groups and companies. For example, in concurrent engineering, several teams of several companies from several disciplines collaborate to carry out a design project. The several teams are regarded as Co-partners who share the decision-makings during the realization of the project. This type of organization is in general dissolved at the end of the project [15]. In this type of organization, the knowledge produced during the realization of the project has a collective dimension that is in general volatile. The documents produced in a project are not sufficient to keep track of this knowledge. In most of the cases, even the project manager cannot explain it accurately. This dynamic character of knowledge is due to the cooperative problem solving where various ideas are confronted to reach a solution. So acquisition of knowledge by interviewing experts or from documents is not sufficient to show different aspects of the projects, especially negotiation [16]. Traceability and direct knowledge capturing are needed to acquire knowledge from this type of organization.

For the same object, people with different background can give different interpretations; concept alters according to different context. As for design project, design decision-making process has always been the main research subject. However, decision-making process can not be fully represented without its context. Normally a decision-making process relies not only on design rationales, but also organizational influence, project constraints etc.. Therefore, we have to focus on design rationale representation as well as its interaction with other parts of a project. In other words, a global representation of all design projects modules as well as interactions between them are needed for design project. We should represent specially:

1. The design rationale (negotiation, argumentation and cooperative decision making)
2. The organization of the project (actors, skills, roles, tasks, etc.)
3. The consequences of problem solving (evolution of the artefact)
4. The context of the project (rules, techniques, resource, etc.)

We called the structure representing this type of knowledge project memory [17]. From the knowledge structure proposed by project memory, we will elaborate our CKD model.

IV. CKD IN DESIGN PROJECT

A. CKD framework

The principle of CDK method is to classify similar concept schemas of cooperative activities to identify certain repetitive ones as routines with a weight factor that indicates their importance. Classification can be defined as the process in which ideas and objects are recognised, differentiated, and understood; classifiers are widely used in biology, documentation, etc. [18]. A routine is defined as a recursive interaction schema of cooperative activity concepts. The weight factor is defined as percentage of recurrence of a routine among past similar project events. Therefore, the result of classification will be an ensemble of interactions between cooperative activity concepts. This result routine can be considered as a knowledge rule for cooperative actors to learn from, and future cooperative activities should pay attention to past knowledge rules. Before classification, cooperative activity information have to be structured, and we believe that semantic network graph is the perfect representation for that. A semantic network graph enable knowledge engineers to communicate with domain experts in language and notations that avoid the jargon of AI and computer science [19].

B. Design project structure

Section 3 has introduced “project memory” that list the four essential parts of design project. Current representation approaches emphasise on organising and structuring project information and expect users to learn from them. The problem is that human can only learn from others by matching to one’s own experience, and the knowledge level or even knowledge context between expert and learner are always not the same. Traditional knowledge engineering method usually doesn’t take project context into consideration (e.g. IBIS, QOC), or they neglect the interaction between different project modules (e.g. CommonKADS, DRCS). Therefore, we have to come up with classification models suited within specific contexts to show organisational knowledge in its specific context [20].

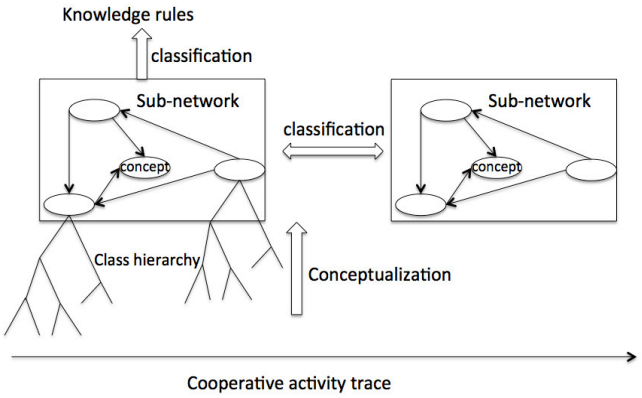


Fig 2. CKD framework

Project memory is be decomposed into smaller modules in order to show project memory in different perspective with different context to provide a better learning angle. The general semantic network of project memory is decomposed into 4 sub-networks:

- Decision-making process: this part represents the core activity of design project, which helps designers to learn from negotiation and decision-making experience.
- Project organisation makes decision: this part represents interaction between organisation and decision, which provides an organisational view of decision-making.
- Project organisation realises project: this part represents arrangement of task and project team organisation, which focuses learning on project management.
- Decision-making and project realisation: this part represents the mutual influence between decision and project realisation, which reveals part of work environment and background.

In each project memory module, a sub-network is built with concepts and relations. These project memory concepts are identified based on the research on engineering design and knowledge representation method for design activities [11] [21] [22] [23]. These concepts are employed and rearranged to represent the elements in project memory. Foundational ontologies serve as a starting point for building

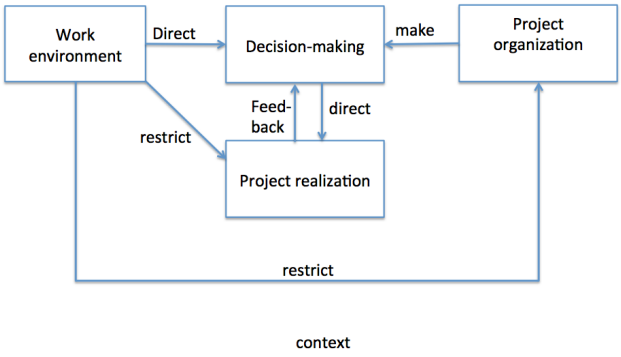


Fig 3. Project memory modules

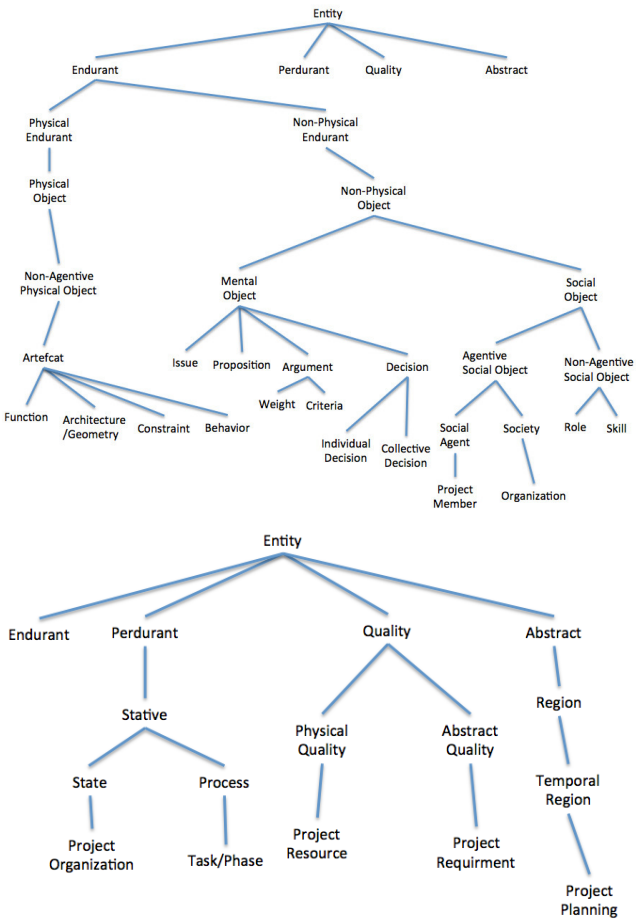


Fig 4. Design project concept aligned with dolce ontology

new domain and application ontologies, provide a reference point for different ontological approaches and create a framework for analysing, harmonising and integrating existing ontologies and metadata [24]. The project memory concepts are aligned with the general Dolce ontology.

Based on these concepts, we are going to build our sub-networks to represent especially interactions between concepts in order to show the cooperative knowledge.

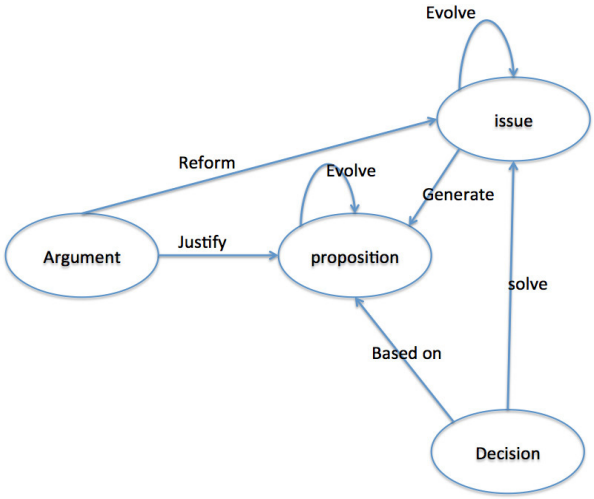


Fig 5. Decision-making sub-network

The first part of project memory is design rational; decision-making process is one of the most important parts in project memory. It contains negotiation process, decision and arguments that can reveal decision-making context. Concepts that are identified in a decision-making process are: issue, proposition, argument and decision. Issue is the major question or problem that we need to address, it can be about product design, organisation arrangement or project realisation etc.; proposition is solution proposed to solve issue by project team member; argument evaluates the proposition by supporting or objecting it, which can push proposal to evolve into another version [4] [23] [25]; argument can also aims at issue which can possibly modify the specification of the issue. Propositions are considered to be possible solutions for issue, and arguments are supposed to explain the reason why. Decision is made by selecting some of the propositions for the issue and setting up a goal for next step of project realisation. Figure 5 shows the decision-making process sub-network.

One of the most important and useful knowledge that we want to represent is the context of design rationale (Moran et al, 1996). This sub-network shows an interaction schema of concepts in decision-making process. Moreover, other project memory modules can also have mutual influences with decision-making process module. Therefore, we connect decision-making to project realisation to show consequences of decision and connect decision-making to project organisation to reveal an organisational influence.

In the sub-network below (figure 6), we want to find a concept that serves as a bridge to connect project organisation and decision-making process. So the concept “member” is introduced into decision-making sub-network to add an organisational dimension into decision-making process. Member is an important concept of project organisation that links to competence, role and task.

The sub-network in figure 7 offers a learning perspective on project realisation with an organisational dimension. Il presents us the interaction schema between task and project organisation. Task is linked to two important attributes of project member: competence and role.

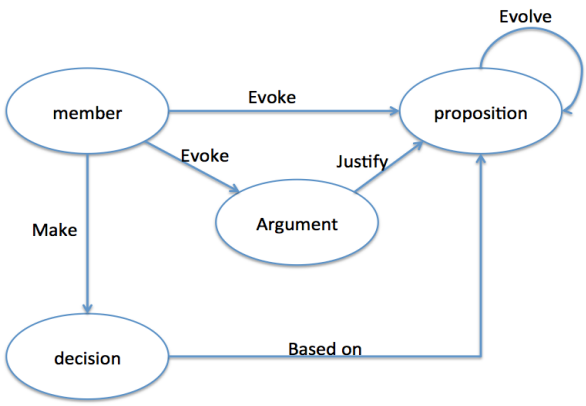


Fig 6. Project organization makes decision

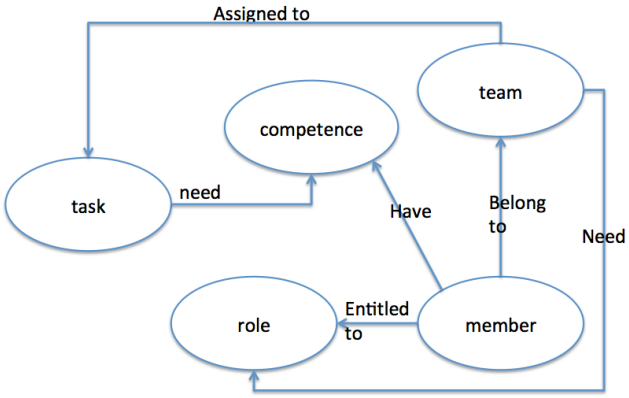


Fig 7. Project organization realizes a project

At last, we want to represent the triangle between task, decision and issue in order to show a mutual influence of task arrangement and decision-making process. A decision sets up a goal for a task; another issue can be evoked during a task, which initiate another decision-making process. The triangle ends by achieving the final result of a task. During a product design, the result of a task can be a new version of a product, and the version of product evolves between decision-making meeting and tasks.

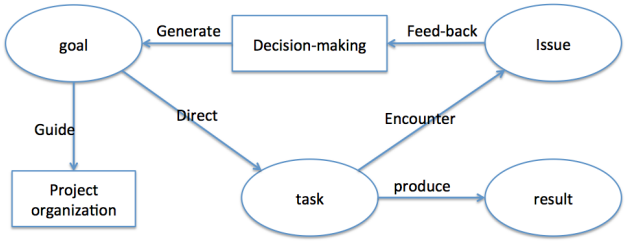


Fig 8. Mutual influence between decision-making and project realization

C. *Knowledge Classification*

The ability to extract general information from example sets is a fundamental characteristic of knowledge acquisition. Machine learning technique is now a hot topic at present, it can figure out how to perform important tasks by generalising from examples. One of the most mature and widely used algorithms is classification [26]. However, design project information are usually not voluminous and quite distinctive; they are highly structured in a computer-aided design environment. Due to these particular characteristics of design project information, present machine learning techniques are not suitable for design project memory classification. We studied four major categories of machine learning algorithms: statistical methods, decision tree, rule based methods and artificial neural network [27] [28] [29] [30]. These methods are not considered for two reasons: 1). Classification process is not transparent to human interpretation. 2). A large recursive training set is needed for classification. The advantage of our classification model in project memory is that it is guided by

semantic networks that indicate knowledge rules resided in interaction schemas. Therefore, according to these semantic networks, we classify interaction schemas instead of concepts. The amount of repetitive interaction schemas is significantly fewer compared to a concept; a large set of instances can be conceptualised into one class, while the probability of similar interaction schemas between concepts is much less. Additionally, the learning process will not ignore non-recursive schemas; on the contrary, they will be put aside as explorative attempts with an explanation.

Two tablet applications have been developed to capture project traces. They can register meeting information and generate XML files (Matta et al, 2013). Project information will be structured according to a XML schema as follow:

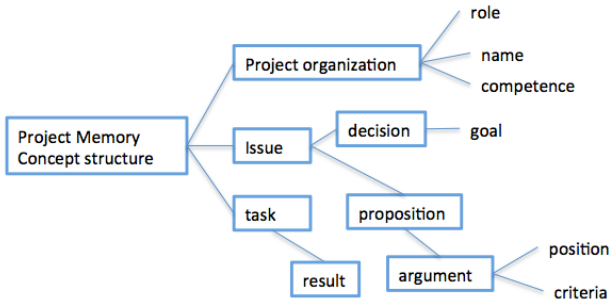


Fig 9. XML schema of project memory

```
<xs:element name="member">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="role" type="xs:string" />
      <xs:element name="competence" type="xs:string" />
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="issue">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="decision" type="xs:string">
      <xs:element name="proposition">
        <xs:complexType>
          <xs:sequence>
            <xs:element name="argument">
              <xs:complexType>
                <xs:sequence>
              <xs:element name="criteria" type="xs:string" />
              <xs:element name="position" type="xs:int" />
                </xs:sequence>
              </xs:complexType>
            </xs:element>
          </xs:sequence>
        </xs:complexType>
      </xs:element>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="task">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="result" type="xs:string" />
    </xs:sequence>
  </xs:complexType>
      </xs:element>
```

Then project information will be classified according to different views to extract knowledge rules. Here we propose three classification views:

1. Problem-solving view: at a specific project phase, we can classify decision-making process for one particular issue. Solutions that are repetitive will be classified as essential solutions, the solutions that are distinctive will be considered as explorative attempt with its precondition as an explanation.

If

$(\text{decision}(d_1) \wedge \ldots \wedge \text{decision}(d_n)) \wedge \text{issue}(i_i) \Rightarrow \text{decision}(d') \wedge \text{issue}(i_i)$,

then

$\text{decision}(d') \wedge \text{issue}(i_i) \Rightarrow \text{essential}(e_i) \wedge \text{issue}(i_i)$

2. Cooperation view: an important subject that we tried to study in our model is cooperation. This classification view allows us to verify whether there are parallel tasks that involve cooperative design or regular meetings concerning whole project team. Projects that are not undertaken concurrently can lead to unsatisfactory results, e.g. solution duplication or excess of project constraint. This rule will reveal the influence of concurrent design on project result.

If

$\exists(\text{issue}(i) \wedge \text{entire\_team}(m)) \wedge \exists(\text{task}(t_1) \wedge \ldots \wedge \text{task}(t_n))$,

then

$\exists \text{cooperation}(m)$

3. Management view: this classification view will focus on project organisation influence on different project memory modules. For example, we can classify project realisation with an organisational dimension to examine how project organisation arrangement can influence project realisation.

A weight factor that indicates recurrence rate will be attributed to each classification result to show the importance of this result. The three aspects proposed above

are the most interesting and practical classification views that we find so far, however we do not exclude the possibility that more useful classification views exist. In the next section, CKD according to these three views will be applied to two example projects.

V. EXAMPLE AND RESULT

Two software design projects were undertaken by two teams in the year 2012 and 2013. The group members are students majored in computer science or mechanic design. The goal of the project is to design a tablet application, which proposes solutions for product maintenance; it should allow a technician to access and modify PLM and ERP information in order to facilitate information flow in supply chain. MMreport and MMrecord were employed to keep track of meetings from the beginning to the end of the project, they can be downloaded in App Store for free. XML documents were generated by these two applications. We analysed these XML documents as well as other documents (email, forum discussion and result) according to the XML schema above. Next we are going to demonstrate three rules extracted by comparison between these two projects.

A problem-solving rule on the issue "function definition" can be extracted by comparing the decision-making process on this issue of both projects. We classify repetitive solutions as essential solutions for the issue function definition, and distinctive solutions as explorative cases with a precondition. The detailed classification is shown in figure 10.

Cooperation rules on this project can be extracted by classifying project planning, which is represented by the sub-network decision-making process and project realisation. If there are tasks concern module integration and

Tablet application for product maintenance				
Year	2012		2013	
Issue	Function definition		Function definition	
Negotiation	Proposition	Argument	Proposition	Argument
	Automatic object reconnaissance	• More efficient • Help operator with little mechanical knowledge • More expensive • Technology obstacle	Manuel object search engine	• Easy to design • Require operator to have certain mechanical knowledge
	ERP and PLM connection	• Reduce data redundancy • Technology obstacle	Tablet connection with PLM and ERP	
	Tablet connection with ERP and PLM			
Decision	• Automatic object reconnaissance • Tablet connection with ERP and PLM		• Manuel object search engine • Tablet connection with ERP and PLM	

Tablet application for product maintenance		
Issue	Function definition	
Essential solutions	Tablet connection with PLM and ERP, object search with tablet applications	
Conditional solutions	Solution	Condition
	Automatic object reconnaissance	Enough budget
	PLM and ERP connection	Feasible technology

Fig 10. Classification on "function definition"

Tablet application for product maintenance				
Year	2012		2013	
Phase	Project realization		Project realization	
Project organization	Three sub-groups for each application module (ERP,PLM, Object reconnaissance)	Competence distribution: ERP(computer science) PLM (mechanical design) Object reconnaissance (computer science)	Three sub-groups for each application module (ERP, PLM, object search engine)	ERP(computer science) PLM (computer science, mechanical design) Object search engine (computer science, mechanical design)
Project planning	• 4 working meetings inside each sub-group to validate project specification • A final meeting to simply collect each sub-group's work		• 12 work meetings of whole project team • Sub-group meetings are organized freely	
Result	• Each module has its own database, the application has 3 databases in total • Automatic image recognition increase the cost drastically		• Client-server architecture that requires only one database • Centralized data management	

Fig 11. Classification on project planning with organizational influence

regular meetings on project specification of whole project team, then this project is undertaken concurrently. If no meetings are held with the whole group or no integration task is assigned to more than one sub-group, then this project is considered failed at concurrent design. We can see from the project information 2012, four meetings were held inside each sub-group and only one final meeting involved the entire project group, but the issue of the final meeting was "collecting each group's work", which means no integration issue was dealt with. Apparently in the project 2012, design activities were not organised concurrently, which leads to the result "database duplication" and "expensive project cost".

Linear project planning leads to bad communication between different sub-group designers, which result in poor integration design. From the management point of view, we can further this classification by adding an organisational dimension to project planning. These two classification is shown in figure 11.

By comparing these two project organisations, we can see that in the project team 2012, competence was distributed homogeneously for each group, members were divided into computer science group and mechanical design group; whereas competence was paired in the project team 2013, computer science and mechanical design both exist in each sub-group. From this classification view, we may draw the conclusion that if designers with different skills are assigned to the same task, project tends to be carried out more concurrently, which leads to a more satisfactory result.

Extraction of these rules are all guided by comparison of structured information according to different project views, rules may change as more project information will be captured. CDK classification will progress in a cumulative manner.

VI. CONCLUSION AND PERSPECTIVE

This paper presented our research work on cooperative knowledge, especially on how to discovery cooperative knowledge in order to reuse them. A CKD method was proposed for this purpose in design project field. It is a knowledge classification guided by semantic network

schemas. Instead of classifying domain expert knowledge, interaction schemas between concepts were classified; it allows us to put each important concept in its interactive context. A CKD classification is semantically expressive and comprehensible by users. Therefore, it is up to users to choose which classification view to use for knowledge extraction. We tested CKD method on two example projects, which shows that cooperative knowledge can be extracted by interaction schema classification, more importantly, the knowledge rules extracted can be quite useful for learning purpose.

No classification can be argued to be a representation of the true structure of knowledge, the design project knowledge classification showed in this paper is a application field of CKD method, class conceptualisation, semantic network structure and knowledge classification views are strictly linked to design project context. In other words, a CKD classification model should be built according to application domain features. In order to enrich this application, we will try to formalise classification rules with programming languages and test our model on more complicated projects.

REFERENCES

[1] R K. Schmidt, L. Bannon, 1992. Taking CSCW seriously, Computer Supported Cooperative Work (CSCW) ,pp. 7-40.

[2] M. Zacklad, 2003. Communities of action: a cognitive and social approach to the design of CSCW systems. In Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work, pp. 190-197.

[3] S. Khoshafian, M. Buckiewicz, 1995. Introduction to Groupware, Workflow and Workgroup Computing. John Wiley & Sons, Inc., New York, NY, USA.

[4] S. Buckingham Shumm , 1997. Representing Hard-to-Formalise, Contextualised, Multidisciplinary, Organisational Knowledge, in AAI Spring Symposium on Artificial Intelligence in Knowledge Management, , pp. 9–16.

[5] I. Nonaka , H. Takeuchi, “The knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation”, Oxford University Press, 1995

[6] A. Newell, 1982. "The knowledge level.” *Artificial intelligence* 18, no. 1 pp: 87-127.

[7] G. Ducellier, N. Matta, Y. Charlot, and F. Tribouillois, 2013. "Traceability and structuring of cooperative Knowledge in design using PLM." *International Journal of Knowledge Management Research and Practices* 11, no. 4 pp: 20.

[8] M. P. V Easterby-Smith, M. Lyles, 2003. “The Blackwell Handbook of Organizational Learning and Knowledge Management.,” Adm. Sci. Q., vol. 48, p. 676.

[9] T. R. Gruber, 1995. "Toward principles for the design of ontologies used for knowledge sharing?", *International journal of human-computer studies,* Vol.43, No.5, pp 907-928.

[10] D. Fensel, 2000. "Ontologies: A silver bullet for Knowledge Management and Electronic-Commerce." Berlin: Spring-Verlag.

[11] G. Pahl, W. Beitz, J. Feldhusen, K.H. Grote, 2007. Engineering design: a systematic approach, pp.1-617.

[12] G. Ducellier, 2008. Thèse aux plateformes PLM, Univ. Troyes, France, 2008.

[13] O. Castillo-Navetty, N. Matta, 2005. "Definition of a practical learning system," *Information Technology Based Higher Education and Training, 2005. ITHET 2005. 6th International Conference on,* vol., no., pp.T4A/1,T4A/6, 7-9.

[14] C. Djaiz, N. Matta, 2006. "Project situations aggregation to identify cooperative problem solving strategies." In *Knowledge-Based Intelligent Information and Engineering Systems,* pp. 687-697. Springer Berlin Heidelberg

[15] S. Bekhti, N. Matta, 2003. "Project memory: An approach of modelling and reusing the context and the design rationale", *Proc. Of IJCAI,* Vol. 3.

[16] N. Matta, M. Ribière, O. Corby, M. Lewkowicz, M. Zacklad, 2001. “Project Memory in Design,” in *Industrial Knolwedge Management,* London Springer, pp. 147-162.

[17] N. Matta, G. Ducellier, 2013. “An approach to keep track of project knowledge in design,” *Proc. IC3K/KMIS, 5th International Conference on Knowledge Management and Information Sharing,* Vilamoura Algarve, Portugal.

[18] H. Cohen, C. Lefebvre, eds, 2005.“Handbook of categorization in cognitive science”, Vol.4, No.9.1, Elsevier, Amsterdam.

[19] J. F. Sowa, 2000.Knowledge representation: logical, philosophical, and computational foundations, Brooks/Cole, Pacific Grove.

[20] J. Mai, 2004. "Classification in context: relativity, reality, and representation", *Knowledge organization,* Vol.31, No.1, pp 39-48.

[21] M. Klein, 1993. “Capturing design rationale in concurrent engineering teams,” *Computer , Calif.,* vol. 26, no. 1, pp. 39–47.

[22] G. Schreiber, B. Wielinga, 1994. Van de Velde W., Anjewierden A., “CML: The CommonKADS Conceptual Modelling Language”, Proceedings of EKAW'94, *Lecture Notes in AI N.867,* L.Steels, G. Schreiber, W.Van de Velde (Eds), Bonn: SpringerVerlag ,pp 1-25.

[23] J. Conklin, M. L. Begeman, 1988. “gIBIS: a hypertext tool for exploratory policy discussion,” *ACM Transactions on Information Systems,* vol. 6., pp. 303–331.

[24] P. Mika, D. Oberle, A. Gangemi, M. Sabou, 2004. "Foundations for service ontologies: aligning OWL-S to dolce." WWW pp. 563-572.

[25] T. P. Moran, J.M. Carroll, eds, 1996. Design rationale: concepts, techniques and use, Routledge, US.

[26] P. Domingos, 2012. “A few useful things to know about machine learning,” *Commun. ACM,* vol. 55, no. 10, p. 78.

[27] R. D. King, F. Cao, A. Sutherland, 1995. "Statlog: comparison of classification algorithms on large real-world problems", *Applied Artificial Intelligence an International Journal,* Vol. 9, No. 3, pp. 289-333.

[28] T. G. Dietterich, 1997. "Machine-learning research", *AI magazine,* Vol.18, No.4, pp 97.

[29] R. M. Goodman, P. Smyth, 1992. "An information theoretic approach to rule induction from databases," *Knowledge and Data Engineering, IEEE transactions,* Vol.4, No. 4 , pp 301-316.

[30] D. Michie , D. J. Spiegelhalter, and C. Taylor. "Machine learning, neural and statistical classification." 1994.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1371–1379
DOI: 10.15439/2014F39
ACSIS, Vol. 2

Social Media And Emotions In Organisational Knowledge Creation

Harri Jalonen
Dr, Adjunct professor, Turku
University of Applied Sciences,
Finland
E-mail: harri.jalonen@turkuamk.fi
Phone: +358 44 907 4964

***Abstract*—Social media increases the connectivity of people inside and outside an organisation. It is not just the implementation of communication technology, but the transformation of working and organisational cultures. The paper presumes that social media provides new opportunities to the organisational knowledge creation process by amplifying knowledge created by individuals as well as crystallising and connecting it to an organisation's knowledge system. The process depends fundamentally on the individual's tacit knowledge and its conversion into organisational explicit knowledge. Knowledge conversion is not a linear and sequential process, but a process which is affected by the individual's emotions. This paper explores the interplay between knowledge and emotions in the organisational knowledge creation process in the context of social media. The paper concludes that knowledge and emotion shared in social media contribute to the social identity, which increases the odds of altruistic behaviour towards others in a way that benefits the organisation.**

I. Introduction

Social media increases the connectivity of people within and across organisational boundaries. It provides new opportunities for acquiring and sharing information to be exploited in strategic decision-making and leadership, innovation, marketing and customer service, and organisational communication. It has been suggested that social media revolutionises the ways information and knowledge are managed in organisations [1]. An extensive literature argues that information requires interpretation to become knowledge [2, 3, 4]. Consistent with previous studies, this paper supposes that the value of information shared through social media depends on the organisation's ability to use it. This is because "information will only acquire meaning for the organisation when meaning is assigned to that information within the receiving organisation" [5]. Information becomes a valuable resource only when it is interpreted and connected to already existing knowledge. Information is never a "pre-given reality" for organisations, but a process of interpretation by the organisation and its individuals [6, 7].

Nonaka [6] has defined *organizational knowledge creation* as a "process of making available and amplifying knowledge created by individuals as well as crystallizing and connecting it to an organization's knowledge system". Turning information into usable knowledge depends fundamentally on the individual's *tacit knowledge* and its conversion into organisational *explicit knowledge*. Knowledge conversion is not a linear and sequential process, but a process which is affected by the individual's *emotions*. Emotion is not seen as opposite of reason, but a different form of it. Along with knowledge, emotion and intuition have an important role in organisational decision-making [8, 9].

From a technological point of view, knowledge conversion presents a challenge because tacit knowledge is difficult to communicate to others as information. Many studies proclaim that tacit knowledge sharing by information technology (IT) is quite impossible [10, 11]. Johannessen et al. [12], among others, have pointed out that by investing in IT, organisations emphasise explicit knowledge at the expense of tacit knowledge. Paradoxically, the mismanagement of tacit knowledge may yield to deterioration of the organisation's competitive advantage, which is reported to be more dependent on tacit than explicit knowledge [13, 14].

The importance of tacit knowledge and emotion in the organisational knowledge creation process on the one hand, and the rapid growth of social media on the other hand, beg to explore and analyse their interoperability. This is not a trivial question as social media transforms the organisation's social practices and therefore enables *or* stifles organisational knowledge creation. This paper explores the interplay between knowledge and emotion in the organisational knowledge creation process in the context of social media.

II. Social media – activities, means, contents and features

There is no single and universally accepted definition of social media. Typically, it is loosely referred to the means of interaction among people in which they create, share, and exchange information in networks. Social media merges technology, people and contents. When emphasising the actions enabled by social media, one possible approach is to characterise social media as a context for communication, collaboration, connecting, completing and combining (5C). The 5C's categorisation is briefly discussed below. Unless otherwise stated, the discussion contained herein is based on work by Vuori [15].

For *communication* purposes, social media provides new tools to share, store and publish contents, discuss and express opinions and influence. Communication is executed through blogs (e.g. Blogger) and microblogs (e.g. Twitter), podcasts (e.g. iTunes) and videocasts (e.g. YouTube), media sharing systems (e.g. SlideShare), discussion forums (e.g.

978-83-60810-58-3/$25.00

Apple Support Communities) and instant messaging (e.g. Skype). Blogs and microblogs – particularly Twitter – have changed our media landscape rapidly. Schultz et al. [16] have pointed out that blogs and microblogs affect the society in which they play a role not only by the content delivered over the media, but also by the characteristics of the communications media themselves. Seemingly, the medium has become the message, as McLuhan once predicted: "the medium is the message" [17]. In *collaboration*, social media enables collective content creation and edition without location and time constraints. Wikis (e.g. Wikipedia) and shared workspaces (e.g. GoogleDocs) are typical social media applications supporting collaboration. They enable collaborative authoring, empowering the users to create, edit and update contents. Empowering the users reminds the prediction made by Alvin Toffler. Toffler predicted as early as 1980 the rise of a society of prosumers. Toffler identified various forms of prosumers but common for all of them is that the roles of producers and consumers are blurring and merging in a way which inevitably transforms the relationship between inside and outside the organisation. It has been suggested, for example, that firms do not create value for customers anymore but with customers [18]. For *connecting* purposes, social media platforms offer new ways of networking with other people (e.g. Facebook), socialising oneself into the community (e.g. LinkedIn) and creating virtual worlds (e.g. Second Life). Social network sites, especially Facebook, are usually seen as a synonym for social media. This is no wonder, as a conservative estimate is that social networks gather worldwide well over one billion users. Social network sites connect people with similar interests and enable the creation of communities around these interests. In *completing*, social media tools are used to complete content by describing, adding or filtering information, tagging contents, and showing a connection between contents. Commercial completing social media applications are, for example, Pinterest, Google Reader and Digg. *Combining* social media tools are developed for mixing and matching contents. The logic behind these tools is simple: users need versatile tolls which able to combine the contents from different applications. Combined social media sites are typically called as mash-ups meaning "a coherent combination of pre-existing web services that allow a certain user within a platform to use another application, in a specific window, without the need to get out of the initial website" [19]. Google Maps, for example, allows geographically pinpoint the locations of hotels and restaurants, and so on.

Due to technological convergence and users' needs, the categorisation of 5C's is only suggestive. Many social media tools support two or more functionalities. As Vuori [15] has pointed out, Facebook and Twitter, for example, make it possible to embed videos and photographs from another location on the Web, whereas wikis can provide RSS feeds to keep up with updates on a certain article. In addition to functionalities, social media has been approached from the point of view of its characteristics. Typically, user-friendliness, interactiveness, openness and transparency, participation and democracy, uncontrollability, velocity, and real-timeness have been mentioned to be the main characteristics of social media [20, 21, 22, 23].

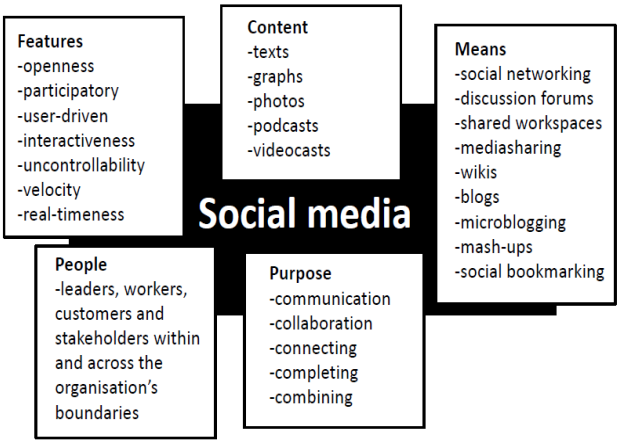


Figure 1. Social media – activities, means, contents, features and actors.

Based on the above mentioned, it is relatively easy to agree with the judgements which claim that social media is not just the implementation of communication technology, but the transformation of working and organisational cultures. Adapting Berthon et al. [24], the transformation can be summarised into three aspects: social media shifts i) the locus of activity from the desktop to the Web, ii) the locus of power from the firm to the collective, and iii) the locus of value production from the firm to the customer. At the heart of this transformation are new possibilities for acquiring, storing, sharing and using information within and across organisational boundaries.

A little pointedly, it can be argued that many behaviours that sociologists study are nowadays taking place in social media. The need to be connected through networked technological applications has become a necessary strategy for postmodern humans [25] and is blurring the boundaries between the reality and virtual and the boundaries between an organisation and its environment. Presumably, it also changes the ways humans deal with knowledge and emotions.

III. Knowledge creation in organisations

Although knowledge has been seen as a virtue to be sought since the days of Plato and Aristotle, it was works by Teece [26] and Nelson & Winter [27] in which knowledge became an important part of organisation science. They were followed by Barney [28] Nonaka [6], Grant [29] and Spender [13], among others, who used knowledge as an explanatory factor for the idiosyncrasies of firms. Barney [28] characterised knowledge as a costly-to-imitate resource and therefore vital for the organisation's competitive advantage. Grant [29] developed ideas further and laid down the premises of "knowledge-based theory of the firm" (KBV). Extending the resource-based view of the firm (RBV), Grant argued that knowledge is not just a generic resource, but the

special one, which is embedded and carried through multiple entities including organisational culture and identity, policies, routines, documents, systems, and employees. The KBV states that heterogeneous knowledge bases and capabilities among firms contribute to competitive advantage and superior performance.

Nonaka [6] contributed to the academic discussion by providing the theory of *organizational knowledge creation*. The theory can be seen as complementing the knowledge-based view of the firm because it explains the dynamic processes of organisational knowledge creation Nonaka et al. [30]. Seemingly Nonaka [6] and many his co-authors thought that knowledge creation is needed to achieve competitiveness through innovation rather than an intrinsic goal per se. Nonaka defined organisational knowledge creation as a dynamic process in which knowledge created by individuals is connected to an organisation's knowledge system. The theory is based on two premises: *tacit* and *explicit knowledge* can be conceptually distinguished along a continuum, and *knowledge conversion* explains the interaction between tacit and explicit knowledge [7].

Tacit knowledge covers knowledge that is "unarticulated and tied to the senses, movement skills, physical experiences, intuition, or implicit rules of thumb" [6]. Knowledge of baking a cake or interpreting the moment of closing a deal are everyday examples of tacit knowledge. In order to succeed in those situations, the baker and the salesperson need tacit knowledge which is not embedded only in individuals, but also encultured in the organisation's practices and procedures [31]. Tacit knowledge differs from "explicit knowledge", which is uttered and captured in documents and stored in certain media [6]. The information contained in manuals and procedures is a typical example of explicit knowledge. Explicit knowledge is encoded [31] – i.e. it is conveyed in signs and symbols (books, manuals, databases, etc.) and decontextualised into codes of practice. It is accessible through consciousness [30]. Compared to tacit knowledge whose locus is in the knower's mind, explicit knowledge can be readily transmitted within and across organisational boundaries.

Knowledge conversion refers to the process through which "one overcomes the individual boundaries and constraints imposed by information and past learning by acquiring a new context, a new view of the world and new knowledge" [30]. Nonaka [6] proposed that in organisational context, knowledge conversion happens in the Socialisation–Externalisation–Combination–Internalisation process (Fig. 2).

Nonaka & von Krogh [7] positioned the organisational knowledge creation theory as an opposite to "correspondence doctrine", which was based on the idea of "pre-given" reality which exists irrespective of the observer. In so doing, correspondence doctrine also implied that an individual's main task was to improve the representation of pre-given reality by processing information about it. Although gathering information improves the organisation's decision-making, the problem is that it cannot improve the organisation's ability to foster creativity, create opportunities and enable innovation [7]. Knowledge has no value in its own right. The value of knowledge comes from its ability to produce change or action.

	TACIT KNOWLEDGE	EXPLICIT KNOWLEDGE
TACIT KNOWLEDGE	**Socialization** -aims at sharing tacit knowledge between individuals	**Externalization** -aims at articulating tacit knowledge into explicit concepts
EXPLICIT KNOWLEDGE	**Internalization** -aims at embodying explicit knowledge into tacit knowledge	**Combination** -aims at embodying explicit knowledge into tacit knowledge

Figure 2. Knowledge conversion [13].

IV. Emotion in organisational knowledge creation

Conventionally, emotion refers to a feeling state involving thoughts, physiological changes, and an outward expression or behaviour. Emotions are expressed in facial reactions, gestures or postures. The behavioural side of emotion means that emotion has a target at which it is intuitively or intentionally directed. [32, 33] Emotions are typically categorised into six universal "basic emotions", which are happiness, surprise, anger, disgust, sadness and fear [34, 35]. Happiness is a positive emotion, whereas anger, disgust, sadness and fear refer to negative valence of emotion. Surprise, in turn, can be either positive or negative. In some circumstances, individual may feel both positive and negative emotions simultaneously [36]. An individual may, for example, evaluate his/her colleague's success positively, while, at the same time, he/she may feel disappointed about his/her own lot. It should also be noted that positive and negative emotions are not necessarily exclusionary to each other. Based on work of Chang et al. [37], Cameron et al. [36] have pointed out that "exclusive focus on negative emotions cannot allow – even by inference – conclusions about positive emotions". Likewise "the possession of conflicting positive and negative reactions by a person experiencing an emotion need not be confined to negative emotions" [36]. Pride and shame, for example, may be both useful for positively motivating individuals in certain instances [38]. Emotion is also biological response triggered by a particular situation or event [33]. Feeling fear and your heart starts beating faster, and your breathing deepens. There are also emotions which are not responses to "external" past happenings. These are called as "anticipated emotions". They refer "to purposive activities concerning goal-directed behaviour" [36]. When the goal is attained, we feel satisfaction, and when the goal is not achieved, we feel dissatisfaction. The locus of emotions is typically seen resided in the individual, although, some authors have argued for collective emotions [39]. Collective emotion refers to the experience reinforced among the community (in contrast to group-based emotions) of large numbers of individuals [39]. This also explains that emotional

experience varies greatly across cultures [40]. Cacioppo & Gardner [33] summarise the above mentioned nicely by stating that emotion "is a short label for a very broad category of experiential, behavioral, socio-developmental, and biological phenomena".

The relationship between emotion and cognition dates back to the ancient Greeks. An assumption has been that "higher forms of human existence – mentation, rationality, foresight, and decision making – can be hijacked by the pirates of emotion" [33]. Throughout the rationalist tradition, emotions have been belittled. 'Reason' has been treated as good for individuals and societies, whereas 'emotions' have been deemed as detrimental and intimidating. Also within organisational studies, emotions have long been seen as disturbing elements and opposite to rational thinking. Studies, however, have changed the understanding of emotions. This is particularly due to works of Herbert Simon, especially his criticisms of human rationality. According to Simon, cognitive limitations of human mind and complexity of situations, the choices made by individuals, are rational only in relation to their own mental models. In other words, individuals' rational actions are limited by irrational elements. Simon [41, 42] has called this kind of rationality as "bounded rationality".

Recent psychological research has questioned the notion of emotions just primitive reflexes. Berntson et al. [43], for example, have found that emotions are more than "disruptive force in rational thought". It has been argued that emotion contributes intelligence [44] and improves decision-making [8, 9]. The opposite of reason is not emotion, but lack of reason [44]. Nowadays, emotion is seen in organisational research as a key resource in (rational) decision-making. It has been shown, for example, that consumers' behaviour and decisions are affected by emotions [45, 46]. Furthermore, there is a growing number of management studies which embrace emotional skills as a key part of management practices [47, 48].

Nonaka & Takeuchi [13] defined knowing as a process in which "knowledge is created by the flow of information (messages) anchored in the beliefs and commitment of its holder... Knowledge is essentially related to human action". Given that emotion strives to action, it can be supposed that emotions also play an important role in organisational knowledge creation. One can expect that emotions affect individuals' willingness to expose themselves into the social situations in which knowledge can be created and shared. Individuals who feel comfortable and not threatened are probably more inclined to share knowledge than those individuals who fear conflict of interests among the individuals. Von Krogh [49], for example, has spoken for the importance of "care" and "empathy" in knowledge creation. Estrada et al. [50] and Isen [51], among others, have shown that positive emotion enhances innovation. Moreover, Isen & Baron [51] have found that positive mood state generally encourages the display of helping behaviour in organisations. It has also been shown in several studies that leaders' mood states affect their followers [52, 53, 54]. Leaders can shape the arousal of their subordinates and hence contribute to organisational knowledge creation.

In addition to positive consequences, emotions can be negative inhibiting organisational knowledge creation. This is the case, for example, when individuals fear that their knowledge can somehow or other be used against them. It is not at all rare, that pure envy, jealousy or anger create an obstacle for knowledge sharing. A bit more rationally, but still very much emotionally, motivated is knowledge hoarding, which happens when an individual prefers to maximise his/her personal pay-off instead of the interest of the organisation [55]. One emotional reason for knowledge hoarding is leaders' behaviour. Leaders who show negative emotions, such as anger and sadness, evoke less enthusiasm within their subordinates [56. Supposedly, individuals who lack of enthusiasm are reluctant to devote their time to participate in knowledge creating and sharing activities within the organisation.

As mentioned before, positive emotion (e.g. happiness) does not necessarily produce positive behaviour or negative emotion (e.g. shame) negative behaviour. This is because individuals are goal oriented. In goal-directed behaviour, individuals "anticipate" what emotions both the goal success and goal failure will awake [57]. According to Bagozzi [57], anticipated positive emotions energise volitions positively (because goal achievement is desirable and motivating), and anticipated negative emotions also energise volitions positively (because goal achievement is a remedy for the bad feeling). The key question, therefore, is how emotions can be "managed" in an organisation in a way which contributes to knowledge creation and sharing.

V. Creating knowledge and sharing emotions through social media

Defining knowledge creation as a dynamic and emotionally affected process of amplifying an individual's knowledge and connecting it to an organisation's knowledge asset, begs the question where does the process take place? Nonaka et al. [58] have introduced a concept of 'Ba' referring it to the context for knowledge sharing, creating and utilising. 'Ba' means not just a physical space, but a specific time and space. Combining physical, virtual and mental spaces, 'ba' provides energy, quality and place to perform not only knowledge sharing activities, but also interpretation of ambiguous information cues [58]. Nonaka et al. [58] have identified four types of 'ba', which are originating 'ba', dialoguing 'ba', systemising 'ba' and exercising 'ba'. The 'ba's are defined by two dimensions (Fig. 3).

Each 'ba' offers a context for specific step in the knowledge-creating process [58]. In the *originating ba*, individuals meet face-to-face sharing emotions, feelings and experiences. It represents socialisation among individuals. Originating 'ba' helps an individual to transcend "the boundary between self and others, by sympathising or empathising with others" [58]. In the *dialoguing ba* individuals' tacit knowledge is shared and articulated through dialogues amongst participants. It differs from originating 'ba', as the dialoguing 'ba' is more consciously constructed. It offers a context for externalisation. *Systemising ba* is defined by collective and virtual interactions. It offers a context for the

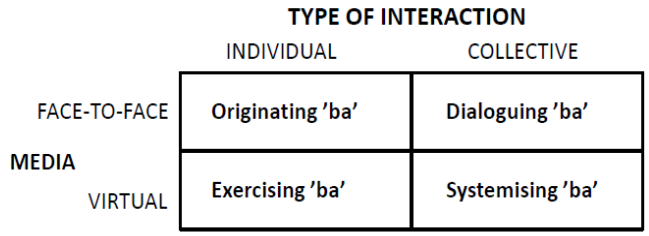


Figure 3. Four types of 'ba' [58]

combination of existing explicit knowledge, as explicit knowledge can be relatively easily transmitted to a large number of people in written form [58]. In the *exercising ba*, individuals embody explicit knowledge that is communicated. Exercising 'ba' offers a context for internalisation.

The classification of 'ba's can be analysed through *media richness theory*. According to media richness theory, communications media can be differentiated based on their abilities to process information and convey meaning [59]. A communications medium is rich when it contains not only factual information, but also provides multiple cues via non-verbal communication and allows immediate feedback. Face-to-face is typically classified as the richest communications medium, whereas numeric documents are seen the less rich [59]. Seemingly, based on the media richness theory, the originating 'ba' and dialoguing 'ba' are rich media as they provide a context for sharing emotions, feelings and experiences. In contrast, as suggested in the introduction, information technology is typically seen quite poor for sharing tacit knowledge. In Nonaka's et al. [58] model, information technology is important particularly in systemising 'ba' and exercising 'ba' because it offers a virtual collaborative environment and an effective way to share explicit knowledge.

Uncertainty and *equivocality* constitute two epistemological forces which exist in organisations that influence information and knowledge processing. The usefulness of IT in reducing uncertainty is well reported in several studies [60]. However, one cannot say the same when it comes to relation between IT and equivocality. It has been shown that IT is a poor medium to share implicit knowledge and emotions as it does not convey important social cues such as body language [61, 62]. In addition, the lack of synchronicity and immediacy inhibits the establishment of mutual understanding to comprehend conversation and knowledge contribution [63].

However, since the advent of social media, application areas of technology-based interaction have significantly expanded. Social media goes beyond systemising and exercising 'ba's by capturing features also from originating and dialoguing 'ba's. Potentially, it offers a context for connections which enable both increasing the amount of available information – i.e. helping to deal with uncertainty – and achieving shared meanings – i.e. helping to deal with equivocality.

This paper argues that enabling organisational knowledge creation through social media, three aspects are especially important. Firstly, social media means a new kind of context, which can be used not only for sharing explicit knowledge but also for making tacit knowledge visible. The argument is based on Michael Polanyi's [64] distinction of tacit knowledge into two separate forms, namely, *proximal* and *distal* [65]. Proximal refers to the thing that is closer to us, while distal thing is further away. This is what is meant with the statement "we know more than we can tell". For an R&D engineer, for example, it is impossible to express all her/his knowledge about the process of new product development. In the words of Polanyi, she/he is not able to "identify particularities" related to product development. However, experienced engineers "know these particularities, without becoming able to identify them" [64]. An R&D engineer's interests (e.g. why a new product is needed, how and who is going to be using it, how success in new product development affects her/his career) and her/his knowledge about R&D techniques, procedures and processes constitute the proximal dimension of his/her tacit knowledge. Given that R&D is typically a knowledge-intensive and often quite complex process, it is expected that the process involves activities which are difficult to communicate outsiders. Searching information, sharing knowledge and assessing others' ideas, to name a few 'innovation activities', represent distal dimension of individuals' (e.g. an R&D engineer) tacit knowledge. By conducting R&D activities, an engineer explicates her/his interests, "without becoming able to identify them" [64]. Social media offers accessibility of an individual's tacit knowledge through features such as keyword searches, personalised content feeds, blogging and microblogging, social bookmarking and mash-ups that identified content relevant to the user. Worth noting is that information communicated through social media is not restricted to pre-given and intended audience, but it might be "overheard" by others, who, in turn, may participate in knowledge creation.

Secondly, emotions cannot be ignored in organisational knowledge creation. Many studies have shown that social media can modulate human *collective emotional* states both in good and bad [66, 67, 68, 69]. Tadic et al. [68], for example, have found out that what happens in social media cannot be explained through real-world events. Seemingly, social media promotes idiosyncratic non-linear dynamics in which individuals contribute to building up a social network, which then propagates the contents of future messages (information *and* emotion), which often escalates into "bursts of emotional messages that involve many users" [67, 68]. Social media offers opportunities for spreading emotionally motivated information in a way which cannot be controlled by the organisation: "everything that can be exposed will be exposed – for all intents and purposes" [23]. This was the case faced by United in 2009 when it failed to appease Canadian amateur musician whose guitar was mishandled by the airline company. Musician posted a YouTube video entitled "United Break Guitars", which became hugely popular, within couple of weeks it was viewed over 3.5 million times. According to Hemsley & Mason [1], United was ill-prepared to deal with "a fast moving story" what became "a symbol of a lone person trying to deal with a large, uncaring corporation". "United Breaks Guitars" is an example of unpleasant event. However, there is nothing like natural law, which ordains that the content of social media is biased to negative emotions.

Thelwall et al. [70], for example, have founded that two thirds of the comments of social network site (Myspace) expressed positive emotion, while only one fifths contained negative emotion. Based on the above mentioned, this paper argues that the organisations should encourage behaviours that induce the emergence of positive collective emotions. In this respect, Nonaka's & Takeuchi's [13] notions of active empathy, leniency in judgement and trust as enabling conditions are extremely relevant also in social media. Collective emotions have been identified as important elements in developing a sense of online community [71]. Worth noting is that, collective positive emotions create an opportunity for the organisation to integrate also external stakeholders into knowledge creation. At best, positive emotions enable unintended collaboration and intensify the effect of knowledge spillover – i.e. diffusion of knowledge across organisational and/or sectorial boundaries.

Thirdly, connections enabled by social media do not only change the information flow within and across organisational boundaries but also affect *social identity*. Social identity consists of three components: 1) a cognitive component refers to self-awareness of organisational membership, 2) an emotional component reflects involvement with the organisation, and 3) an evaluative component means value connotations attached to the organisation [36]. Social identity evolves in a process of social identification referring to "perception of oneness with a group of persons" which, in turn, may "lead to the activities that are congruent with the identity" and "reinforce the antecedents of identification" [72]. In so doing, social identification is both the cause and the result. Several studies have suggested that social identity affects organisations' behaviour. However, as suggested by Bagozzi [57] and many others, the relation between social identity and organisational behaviour is not direct, but mediated by emotions. Social identity increases an individual's emotional commitment to the organisation. Bergami & Bagozzi [73], for example, have found that social identification leads to positive emotions toward the organisation, positive emotions from the organisation, and positive self-esteem as a consequence of organisational membership. Most importantly, from this paper's perspective, these personally felt emotions may induce individuals "to perform discretionary acts that are not part of the job description but that benefit other employees directly and the organization indirectly" [57]. Presumably, personal experiences also bring about the change how explicit and, especially, tacit knowledge are managed in the organisation. As social media potentially changes the process of social identification, it can also promote altruistic organisational culture. Altruistic organisational culture may have different manifestations, but one of the most obvious ones is that knowledge creation and sharing are preferred to knowledge hoarding and egoist behaviour. Social media is convenient for organisational identity as it makes individuals' identities visible to others through the conscious or unconscious 'self-disclosure' of subjective information such as feelings, likes, and dislikes [20, 22]. Although the identification of a collective can arise without interaction (as an individual need only perceive himself or herself as psychologically intertwined with the fate of the group, however, this paper argues that social media changes the process of identity formation within organisations. Adapting the concept of symbolic interactions [74], it is argued that social media offers a context for verbal and nonverbal interactions of individuals. Congruently with Weick's [75] thoughts, the meaning is not a given but evolved and emerged from these individual acts. Worth noting is that individuals "cannot not communicate" [76]. Social identity is continuously created and re-created through intentional and conscious *and* unintentional and unconscious interactions inside and outside the organisation.

The relationship between knowledge, emotion and social identity in organisational knowledge creation is presented in the Figure 4.

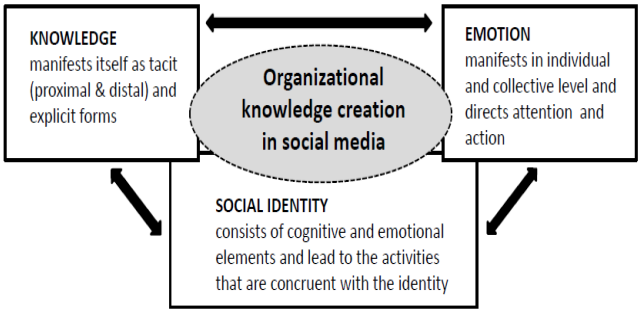


Figure 4. Knowledge, emotion and social identity in organisational knowledge creation in social media.

It is important to notice that the elements of organisational knowledge creation are not independent, but instead related to one another.

VI. Conclusions

This paper has discussed the interplay between (tacit and explicit) knowledge and emotion in the organisational knowledge creation process in the context of social media. The paper argues that social media provides new opportunities to the organisational knowledge creation process amplifying knowledge created by individuals as well as crystallising and connecting it to an organisation's knowledge system. However, the organisational knowledge creation is not a rational process, but a process that entails emotional elements. Emotions influence organisational knowledge creation. This is because emotions act as mediators between social identity and organisational behaviour. Therefore, the paper concludes that knowledge and emotion shared in social media contributes to the social identity, which reduces individuals' risk of being abused, which, in turn, increases the odds of altruistic behaviour that benefits the organisation.

However, as described earlier, social media consists of a bunch of activities, means, contents and features, it is obvious that different social media platforms have different abilities to contribute to organisational knowledge creation. More research is needed into uncovering which platforms are the most useful in organisational knowledge creation.

ACKNOWLEDGEMENT

This paper is part of research project NEMO (Business Value from Negative Emotions) and has been funded by TEKES – The Finnish Funding Agency for Innovation and Turku University of Applied Sciences.

REFERENCES

[1] Hemsley, J. & Mason. R. M. (2013). "Knowledge and knowledge management in the social media age", *Journal of Organizational Computing and Electronic Commerce*, 23(1-2), 138–167. DOI: 10.1080/10919392.2013.748614

[2] Choo, C. (1996). "The Knowing organization: how organizations use information to construct meaning, create knowledge and make decisions", *International Journal of Information Management*, 16(5), 329–340. DOI: 10.1016/0268-4012(96)00020-5

[3] Davenport, T. H. & Prusak, L. (1998). *Working Knowledge*. *How Organizations Manage What They Know.* Harvard University Press, Cambridge, MA.

[4] Tsoukas, H. & Vladimirou, E. (2005). "What is Organizational Knowledge?" In Tsoukas, H. (eds) *Complex Knowledge. Studies in Organizational Epistemology*. Oxford University Press, Oxford, UK.

[5] van Lier, B. (2013). "Luhmann meets Weick: Information interoperability and situational awareness", *E:CO*, 15(1), 71–95.

[6] Nonaka, I. (1994). "A dynamic theory of organizational knowledge creation", *Organization Science*, 5(1), 14–37.

[7] Nonaka, I. & von Krogh, G. (2009). "Tacit Knowledge and Knowledge Conversion: Controversy and Advancement in Organizational Knowledge Creation Theory", *Organization Science*, 20(3), 635–652. DOI: 10.1287/orsc.1080.0412

[8] Kahnemann, D. (2003). "A Perspective on Judgement and Choice Mapping Bounded Rationality", *American Psychologist*, 58(9), 697–720. DOI: 10.1037/0003-066X.58.9.697

[9] Dane, E. & Pratt, M. G. (2007). "Exploring Intuition and Its Role in Managerial Decision Making", *The Academy of Management Review*, 32(1), 33–54. DOI: 10.5465/AMR.2007.23463682

[10] Hislop, D. (2001). "Mission impossible? Communicating and sharing knowledge via information technology", *Journal of Information Technology,* 17(3), 165–177. DOI: 10.1080/02683960210161230

[11] Flanagin, A. J. (2002). "The elusive benefits of the technological support of knowledge management", *Management Communication Quarterly,* 16, 242–248. DOI: 1 0.1177/089331802237237

[12] Johannessen, J. A., Olaisen, J. & Olsen, B. (2001). "Mismanagement of tacit knowledge: The importance of tacit knowledge, the danger of information technology, and what to do about it", *International Journal of Information Management,* 21(1), 3–20. DOI: 10.1016/S0268-4012(00)00047-5

[13] Nonaka, I. & Takeuchi, H. (1995). *The Knowledge Creating Company – How Japanese Companies Create the Dynamics Innovation*. Oxford University Press, Oxford, UK.

[14] Spender, J. C. (1996). "Making knowledge the basis of a dynamic theory of the firm", *Strategic Management Journal*, 17, Special issue: Knowledge and the firm, 45–62. DOI: 10.1002/smj.4250171106

[15] Vuori, V. (2011). *Social Media Changing the Competitive Intelligence Process: Elicitation of Employees' Competitive Knowledge*, Academic Dissertation, Publication 1001, Tampere University of Technology, Tampere.

[16] Schultz, F., Utz, S. & Göritz, A. (2011). "Is the medium the message? Perceptions of and reactions to crisis communication via twitter, blogs and traditional media", *Public Relations Review*, 37(1), 20–27. DOI: 10.1016/j.pubrev.2010.12.001

[17] McLuhan, M. (1964) *Understanding Media: The Extensions of Man*, McGraw Hill, NY.

[18] Bowman, C. & Ambrosini, V. (2000). "Value creation versus value capture: Towards a coherent definition of value in strategy", *British Journal of Management*, 11(1), 1–15. DOI: 10.1111/1467-8551.00147

[19] Bonsón, E. & Flores, F. (2011). "Social media and corporate dialogue: The response of global financial institutions", *Online Information Review*, 35(1), 34–49. DOI: 10.1108/14684521111113579

[20] Kaplan, A. M. & Haenlain, M. (2010). "Users of the world, unite! The challenges and opportunities of social media", *Business Horizons*, 53, 59–68. DOI: 10.1016/j.bushor.2009.09.003

[21] Denyer, D., Parry, E. & Flowers, P. (2011). "Social", "Open" and "Participative"? Exploring Personal Experiences and Organizational Effects of Enterprise 2.0 Use", *Long Range Planning*, 44, 375–396. DOI: 10.1016/j.lrp.2011.09.007

[22] Kietzmann, J. H., Hermkens, K., McCarthy, I. P. & Silvestere, B. C. (2011). "Social media? Get serious! Understanding the functional blocks of social media", *Business Horizons*, 54(3), 241–251. DOI: 10.1016/j.bushor.2011.01.005

[23] Fournier, S. and Avery, J. (2011). "The uninvited brand", *Business Horizons*, 54(2), 193–207. DOI: 10.1016/j.bushor.2011.01.001

[24] Berthon, P. R., Pitt, L. F., Plangger, K. & Shapiro, D. (2012). "Marketing meets Web 2.0, social media, and creative consumers: Implications for international marketing strategy", *Business Horizons*, 55(3), 261–271. DOI: 10.1016/j.bushor.2012.01.007

[25] Cooper, R. (2005). "Peripheral vision: Relationality", *Organization Studies*, 26, 1689–1710. DOI: 10.1177/0170840605056398

[26] Teece, D. J. (1982). "Towards an economic theory of the multiproduct firm", *Journal of Economic Behavior & Organization*, 3(1), 39–63. DOI: 10.1016/S0742-3322(00)17002-0

[27] Nelson, R. & Winter, S. (1982). *An evolutionary theory of economic change*, Harvard University Press, Cambridge, MA.

[28] Barney, J. (1991). "Firm resources and sustained competitive advantage", *Journal of Management*, 77(1), 99–120. DOI: 10.1177/014920639101700108

[29] Grant, R. (1996). "Toward a knowledge-based theory of the firm", *Strategic Management Journal*, 17, 109–122. DOI: 10.2307/2486994

[30] Nonaka, I., von Krogh, G. & Voelpel, S. (2006). "Organizational knowledge theory: Evolutionary paths and future advances", *Organization Studies*, 27(8), 1179–1208. DOI: 10.1177/0170840606066312

[31] Blackler, F. 1995. "Knowledge, Knowledge Work and Organizations: An Overview and Interpretation", *Organization Studies*, 16(6), 1021–1046. DOI: 10.1177/017084069501600605

[32] Brehm, J. W. (1999). "The intensity of emotion", *Personality and Social Psychology Review*, 3(1), 2–22. DOI: 10.1207/s15327957pspr0301_1

[33] Cacioppo, J. T. & Gardner, W. L. (1999) “Emotion”, *Annual Review of Psychology*, 50, 191–214. DOI: 10.1146/annurev.psych.50.1.191

[34] Ekman, P. (1992a). “An argument for basic emotions”, *Cognition & Emotion,* 6(3-4), 169–200. DOI: 10.1080/02699939208411068

[35] Ekman P. (1992b). ”Are there basic emotions?”, *Psychological Review,* 99, 550–553. DOI: 10.1080/02699931003612114

[36] Cameron, K. S., Dutton, J. E. & Quinn, R. E. (2003). Positive organizational scholarship - Foundations of new discipline, Berrett-Koehler Publishers, San Francisco, US.

[37] Chang, E. C., D'Zurilla, T. J. & Maydeu-Olivares, A. (1994). “Assessing the dimensionality of optimism and pessimism using a multimeasure approach”, *Cognitive Therapy and Research*, 18, 143-160. DOI: 10.1007/BF02357221

[38] Verbeke, W. & Bagozzi, R. B. (2002). “A situational analysis on how salespeople experience and cope with shame and embarrassment”, *Psychology & Marketing*, 19, 713–741. DOI: 10.1002/mar.10032

[39] Bar-Tal, D., Halperin, E. & De Rivera, J. (2007) ”Collective emotions in conflict situations: Societal implications”, *Journal of Social Issues*, 63(2), 441–460. DOI: 10.1111/j.1540-4560.2007.00518.x

[40] Matsumoto, D. & Hwang, H. S. (2012). “Culture and emotion: The integration of biological and cultural contributions“, *Journal of Cross-Cultural Psychology*, 43(1), 91–118. DOI: 10.1177/0022022111420147

[41] Simon, H. A. (1956) “Repy: Surrogates for uncertain decision problems”, *Office of Naval Research*, January 1956.

[42] Simon, H. A. (1982). *Models of Bounded Rationality: Empirically grounded economic reason*, MIT Press.

[43] Berntson, G. G., Boysen, S. T. & Cacioppo, J. T. (1993) “Neurobehavioral organization and the cardinal principle of evaluative bivalence”, *Annals of New York Academy of Science,* 702, 75–102. DOI: 10.1111/j.1749-6632.1993.tb17243.x

[44] Goleman, D. (1995). *Emotional Intelligence*, Bantam Books.

[45] Jacoby, J., Johar, G. V. & Morrin, M. (1998). “Consumer behavior: A Quadrennium”, *Annual Review of Psychology*, 49, 319–344. DOI: 10.1146/annurev.psych.49.1.319

[46] Laros, F. J. M. & Steenkamp, J-B. E. M. (2005). ”Emotions in consumer behavior: a hierarchical approach”, *Journal of Business Research*, 58, 1437–1445. DOI: 10.1016/j.jbusres.2003.09.013

[47] Shepherd, D. A. & Cardon, M. S. (2009). “Negative emotional reactions to project failure and the self-compassion to learn from the experience”, *Journal of Management Studies*, 46(6), 923–949. DOI: 10.1111/j.1467-6486.2009.00821.x

[48] Liu, F. & Maitlis, S. (2013). “Emotional dynamics and strategizing processes: A study of strategic conversations in top team meetings”, *Journal of Management Studies*, 51(2), 202–234. DOI: 10.1111/j.1467-6486.2012.01087.x

[49] von Krogh, G. (1998). “Care in knowledge creation”, *California Management Review*, 40(3), 133-153

[50] Estrada, C. A., Isen, A. M. & Young, M. J. (1997). “Positive affect facilitates integration of information and decreases anchoring in reasoning among physicians”, *Organizational Behavior and Human Decision Processes*, 72, 117–135. DOI: 10.1006/obhd.1997.2734

[51] Isen, A. M. (1999). “Positive affect and creativity”, in Russ, S. (eds) *Affect, Creative Experience, and Psychological Adjustment*, 3–17, Bruner/Mazel, Philadelphia.

[52] Ashforth, B. F. & Humphrey, R. H. (1995). “Emotion in the workplace: A reappraisal”, *Human Relations*, 48, 97–125. DOI: 10.1177/001872679504800201

[53] George, J. M. (1996). “Group affective tone”, in West, M. (eds) *Handbook of Work Group*, Wiley, Sussex, UK.

[54] Conger, J. A. & Kanungo, R. N. (1998). *Charismatic Leadership in Organizations*, Sage, Thousand Oaks.

[55] Kollok, P. (1998). “Social dilemmas: The anatomy of cooperation”, *Annual Review of Sociology*, 22, 183–205. DOI: 10.1146/annurev.soc.24.1.183

[56] Lewis, K. M. (2000). “When leaders display emotion: How followers respond to negative emotional expression of male and female leaders”, *Journal of Organizational Behavior*, 21, 221–234. DOI: 10.1002/(SICI)1099-1379(200003)21:2

[57] Bagozzi, R. B. (2003). "Positive and negative emotions in organizations", in Cameron, K. S., Dutton, J. E. & Quinn, R. E. (eds) Positive organizational scholarship - Foundations of new discipline, 176–183, Berrett-Koehler Publishers, San Francisco, US.

[58] Nonaka, I., Toyoma, R. & Konno, N. (2000). ”SECI, Ba, and leadership: a unified model of dynamic knowledge creation”, *Long Range Planning*, 33, 5–34. DOI: 10.1016/S0024-6301(99)00115-6

[59] Daft, R. L. & Lengel, R. H. (1986). “Organizational information requirements, media richness and structural design”, *Management Science*, 32(5), 554–571. DOI: 10.1287/ mnsc.32.5.554

[60] Alavi, M. & Leidner, D. E. (2001). “Knowledge Management and Knowledge Management System: Conceptual Foundations and Research Issues”, *MIS Quarterly*, 25(1), 107–136. DOI: 10.2307/3250961

[61] Ma, M. & Agarwal, R. (2007). “Through a glass darkly: Information technology design, identity verification, and knowledge contribution in online community”, *Information Systems Research*, 18(1), 42–67. DOI: 10.1287/isre

[62] Wang, J. C. & Chiang, M. J. (2009). “Social interaction and continuance intention in online auctions: A social capital perspective”, *Decision Support Systems*, 47(4), 466–476. DOI: 10.1016/j.dss.2009.04.013

[63] Chou, S-W. (2010). “Why do members contribute knowledge to online communities?”, *Online Information Review*, 24(6), 829–854. DOI: 10.1108/14684521011099360

[64] Polanyi, M. (1966). *The tacit dimension*, Routledge and Kegan Paul, London, UK.

[65] Stenmark, D. (1999). “The tacit knowledge of interests”, in Workshop on Beyond Knowledge Management: Managing Expertise, at the ECSCW 1999.

[66] Schweitzer, F. & Garcia, D. (2010). “An agent-based model of collective emotions in online communities”, *European Physical Journal B*, 77, 533–545. DOI: 10.1140/epjb/e2010-00292-1

[67] Chmiel, A., Sienkiewicz, J., Thelwall, M., Paltoglou, G., Buckley, K., Kappas, A. & Holyst, J. A. (2011). “Collective emotions online and their influence on community life”, *PLoS ONE*, 6(7), 1–9. DOI: 10.1371/journal.pone.0022207

[68] Tadic, B., Gligorijevic, V., Mitrovic, M. & Suvakov, M. (2013). “Co-evolutionary mechanisms of emotional bursts in online social dynamics and networks”, *Entropy*, 15, 5084–5120. DOI: 10.3390/e15125084

[69] Kwon, O., Kim, C-R. & Kim, G. (2013) ”Factors affecting the intensity of emotional expressions in mobile communications”, *Online Information Review*, 37(1), 114–131. DOI: 10.1108/14684521311311667

[70] Thelwall, M., Wilkinson, D. & Uppal, S. (2010). “Data mining emotion in social network communication: Gender differences in MySpace”, *Journal of the American Society*

for Information Science & Technology, 61(1), 190–199. DOI: 10.1002/asi.21180

[71] Martin-Niemi, F. & Greatbanks, R. (2009). "The ba of blogs – Enabling conditions for knowledge conversion in blog communities", *VINE: The Journal of Information and Knowledge Management Issues*, 40(1), 7–23. DOI: 10.1108/03055721011024892

[72] Ashforth, B. E., & Mael, F. (1989). "Social identity and the organization. *Academy of Management Review",* 14, 20–39.

[73] Bergami, M., & Bagozzi, R. P. (2000). "Self-categorization, affective commitment and group self-esteem as distinct aspects of social identity in the organization", *British Journal of Social Psychology,* 39, 555–577. DOI: 10.1348/014466600164633

[74] Ashforth, B. E. (1985) "Climate formation: Issues and extensions", *Academy of Management Review*, 10, 837–847.

[75] Weick, K.E. (1995). *Sensemaking in organizations*, Thousand Oaks, Sage, CA.

[76] Watzlawick, P., Bavelas, J. & Jackson, D. (1967). *The Pragmatics of Human Communication*, W. W. Norton, NY.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1381–1390
DOI: 10.15439/2014F158
ACSIS, Vol. 2

Application of selected classification schemes for fault diagnosis of actuator systems

Mateusz Kalisch, Piotr Przystałka, Anna Timofiejczuk
Silesian University of Technology,
Institute of Fundamentals of Machinery Design,
18a Konarskiego Street,
44-100, Gliwice, Poland,
Telephone: +48 32 237 10 69
Email: {mateusz.kalisch, piotr.przystalka, anna.timofiejczuk}@polsl.pl

***Abstract*—The paper presents the application of various classification schemes for actuator fault diagnosis in industrial systems. The main objective of this study is to compare either single or meta-classification strategies that can be successfully used as reasoning means in off-line as well as on-line diagnostic expert systems. The applied research was conducted on the assumption that only classic and well-practised classification methods would be adopted. The comparison study was carried out within the DAMADICS benchmark problem which provides a popular framework for confronting different approaches in the development of fault diagnosis systems.**

I. Introduction

THE INCREASING complexity of recent industrial objects makes the issue of fault diagnosis one of the most important directions of research in modern automatic control and robotics [1], [2], [3]. Technical systems and processes are required to be safely and reliably operated due to the protection of human life and health, the quality of the environment, as well as the economic interests. It is possible to specify numerous areas of interdependence of human and technical means, where safety plays a key role, such as aircraft, spaceship, automotive, power or chemical industry. The above mentioned factors cause that new developments in control theory such as passive and active fault-tolerant control approaches are more often applied in these areas of the industry [4], [5], [6]. A special attention is currently paid on the second type of the advanced control methodologies, where fault diagnosis methods hold a critical importance. The present state of the art in the field of fault diagnosis shows the really need for development of fault diagnosis expert systems. The goal is to elaborate general-purposes systems with multi-domain knowledge representations and multi-inference engines [7], [8], [9]. Generally, the fault diagnosis can be divided into three steps [10]: fault detection, fault isolation and fault identification. Moreover, each of them can be developed by means of model-free (based on data), model-based and knowledge-based approaches [4]. In this paper the first approach, where experimental data are exploited was discussed. In this kind of methods data that represents normal and faulty situations can be obtained from historical databases or from simulators as well as laboratory stands.

There are many types of classifiers available in the literature, as well as different concepts are introduced [11]. Examples are methods based on the similarity between objects in the feature space, probabilistic methods or methods based on black box models. Generally, the classification problems can be divided into two groups including approaches of the machine learning techniques: supervised learning and unsupervised learning. In the paper, the authors concentrated the attention only on methods belonging to the first group.

Currently, the information fusion and meta-classification problems are recognized as the most important directions of the research in the domain of supervised learning. The main idea in this approach is the application of simple classifiers working together to solve a problem with better results than it can by means of single one or more complicated classifiers. There are a lot of different kinds of information fusion methods, but the most popular are majority voting, weighted voting, boosting, AdaBoost [11]. On the other hand, meta-classifiers are very often used for the same reason that its efficiency is higher, than the efficiency of the best single classifier [12].

The current research trends in developing machine learning methods are focused on ideas of improving the general efficiency of different classification and meta-classification methods. The main directions are concentrated on optimization techniques which are used to tune relevant parameters of the classical methods, e.g. with the use of evolutionary and particle swarm algorithms [13], [14], [15], [16], [17], [18]. A number of results included in the works show the benefits of using these methods. In case of a task of fault detection and isolation the key features of the signals in time or frequency domains are most commonly used. Industrial actuators may be characterized by a very high complexity which affects the large number of measuring signals and their features. Therefore, another approach aimed at improving the efficiency of the classifier, and often also shortening the time of its learning, is to remove irrelevant variables [19]. There are various methods that can be used in this procedure, e.g. forward or backward selection methods, as well as elimination methods based on

978-83-60810-58-3/$25.00

statistical measures. Another group of methods are known as fusion methods such as bagging, boosting, and development of these concepts that is AdaBoost method [20], [21]. These methods are often more effective than simple classifiers but also show some drawbacks. Some advanced concepts were developed to take advantage of positive aspects of classic methods and to eliminate their limitations [22]. There are also attempts to connect together several different methods such as selection of relevant features and usage of boosting into one algorithm [23]. Such approach may lead to the final result that should be better than the results of the methods applied separately.

The paper is organized as follows. In Section I a brief introduction to the problem is given. Section II illustrates the issue of fault diagnosis using the model-free fault detection and isolation methodology. The next section deals with several classification schemes that can be applied to develop fault detection and isolation systems. A case study is included in Section IV. This example shows the comparison research of the classification schemes for creating fault diagnosis system of the benchmark actuator [24] which were elaborated on the basis of the activity of the DAMADICS (Development and Application of Methods for Actuator Diagnosis in Industrial Control Systems) Research Training Network funded by the European Commission. The last section is focused on conclusions.

II. Model-free fault detection and isolation

One of the most often used model-free fault detection and isolation methods is presented in Fig. 1. It can be seen, that faults are detected and distinguished using primary and redundant process variables. In this method two separated classifiers must be created. The first classifier uses the subset of process variables ($U' \cup Y'$) as its input and it is dedicated for generating diagnostic signals (S), whereas the second one has the same set of input variables but its task is to calculate a fault signature (F). This classifier is triggered in case when the diagnostic signal indicates a fault scenario.

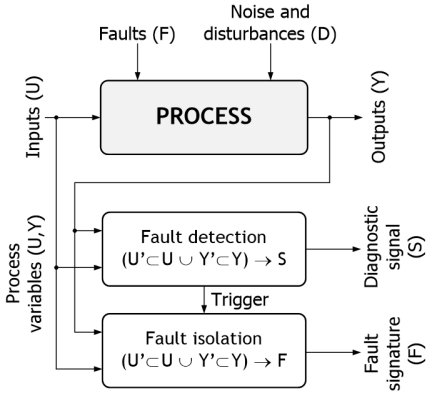


Fig. 1: A diagram of model-free fault detection and isolation

The algorithms corresponding to the diagram presented in Fig. 1 can be designed using different classification methods [10], [4], [25]. Generally, it is possible to apply so-called classical (e.g. decision trees, k-nearest neighbour, naive Bayes, etc.) or soft computing approaches (e.g. neural networks, bayesian networks, fuzzy systems, neuro-fuzzy systems, etc.). The paper deals with the first group of the methods only.

III. Model-free fault diagnosis using different classification schemes

In the next part of the article, model-free fault detection and isolation approaches with the use of different classification schemes were described. As it was mentioned above, these kinds of methods require data (process variables) corresponding to regular (fault-free) and faulty states of the system. In this section, different variants of three basic concepts with a single classifier, meta-classifier and bank of classifiers were applied in order to provide the fault detection and isolation system that is directly based on the process variables.

A. Fault detection

The first concept of the fault detection was presented in Fig. 2 and was elaborated basing on a single classifier, which returns a diagnostic signal corresponding to fault or faultless states of the device. In this method, the process variables were converted by a moving window in order to compute scalar features of the measuring signals. These values were used as input of a single classifier, which generates directly the diagnostic signal. The second detection method was presented in Fig. 3. In this approach a series of two-state classifiers was applied and their task was to determine the degree of the belief for fault detection. The level of belief about faults occurring was a numerical value from 0 to 1. The signal values returned from each classifier were connected to the meta-classifier as its input. The features of the process variables were also connected to the meta-classifier, as the additional input.

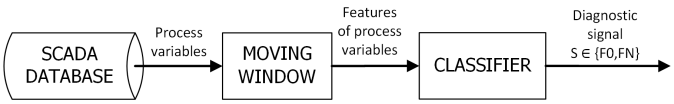


Fig. 2: A scheme of fault detection using the global classifier

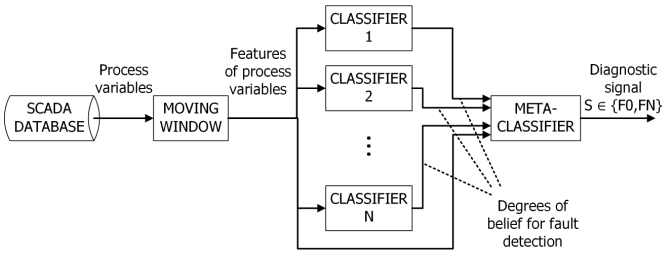


Fig. 3: A scheme of fault detection using the set of various classifiers and meta-classifier

The result of both methods was a diagnostic signal, which indicated fault occurrence. When a classifier or a meta-

classifier detects a fault, the second part of the fault diagnosis system was run in order to isolate the faults.

B. Fault isolation

The first method of fault isolation was comparable to the method that was proposed for the fault detection. It was presented in Fig. 4. As one could see it was a single global classifier. Its task was to determine a type of the fault. Similarly to the previous method, in this case the process variables were calculated in the moving window to obtain scalar features of the measuring signals. The preprocessed signals were connected to the input of a global classifier. This classifier returns a fault signature.

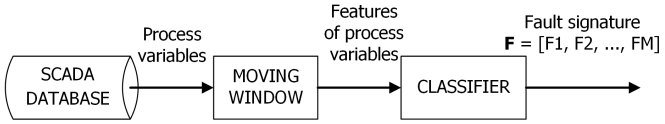


Fig. 4: A scheme of fault isolation using the global classifier

The next fault isolation scheme was presented in Fig. 5. In this approach a set of classifiers of different types was used in order to calculate the degrees of beliefs that were related to fault signatures. These values were given to the input of the meta-classifier and the final decision (fault signature) was obtained.

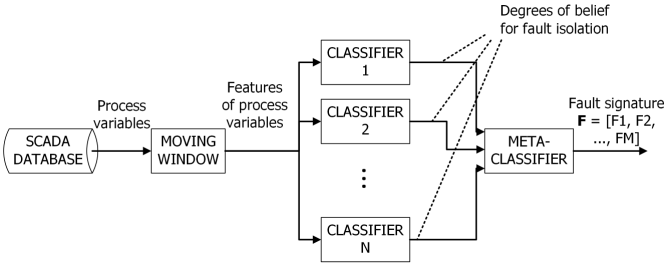


Fig. 5: A scheme of fault isolation using the set of different classifiers and meta-classifier

The last concept of fault isolation was shown in Fig. 6. The main idea was based on a bank of classifiers that were used to calculate degrees of beliefs for specific faults and unknown states of a device. In this case, M single classifiers must be created for M faulty states. Each classifier was dedicated for one state only (it was used for detection one fault solely). In the next step, all available variables (features of the process variables and outputs from base classifiers) are linked to a single dataset. The prepared signals were sent to the input of the meta-classifier which was employed to return the final decision.

C. Used classifiers

The schemes of fault detection and isolation presented in Sections III.A and III.B can be elaborated with use of basic classification methods. The classification problem is possible

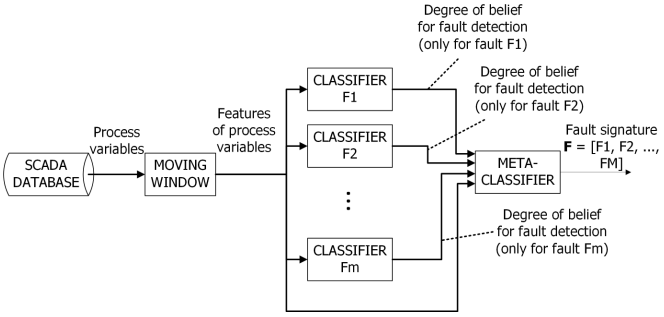


Fig. 6: A scheme of fault isolation using the set of local classifiers (fault detectors) and meta-classifier

to be solved using many known approaches, however, in this research the following methods were applied:

- k - nearest neighbour,
- naive Bayes,
- decision tree,
- rules inductions.

Each of these classifiers returns a label of a chosen class and the degrees of belief for all predicted classes. The best solution is when one of the class is characterised by the belief level equal to 1 and the rest of them are equal to 0. It gives us 100% certainty that a new element should be classified as this particular class. In the next subsections a more precise description of the selected methods was given.

1) k - Nearest neighbour: This is one of the simplest classification techniques. The class label assigned to an example is based on the similarity of this example to one or more prototypes. Typically, the similarity is defined in a geometrical sense using a certain distance. The smaller distance between new object and classified element means the higher similarity between new element and the class represented by the known object. The classifier looks for one nearest neighbour, then it is called 1NN or search for more nearest neighbours, then decision is made by voting. In kNN method, different types of distance measures can be used, e.g. Euclidean (1), Manhattan (2) or Chebyshev (3) distances [26].

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

$$D(x,y) = \sum_{i=1}^{n}|x_i - y_i| \tag{2}$$

$$D(x,y) = \max_{i=1:n}(|x_i - y_i|) \tag{3}$$

The level of confidence about a result of classification depends on a value of k parameter. For k equal to 1, the confidence level for the classification result is always equal to 1. For k greater then 1, when the final result is determined by majority voting, the confidence level depends on the number and types of classified elements in the considered group [27].

2) Naive Bayes [28]: This is a simple probabilistic classification method which is based on bayesian theory. However the naive Bayes classifier considers each of existing features independently.

$$P(d_i|V_1,\cdots,V_n) = \frac{P(V_1,\cdots,V_n|d_i)P(d_i)}{P(V_1,\cdots,V_n)} \quad (4)$$

Taking into account this assumption, the bayesian equation (4) can be transformed to (5), where the denominator of the equation is replaced by a constant C and the conditional probability is calculated by the multiplication.

$$P(d_i|V_1,\cdots,V_n) = C \cdot P(V_1|d_i) \cdot ... \cdot P(V_n|d_i) \cdot P(d_i) \quad (5)$$

The degrees of beliefs for the classification results are equal to probability values obtained from the bayesian equation.

3) Decision tree: This is the classifier based on the tree-like graph created by nodes and connections between them, where the end nodes are called *leafs* and the rest of them have conditions. The result of a decision tree application depends on a chosen leaf. In the algorithm different split evaluation criteria (e.g. ratio gain in C4.5, information gain in ID3, the Gini impurity measure in CART, etc.) can be used [29], [27]. The confidence levels about the classification results are calculated separately for all leafs of the tree during the learning process. Sometimes, when learning data is very complex, the results of the decision tree may be uncertain since some of the leafs may be connected to more then one class. The class which is described by more elements then others (in specific leaf) is chosen as the main class for this leaf. The ratio between the number of elements for available classes is used to calculate the probability for each class in the leaf.

4) Rules induction: The method is based on Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [30] algorithm. The confidence level is calculated in the same way as in the decision tree method.

IV. Verification studies

The proposed schemes of fault detection and isolation were implemented using RapidMiner software. It is an open source software created for solving data mining problems. The verification studies were conducted on data generated using the DAMADICS simulator [31] in order to investigate selected classification schemes. The research problem was actuator fault diagnosis.

A. Benchmark problem

DAMADICS was elaborated for scientists and engineers to simplify the process of evaluating and comparing different methods of fault detection and isolation for industrial systems. In the literature there were available several papers where case study results deal with this problem were presented [32], [33], [34]. The numeric model is used to simulate an electro-pneumatic valve (Fig. 7) which is a part of the production line in Lublin sugar factory in Poland. The model was created in MatLAB/Simulink®software and was on a careful study of the physical phenomena that gave the origin to faults in

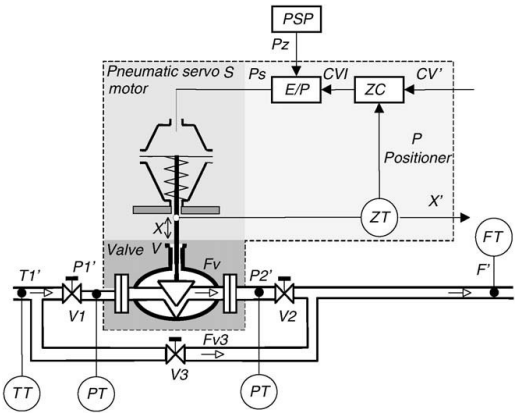


Fig. 7: Structure of benchmark actuator system [24]

the actuator system. This simulator generated the following signals of the process variables:

- CV - process control external signal,
- P1 - inlet pressures on valve,
- P2 - outlet pressures on valve,
- X - valve plug displacement,
- F - main pipeline flow rate,
- T - liquid temperature,
- f - standard diagnostic signal.

All of these signals are normalized to the range between 0 and 1. The DAMADICS simulator allows to choose only one from nineteen available faults (due to this, only scenarios with single faults were taken into account). A part of them is considered only as incipient faults or as abrupt faults (there are three sizes of abrupt faults: small, medium and big) and some of them as both. In this paper the authors decided to investigate only abrupt faults, such as:

- f1 - valve clogging,
- f2 - valve or valve seat sedimentation,
- f7 - medium evaporation or critical flow,
- f8 - twisted servo-motor stem,
- f10 - servomotor diaphragm perforation,
- f11 - servomotor spring fault,
- f12 - electro-pneumatic transducer fault,
- f13 - stem displacement sensor fault,
- f14 - pressure sensor fault,
- f15 - positioner spring fault,
- f16 - positioner supply pressure drop,
- f17 - unexpected pressure change across valve,
- f18 - fully or partly opened bypass valves,
- f19 - flow rate sensor fault.

The list does not have some faults, because the incipient faults such as f3 - *Valve or valve seat erosion* or f4 - *Increase of valve friction* were not considered. The verification tests were performed basing on the process variables generated by the DAMADICS simulator for fault-free and faulty scenarios.

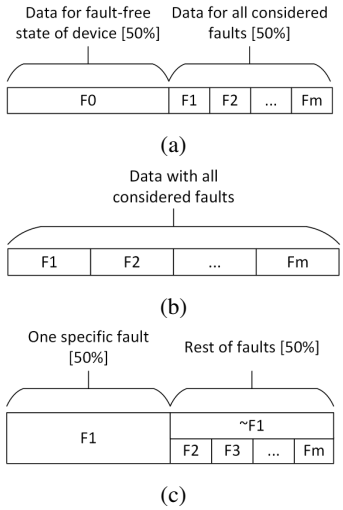


Fig. 8: Examples of data groups for specific methods

B. Data preparation

Collections of data for the training, testing and verification of classifiers were prepared in such a way that the results of classifiers were very similar to classifiers working in a real environment. The process of learning (training and testing) and verification for the applied classifiers was described in this section.

1) Data for fault detection classifiers: Data preparation is a very important part of the classifier learning process. The dataset should be divided into two equal parts, where the first part describes correctly working device (fault-free state) and the second part corresponds to situation when fault occurs (Fig. 8a). It was important to divide the prepared data again to two separated groups (learning group and verification group). For the meta-classifier the number of groups was extended to four, because the first and the second groups were used in the learning and verification process for the base classifier. The other two groups were used for a meta-classifier. In this approach, the size of the dataset for each group was equal 10000 samples, where 5000 samples were prepared from data without faults and rest of them contained data with all considered faults.

2) Data for fault isolation classifiers: The data prepared for the first two fault isolation methods (Fig. 2, 3) consists of characteristic process values for all chosen faults (Fig. 8b). The number of elements for each fault was the same for all sets. For learning and verification process four independent groups of data were prepared (like in fault detection methods, two for base classifiers and two for meta-classifier). The dataset for a single fault for one group contains approximately 600 samples, while the full dataset size is about 8000 samples. The third method of fault isolation requires a different type of data. The initial classifier needs data, where a half of the elements describes an actuator device working with one specific fault and the rest of the elements describe the device working with the other faults (Fig. 8c). In this case, a classifier can generate a two-state signal where the first state defines one specific fault and the other ones are correlated with unknown faults. The size of the dataset in this approach is similar as in case for the method of fault detection. The size of the dataset for the considered fault is equal to 5000 samples and it is equal to the rest of a dataset which contains samples corresponding to other faults.

C. Statistical analysis

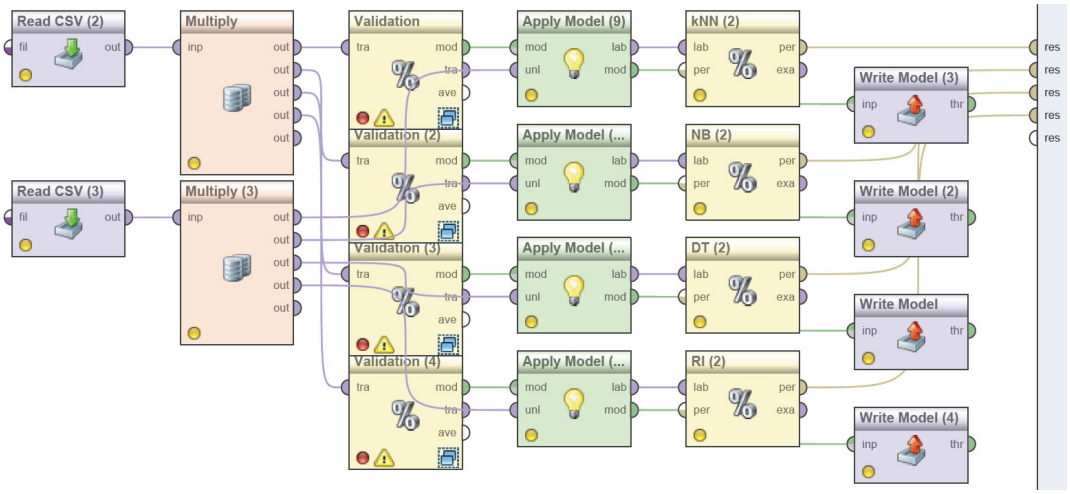
Linear correlation and mutual information analysis were used for choosing relevant process variables and a proper value of the width of a moving window function. In the analysis, all of available process variables were compared between themselves for different device status (e.g. device without faults and with a chosen fault). The results of these tests showed very strong correlations between states F8, F14 and F0. A group of useful process signals was prepared on the basis of results of these tests. Most of the process signals had very difficult character for model-free fault detection and isolation methods. Therefore, the authors decided to apply scalar features of the process variables. Among all available functions in RapidMiner software, four of them were tested: moving average, moving median, maximal value, minimal value.

The scalar features were computed using a moving window of 100 samples width. Such the width value was assumed on the basis of frequency of the harmonic control signal of the valve which was equal to $0,01$ Hz. The authors also studied other values of the window width, however, expected effects in increasing of the efficiencies of the models were not observed. In Fig. 9 exemplary process variables and their features as the time function are plotted. Figures 9a, 9b and 9c show time series of measured signals X, F, P2 for fault F17, whereas figure 9d presents the change in the temperature signal T1 as a result of the fault F7. The sudden change of this signal and its scalar features can be observed at around 800 second when the fault F7 starts affecting the process.

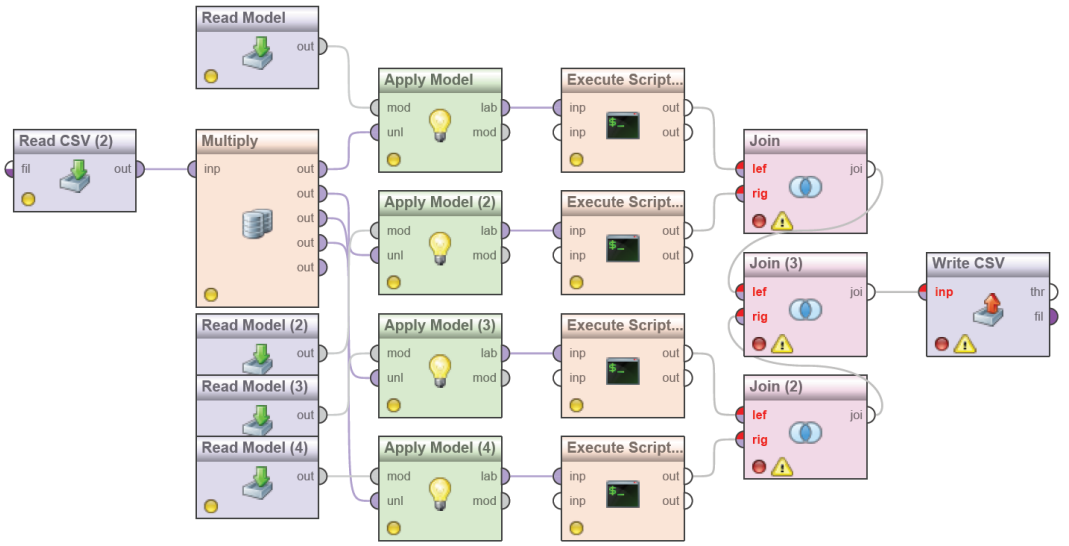
D. Classification schemes implementation

The RapidMiner®software allows to create data mining processes with the use of a visual programming language. This tool gives the opportunity for developing different classification schemes using so-called drag and drop methodology. In this way the classification processes can be viewed as dataflow graphs (Fig. 10).

Fig. 10a presents the scheme of learning and verification processes using four different classifiers. In the first step, the data is red from CSV files by means of *Read CSV* blocks. Next, the learning dataset from the first step is sent to validation blocks, where the learning and evaluation processes for each classifiers is run. The output of these blocks is the ready-to-use classifier, which is applied in *Apply model* block using another dataset read in *Read CSV (2)* block. In the *Performance* block, the process of classifier evaluation is again carried out, but the data are completely different than those in the learning process. *Write Model* blocks are



(a)



(b)

Fig. 10: Selected examples of RapidMiner processes

used in order to write the models into files, because these classifiers are necessary during the process of meta-classifier learning. This scheme can be employed for creating single and meta classifiers. Although, there is a small difference between these two cases in the manner of pre-processing input data. Figure 10b illustrates the process that is applied to generate either learning or verification datasets for meta-classifiers. The data for meta-classifiers is forwarded to their inputs. These samples are totally different from those used at the learning and verification stage of the base classifiers. *Read Model* blocks are employed for reading models of base classifiers. The results of these classifiers are degrees of belief for faultless and faulty states of the actuator. In order to use the data in the next part of the classification process, the pre-processing operations must be done in the *Execute Script* blocks. In the last phase of the process relevant information generated by base classifiers is put together in the form of a coherent dataset which is written in the other file.

E. Results of verification studies

The learning process for the whole set of applied classifiers was conducted using the X-validation method. At firs the case study tests were carried out to choose the most important parameters for X-validation method. The described schemes of fault detection and isolation were examined with all types of classifiers.

1) Results of fault detection: In the first concept of fault detection (see Fig. 2) four single classifiers were compared. In each table the following notation was assumed: kNN - k-nearest neighbours, NB - Naive Bayes, DT - Decision Tree, RI - Rule Induction. The letter M before each label (kNN, NB, etc.) means meta-version of a classifier, for example, the

TABLE I: Results of fault detection for global classifiers and meta-classifiers

	All	F0	F1	F2	F7	F8	F10	F11	F12	F13	F14	F15	F16	F17	F18	F19
kNN	**0,869**	0,987	**0,752**													
			0,791	1,000	1,000	0,005	0,946	1,000	0,889	1,000	0,003	0,771	0,451	1,000	1,000	1,000
NB	**0,857**	1,000	**0,717**													
			0,667	1,000	1,000	0,000	0,637	1,000	1,000	1,000	0,000	1,000	0,333	1,000	1,000	1,000
DT	**0,864**	0,947	**0,783**													
			0,836	1,000	1,000	0,007	0,991	1,000	1,000	1,000	0,007	1,000	0,429	1,000	1,000	1,000
RI	**0,864**	0,958	**0,772**													
			0,964	1,000	0,787	0,029	1,000	1,000	1,000	1,000	0,030	1,000	0,602	1,000	1,000	1,000
MkNN	**0,867**	0,959	**0,776**													
			1,000	1,000	1,000	0,007	1,000	1,000	1,000	1,000	0,007	1,000	0,450	1,000	1,000	1,000
MNB	**0,876**	0,989	**0,764**													
			1,000	1,000	1,000	0,000	1,000	1,000	1,000	1,000	0,000	1,000	0,334	1,000	1,000	1,000
MDT	**0,871**	0,935	**0,809**													
			1,000	1,000	1,000	0,086	1,000	1,000	1,000	1,000	0,086	1,000	0,667	1,000	1,000	1,000
MRI	**0,869**	0,952	**0,788**													
			1,000	1,000	1,000	0,086	1,000	1,000	1,000	1,000	0,086	1,000	0,425	1,000	1,000	1,000

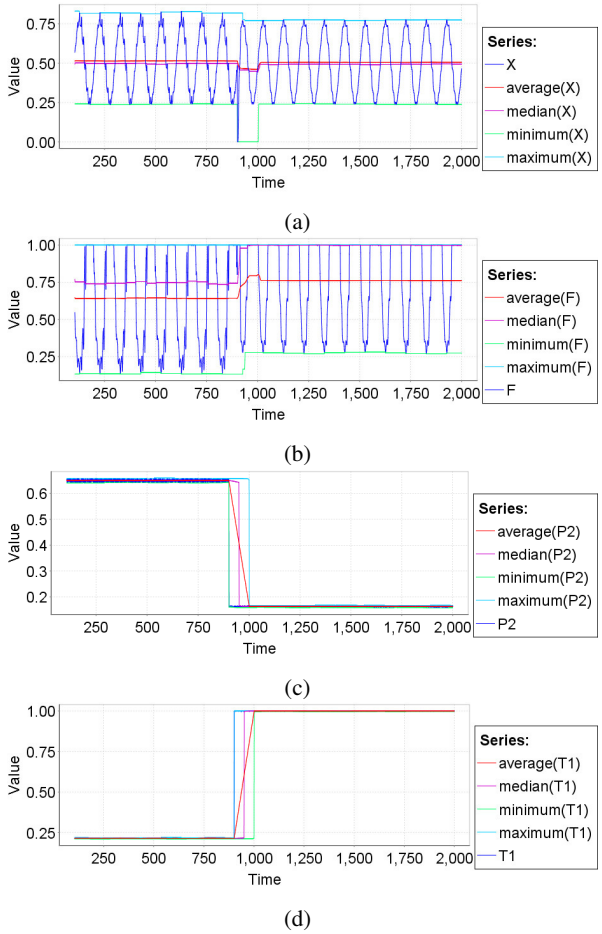


Fig. 9: Examples of process variables and their features

label MNB denotes a meta-classifier which is based on a naive Bayes classifier.

Table I shows results of two concepts of fault detection realized by schema presented in Fig. 2 and 3. The effectiveness of the classifiers is given in columns. The column indicated as "All" includes the general efficiency calculated on the basis of the confusion matrix which was generated after the classifier verification process. The next column (F0) includes the efficiency obtained for faultless states. The rest of the columns (F1-F19) show the efficiency of fault detection for all considered faults, separately. Above these columns the general result of the efficiency of fault detection is presented. Rows from 1 to 4 show results for single classifiers, whereas the next four rows show results for considered meta-classifiers.

2) Results of fault isolation: The results of comparison of the first fault isolation method (Fig. 4) are included in Tab. II. The same types of classifiers were used. The column indicated as "All" includes values of the global efficiency for single classifiers of different types. The rest of the columns show information about the efficiency of fault isolation for each scenario.

The second method presented in Fig. 5 use four classifiers as in the previous method but the outputs of these classifiers are connected to a meta-classifier. The results obtained for the meta-classifier are compared in Tab. III. Tab. IV presents the general efficiency of single classifiers (rows from 1 to 4) and meta-classifiers (rows from 5 to 8) for the fault isolation process.

TABLE IV: Comparison between all methods of fault detection for the learning and verification stages

	Learning and testing stage	Verification stage
kNN	0,988	0,844
NB	0,722	0,661
DT	0,176	0,177
RI	0,989	0,808
MkNN	0,990	0,837
MNB	0,860	0,835
MDT	0,890	0,777
MRI	0,982	0,812

The last method of fault isolation (Fig. 6) is based on series of single classifiers, where each classifier is used for detecting a single fault. The first task of the verification process was to choose a single classifier (from four available) for the fault

TABLE II: Results of fault isolation for global classifiers

	All	F1	F2	F7	F8	F10	F11	F12	F13	F14	F15	F16	F17	F18	F19
kNN	**0,844**	**1,000**	**1,000**	**1,000**	**0,516**	**1,000**	**1,000**	**1,000**	**0,876**	0,007	**1,000**	0,829	**1,000**	**1,000**	**1,000**
NB	0,661	0,667	**1,000**	**1,000**	0,347	0,333	**1,000**	0,669	0,153	**0,608**	**1,000**	0,383	**1,000**	**1,000**	**1,000**
DT	0,177	**1,000**	0,000	**1,000**	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
RI	0,808	0,885	**1,000**	**1,000**	0,443	0,947	0,754	**1,000**	0,822	0,003	**1,000**	**0,917**	**1,000**	**1,000**	0,893

TABLE III: Results of fault isolation for meta-classifiers

	All	F1	F2	F7	F8	F10	F11	F12	F13	F14	F15	F16	F17	F18	F19
MkNN	**0,835**	**1,000**	**1,000**	**1,000**	0,672	**1,000**	**1,000**	**1,000**	**1,000**	0,086	**1,000**	**0,708**	**1,000**	**0,668**	**1,000**
MNB	0,802	**1,000**	**1,000**	**1,000**	**0,729**	**1,000**	**1,000**	**1,000**	**1,000**	0,253	**1,000**	0,362	**1,000**	0,416	**1,000**
MDT	0,777	**1,000**	**1,000**	**1,000**	0,175	**1,000**	**1,000**	**1,000**	0,540	**1,000**	**1,000**	0,391	**1,000**	0,369	**1,000**
MRI	0,812	**1,000**	**1,000**	**1,000**	0,621	**1,000**	**1,000**	**1,000**	**1,000**	0,339	**1,000**	0,578	**1,000**	0,326	**1,000**

detection purpose. To solve this problem the authors tested all classifiers for all available faults. The results are presented in Tab. V. The values included in the table present the general efficiency of each classifier. The bolded values are related to the classifiers which were chosen as the basic classifiers for the meta-classifier.

In the next step of the method the meta-classifier is used. Its inputs are connected to the outputs of basic classifiers (the degrees of the belief for single fault detection). The main task of this meta-classifier is to compute the final result. Table VI presents results of different types of classifiers which are presented in the same form as in the second method of fault isolation (Tab. III). In the first column indicated as "All" there are included values of the general efficiency of the meta-classifiers. In the next columns the efficiency values of single fault isolation obtained by means of meta-classifiers are included.

3) Discussion of the results: The comparison of two schemes of the fault detection (Fig. 2 and 3) as well as different types of classifiers (kNN, NB, DT, RI) showed that results which were obtained using these schemes were very similar to each other. Moreover, the comparison between the same classifiers and other methods showed that the meta-classifier was characterized by the highest fault detection efficiency at the expense of the fault-free state. Both methods revealed the some problems with detection of faults F8 and F14, because these faults were very strong correlated with faultless state. The best meta-classifier returned more accurate results then the best single classifier, but the difference between them was too small to conclude that this meta-classifier could be better than the second one. The last three methods concerned fault isolation without taking into account a faultless scenario. In the first method (Fig. 4) the global single classifier was used in order to qualify the current fault class of the device. The results of tested classifiers were different, e.g. the global efficiency of DT was 0.177 and the second one (sorted by their efficiency values) was 0.661. Decision tree was able to recognize fault F7. However, value 1.000 for fault F1 indicated 100% efficiency of detection fault F1, but on the other hand the rest of faults were also recognized by this classifier as fault F1. The next classifier (NB) had better results than a decision tree algorithm. The last two classifiers showed a similar global efficiency. The k-Nearest Neighbour classifier (kNN) provided the best basic efficiency for most of single faults. The exception was fault F16 which was better recognized by the rule induction classifier (RI) and fault F14 which was better detected by the Naive Bayes (NB). The results obtained for the second scheme of fault isolation (Fig. 5), which was based on the meta-classifier were more similar to each other than in the first case (Fig. 4). There was a group of faults which were very easy to isolate by all meta-classifiers (F1, F2, F7, F10, F11, F12, F15, F17, F19). The general level of the efficiency in this method was increased significantly in some cases. But the comparison between the best single classifier and the best meta-classifier showed that in the second case, the global efficiency was worse. The more careful analysis of the results for these two classifiers showed that most of the best meta-classifier results were better than the results of the best single classifier. Some exceptions were faults F16 and F18. In the second case deterioration was very sizeable and it was the main reason of the worse result for this classifier. The third scheme of fault isolation (Fig. 6) was divided into two parts. The first part was dealt with the selection of the basic classifiers, applied to isolate a single fault. After the analysis of the results presented in Table V the authors nominated the classifiers for single fault detection. These classifiers were chosen on the basis of general results. In case more than one classifier had the same efficiency value (more classifiers with the efficiency equal to $1,000$ for the fault at the same time) the authors pointed out a classifier with more stable results in the time domain. The previous analysis showed that this type of classifiers carried with it the increase of the efficiency of the meta-classifier. In the second part different types of meta-classifiers were compared. The final results of this method showed that the difference between the efficiency of specific fault isolation for a single meta-classifier was sizeable. Also the difference for the global efficiency among all met-classifiers was sizeable. The best result was obtained for the decision tree (0.840) and it was comparable with a case of the best efficiency level for the single global classifier (k-nearest neighbours: 0.844) in the first method (Fig. 4). The meta-classifier based on the naive Bayes method demonstrated the smallest efficiency for this approach.

TABLE V: Comparison results of base classifiers for fault isolation of single faults

	F1	F2	F7	F8	F10	F11	F12	F13	F14	F15	F16	F17	F18	F19
kNN	**1,000**	**1,000**	**1,000**	**0,930**	**0,992**	0,999	0,986	**0,981**	0,899	**1,000**	**0,908**	**1,000**	**1,000**	**1,000**
NB	0,643	1,000	1,000	0,904	0,696	0,974	0,894	0,934	0,904	1,000	0,709	1,000	0,934	0,845
DT	1,000	0,987	1,000	0,880	0,889	**0,999**	0,987	0,913	0,743	1,000	0,853	1,000	0,960	0,987
RI	0,947	1,000	1,000	0,879	0,954	0,987	**0,999**	0,949	**0,905**	1,000	0,875	1,000	0,973	1,000

TABLE VI: Results of fault isolation for meta-classifiers with a bank of classifiers for isolating single faults

	All	F1	F2	F7	F8	F10	F11	F12	F13	F14	F15	F16	F17	F18	F19
MkNN	0,824	**1,000**	**1,000**	**1,000**	0,517	0,870	**1,000**	**1,000**	**1,000**	0,010	**1,000**	**0,825**	**1,000**	**0,777**	**1,000**
MNB	0,359	**1,000**	**1,000**	**1,000**	**0,578**	**1,000**	0,472	0,000	0,000	0,000	**1,000**	0,000	**1,000**	0,000	0,000
MDT	**0,840**	**1,000**	**1,000**	**1,000**	0,328	0,889	**1,000**	**1,000**	0,889	**0,847**	0,000	0,459	0,000	**0,777**	**1,000**
MRI	0,748	**1,000**	**1,000**	**1,000**	0,718	0,718	0,914	0,000	**1,000**	0,086	**1,000**	0,595	**1,000**	**0,777**	**1,000**

V. CONCLUSION

In the paper the application of selected classification schemes for fault diagnosis of the actuator systems was presented. The main purpose of the paper was to compare single and meta-classification strategies that could be successfully used as reasoning approaches in off-line as well as on-line diagnostic expert systems. The research was realized basing on the classical and well-practised classification methods. The comparison study was carried out within the DAMADICS benchmark problem. The classification schemes were implemented in RapidMiner software which is a well-known open source system for data mining and knowledge discovery. The particular results of the fault detection study showed that for simple industrial actuators it is possible to apply simple classification schemes without the necessity of using more advanced methods which are based on meta-classifiers. Significant differences can be observed in case of the results that are related to fault isolation schemes. The best evaluation results obtained from the three classification methods are ranged from 0.835 to 0.844. It should be stated that it is possible to observe some important differences in outcomes obtained using simple classification methods in the first fault isolation scheme (see Fig. 4) and similar results in the second one (see Fig. 5). The third concept (see Fig. 6) leads to the varied results of the classification process. The merits in the case of using meta-classifier (the second method applied according to Fig. 5) can be seen for several faults, especially when compared this to the best single classifier. The last scheme (see Fig. 6) is the most complicated and there is the need to test various classifiers and to have additional learning datasets. Moreover, in this scheme the general efficiency of fault isolation is close to the result achieved by means of the single classifier.

In this study, the authors used a confusion matrix in order to evaluate fault diagnosis systems that were created applying different classification schemes. Nevertheless, this measure can be directly compared with false and true detection/isolation rates proposed by the authors of the DAMADICS simulator [24]. The results of fault detection and isolation using single or meta-classification strategies that were achieved in this study are comparable to even more advanced methods described in the literature [35], [36]. Furthermore, in this study the whole set of potential faults were investigated, whereas in the related papers only selected states were taken into consideration.

Overall, the application of single or meta-classification strategies allows to create effective as well as relatively less-complicated computational fault detection and isolation systems that can be successfully employed for on-line and off-line fault diagnosis of industrial actuators.

ACKNOWLEDGEMENT

The research presented in the paper was partially financed by the National Centre of Research and Development (Poland) within the frame of the project titled "Zintegrowany, szkieletowy system wspomagania decyzji dla systemów monitorowania procesów, urządzeń i zagrożeń" (in Polish) carried out in the path B of Applied Research Programme - grant No. PBS2/B9/20/2013. The part of the research was also financed from the statutory funds of the Institute of Fundamentals of Machinery Design.

REFERENCES

[1] F. Caccavale and L. Villani, *Fault Diagnosis and Fault Tolerance for Mechatronic Systems: Recent Advances*, ser. Springer Tracts in Advanced Robotics. Springer Berlin/Heidelberg, 2003.

[2] J. M. Kościelny, *Diagnostyka zautomatyzowanych procesów przemysłowych*. Warszawa: Akademicka Oficyna Wydawnicza EXIT, 2001.

[3] R. J. Patton, P. M. Frank, and R. N. Clark, *Issues of Fault Diagnosis for Dynamic Systems*. Springer-Verlag Berlin and Heidelberg, 2000.

[4] J. Korbicz, J. M. Kościelny, Z. Kowalczuk, and Cholewa, W. (Eds.), *Fault diagnosis. Models, artificial intelligence, applications*. Springer Berlin/Heidelberg, 2004. [Online]. Available: http://dx.doi.org/10.1017/S0263574704241133

[5] R. Isermann, "Model-based fault detection and diagnosis - status and applications," *Annual Reviews in Control*, vol. 29, no. 1, pp. 71–85, 2005. [Online]. Available: http://dx.doi.org/10.1016/j.arcontrol.2004.12.002

[6] M. Blanke, M. Kinnaert, J. Lunze, and M. Staroświecki, *Diagnosis and Fault-Tolerant Control*. Springer-Verlag Berlin Heidelberg, 2006.

[7] W. Moczulski, "Inductive acquisition of diagnostic knowledge for states tree with complex structure," *Mech. Syst. Signal Process.*, vol. 15, no. 4, pp. 813–825, 2001. [Online]. Available: http://dx.doi.org/10.1006/mssp.2001.1389

[8] W. Moczulski, *Diagnostyka techniczna. Metody pozyskiwania wiedzy*. Gliwice: Wydawnictwo Politechniki Śląskiej, 2002.

[9] W. Cholewa, "Real-time diagnostic expert systems," *CAMES*, vol. 9, no. 1, pp. 21–40, 2002.

[10] R. Isermann, *Fault-Diagnosis Systems. An Introduction from Fault Detection to Fault Tolerance*. Springer, 2006.

[11] L. Kuncheva, *Combining Pattern Classifier: Methods and Algorithms*. New Jersey: Wiley-Interscience, 2004.

[12] L. Lam, "Classifier combinations: Implementations and theoretical issue," *Lecture Notes in Computer Science*, vol. 1857, pp. 77–86, 2000. [Online]. Available: http://dx.doi.org/10.1007/3-540-45014-9_7
[13] M. Namdari, H. Jazayeri-Rad, and S.-J. Hashemi, "Process fault diagnosis using support vector machines with a genetic algorithm based parameter tuning," *Journal of Automation and Control*, vol. 2, no. 1, pp. 1–7, 2014.
[14] B.-S. Yang, X. Di, and T. Han, "Random forests classifier for machine fault diagnosis," *Journal of Mechanical Science and Technology*, vol. 22, pp. 1716–1725, 2008. [Online]. Available: http://dx.doi.org/10.1007/s12206-008-0603-6
[15] S. Pöyhönen, "Support vector machines in fault diagnostics of electrical motors," Helsinki University of Technology Control Engineering Laboratory, Tech. Rep., 2002.
[16] Q. Wu and Z. Ni, "Car assembly line fault diagnosis based on triangular fuzzy support vector classifier machine and particle swarm optimization," *Expert Systems with Application*, vol. 38, pp. 4727–4733, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2010.08.099
[17] T. Bhadra, S. Bandyopadhyay, and U. Maulik, "Differential evolution based optimization of svm parameters for meta classifier design," *Procedia Technology*, vol. 4, pp. 50–57, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.protcy.2012.05.006
[18] R. Cretulescu, D. Morariu, M. Breazu, and L. Vintan, "Weights space exploration using genetic algorithms for meta-classifier in text document classification," *Studies in Informatics and Control*, vol. 21, no. 2, pp. 147–154, 2012.
[19] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.
[20] T. Jingyuan, S. Yibing, Z. Longfu, and Z. Wei, "Analog circuit fault diagnosis using adaboost and svm," in *Communications, Circuits and Systems*, 2008, pp. 1184–1187. [Online]. Available: http://dx.doi.org/10.1109/ICCCAS.2008.4657978
[21] P. Yao, Z. Liu, Z. Wang, and S. Bu, "Fault signal classification using adaptive boosting algorithm," *Elektronika ir Elektrotechnika*, vol. 18, no. 8, 2012. [Online]. Available: http://dx.doi.org/10.5755/j01.eee.18.8.2635
[22] M. Woźniak, *Metody fuzji informacji dla komputerowych systemów rozpoznawania*. Oficyna Wydawnicza Politechniki Wrocławskiej, 2006.
[23] K. Kerdprasop and N. Kerdprasop, "Feature selection and boosting techniques to improve fault detection accuracy in the semiconductor manufacturing process," in *Proceedings of the International MultiConference of Engineers and Computer Scientist*, 2011.
[24] M. Bartyś, R. Patton, M. Syfert, Salvador de las Heras, and J. Quevedo, "Introduction to the DAMADICS actuator FDI benchmark study," *Control Engineering Practice*, vol. 14, no. 6, pp. 577–596, June 2006. [Online]. Available: http://dx.doi.org/10.1016/j.conengprac.2005.06.015
[25] R. Patton, F. Uppal, and C. Lopez-Toribio, "Soft computing approaches to fault diagnosis for dynamic systems: A survey," in *IFAC Symposium SAFEPROCESS*, June 2000, pp. 298–311.
[26] H. B. Kekre, T. K. Sarode, and J. K. Save, "Gender classification of human faces using class based pca," *International Journal of Scientific and Research Publications*, vol. 4, no. 2, 2014.
[27] F. Akthar and C. Hahne, *RapidMiner 5, Operator Reference*. www.rapid-i.com, 2012.
[28] P. Cichosz, *Systemy uczące się*. Warszawa: WNT, 2000.
[29] A. Lile, "Analyzing e-learning systems using educational data mining techniques," *Mediterranean Journal of Social Sciences*, vol. 2, no. 3, pp. 403–419, 2011. [Online]. Available: http://dx.doi.org/10.5901/mjss.2011.v2n3p403
[30] W. W. Cohen, "Fast effective rule induction," in *Twelfth International Conference on Machine Learning*, 1995.
[31] M. Bartyś and M. Syfert. Using damadics actuator benchmark library (dablib). [Online]. Available: http://diag.mchtr.pw.edu.pl/damadics/
[32] V. Puig, M. Witczak, F. Nejjari, J. Quevedo, and J. Korbicz, "A GMDH neural network-based approach to passive robust fault detection using a constraint satisfaction backward test," *Engineering Applications of Artificial Inteligence*, vol. 20, pp. 886–897, 2007. [Online]. Available: http://dx.doi.org/10.1016/j.engappai.2006.12.005
[33] M. Mrugalski, M. Witczak, and J. Korbicz, "Confidence estimation of the multi-layer perceptron and its application in fault detection systems," *Engineering Applications of Artificial Inteligence*, vol. 21, pp. 895–906, 2008. [Online]. Available: http://dx.doi.org/10.1016/j.engappai.2007.09.008
[34] J. Korbicz and M. Kowal, "Neuro-fuzzy networks and their application to fault detection of dynamical systems," *Engineering Applications of Artificial Inteligence*, vol. 20, pp. 609–617, 2007. [Online]. Available: http://dx.doi.org/10.1016/j.engappai.2006.11.009
[35] J. Calado, J. Sá da Costa, M. Bartyś, and J. Korbicz, "FDI approach to the damadics benchmark problem based on qualitative reasoning coupled with fuzzy neural networks," *Control Engineering Practice*, vol. 14, pp. 685–698, 2006. [Online]. Available: http://dx.doi.org/10.1016/j.conengprac.2005.03.025
[36] F. Previdi and T. Parisini, "Model-free actuator fault detection using a spectral estimation approach: the case of the damadics benchmark problem," *Control Engineering Practice*, vol. 14, pp. 635–644, 2006. [Online]. Available: http://dx.doi.org/10.1016/j.conengprac.2005.04.001

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1391–1396
DOI: 10.15439/2014F289
ACSIS, Vol. 2

Intelligent Association Rules for Innovative SME Collaboration

Gulgun Kayakutlu *
Industrial Engineering Dept.
Istanbul Technical University
Macka 34367
Istanbul, Turkey
Email: {kayakutlu@itu.edu.tr}

Irem Duzdar
Industrial Engineering Dept.
Istanbul Arel University
Türkoba Mahallesi Erguvan Sokak
34537, Tepekent İstanbul-Türkiye
Email: {iremduzdar@arel.edu.tr}

Eunika Mercier-Laurent
IAEUniversity Lyon 3
6, cours Albert Thomas
Lyon,France
Email: {eunika@innovation3d.fr}

***Abstract*—SMEs are encouraged to collaborate for research and innovation in order to survive in tough global competition. Even the technology SMEs with high knowledge capital have the fear to collaborate with other SMEs or bigger companies. This study aims to illuminate the preferences in customer, supplier and competitor collaboration within industry or inter industry. A survey is run on more than 110 companies and Machine Learning methods are used to define the association rules that will lead for success.**

***Index Terms*—Collaborative Innovation, Association Rules, SVM, SOM**

I. Introduction

KNOWLEDGE based SMEs need to construct successful alliances in order to have sustainable business in a competitive environment. Global experiences with randomly chosen collaborators have shown failures that caused the fear of new collaborative work. Causes of failure based on the culture and the type of collaboration are studied [1]. Alliance in new product development has been the focus of industrial researchers [2][3][4].

This study aims to provide a pre-analysis of the path for successful alliances that will lead improvements in innovative power. Both qualitative and quantitative analysis of the SME alliances is realized to find the conditions causing failures and supporting the success in innovation. Support Vector Machine and Self Organized Maps are used to define the most frequent patterns that will give the support and confidence to identify the relationships. Association rules achieved will determine the optimal use of resources.

This paper is so organized that the literature review will be given in the second section and the methodology definition will follow. The fourth section will be reserved for presenting the survey and the results. The conclusion will be given in the fifth and last section.

The implication of the study is generic enough to help any SME or research organization or large business to reduce risks in future alliances.

II. Background

The first research on Association Rule Mining and Methods is found in 1996 [5] trying to find the most frequent occurrences of events to support the linked processes. The research in the field followed the timeline shown in Figure 1.

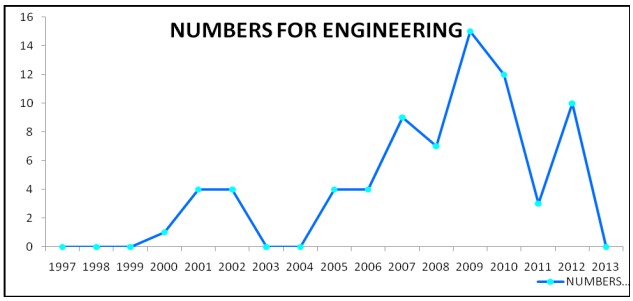


Figure 1. Research timeline on Association Rules

Post et Al. showed the fact that SMEs would like to collaborate only for developing new products [6]. However, Kakabadse et Al. showed that using improved communication and information technologies will improve the SME collaboration[1]. New product based collaboration has evolved fast [7]. Corporation and competition are found as flaming collaboration types that feed the SME improvements in innovation [8]. Association rules defined for the failure types have opened a new dimension for the research on failure of collaboration [9]. The first study on mining the SME innovation by Wang et Al has found some patterns for allocating the R&D resources [10]. Suh & Kim have detailed the R&D collaboration in service industries detected the positive relations of technology and the product or process innovation[11]. Swarnkar et Al. analyzed when and how the collaboration strategies will be used in virtual organizations[12]. Wiltsey et Al. claimed that extent, nature or impact of R&D programs are studied rarely. The interactions among the influences must be given in multiple levels and fidelity and changes must be observed in time [13]. Woodland & Hutton introduced the social dimension

on the collaborative success [14]. Both the fear issues and the success causes studied by Bouncken et Al defined technology influencers, sharing the knowledge and learning from the partner as the main influencers [2].

Knowledge management and data mining overviews [15] and Knowledge Management performance studies [16] realized recently do not show any association rule study for the collaborative innovation success and failure.

METHODOLOGIES

A. Association Rules

Given a set of transactions, rules are defined that will exhibit that the occurrence of an item based on the occurrences of other items in the transaction. This is the association analysis. It is useful to explore the interesting relations, which are embedded in the huge data sets. These hidden interactions can be stated in the form of association rules[17]. The strength of an association rule is measured with its support and confidence values. Support shows the how often that rule is applicable to a given dataset. The Confidence is the occurrence frequency of the item in that transaction [5].

Support (s) is the fraction of transactions that contain an itemset

$$\text{Support, } s(X \longrightarrow Y) = \frac{\sigma(X \cup Y)}{N}; \quad (1)$$

Confidence (c) measures how often items in Y appear in transactions that contain X

$$\text{Confidence, } c(X \longrightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}. \quad (2)$$

The item set patterns are found in various methods which could be apriori or aposteriori. The overview of all the methods used in association srule studies are given in Figure 2.

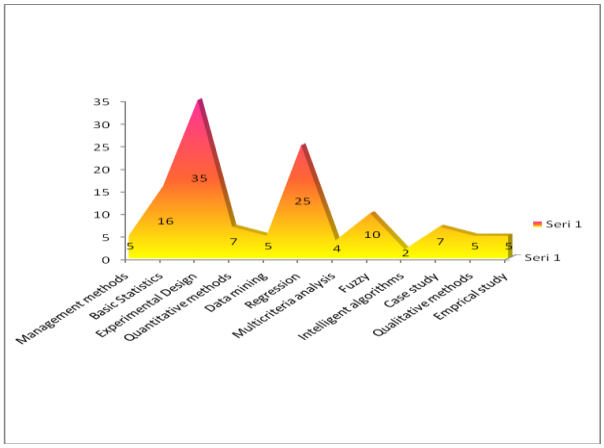


Figure 2. Overview of Association rule techniques

B. Support Vector Machines

It is a machine learning technique, which is mainly introduced for classification in two classes [18] but further used in clustering[19].

It can be analyzed as an optimization problem as in equation 3 [20] relaxed with Lagrange multipliers in objective function as in equation 4.

$$\min z = \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad r_i(w_i x_i + w_0) \geq 1 \quad (3)$$

Data is separated with a hyper-plane multiplied by -1 or +1.

$$L_p = \frac{1}{2}\|w\|^2 - \sum_i \lambda_i r_i (w_i x_i + w_0) + \sum_i \lambda_i \quad (4)$$

Using Using a Gaussian Kernel as defined in Eq. 5. will increase the reliability on dissimilarities [22].

$$K(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2}\right) \quad (5)$$

C. Self Organized Maps

Self-Organizing Map (SOM) is a widely used artificial neural network technique in clustering with unsupervised learning algorithm. This technique clusters according to the similarities to the input data [23]. SOMs structure the output with individual node similarity as well as cluster center distance. This technique is based on competitive learning, where the output nodes are made of the winning node activated by one input node. The output nodes would have scoring values using a function, most commonly Euclidean distance between the inputs and weights. For each input vector x, and for each output node j, the value D (w_j, x_n) of the scoring function. Euclidean distance function is shown in Eq. 6

$$D_{w_j, x_n} = (w_{ij} - x_{ni})^2 \quad (6)$$

The winning node therefore becomes the center of a neighborhood of excited nodes. In self- organizing maps, all nodes in the given neighborhood share competition. Therefore, even if the nodes in the output layer are not connected directly to the input layer, they tend to share common features, of the neighborhood [24]. The nodes in the neighborhood of the winning node participate in adaptation, which is, learning. The weights of these nodes are adjusted to improve the weights defined in Eq. 7., until a threshold is reached.

$$w_{ij}\ new = w_{ij}\ current + \alpha\ x_{ni} - w_{ij}\ current \quad (7)$$

In Eq 7. α is the learning rate. If it is necessary, the learning rate and neighborhood size are adjusted.

APPLICATION

A survey is run with the technology firms sited in Techno-parks of linked , 5 are about competences and 4 four the technologychoices. 130 firms responded but only 105 are included in the analysis. 14% of the companies were medium size and 37 % of them were aged more than 10 years. They have chosen the type of collaboration among the SME and Big firms as well as among the customers, suppliers and competitors as shown in Figure 3.

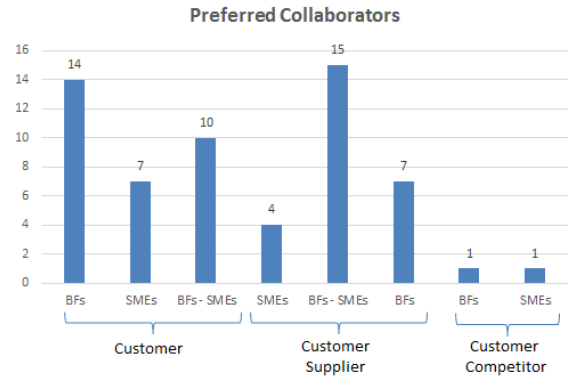


Figure 3. Choice of collaboration according to size and relation

The reason for innovative collaboration is stated as shown in Figure 4.

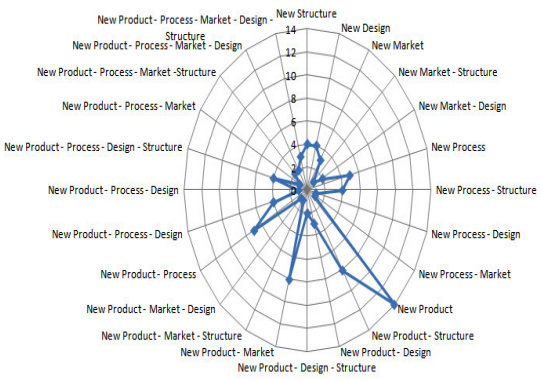


Figure 4. Innovation causes in collaboration

SVM is applied to classify the collaboration made specifically for innovation, validated by Cronbach alpha, that resulted as more than 0.7. SOM is used to cluster those to show the supporting frequencies of choices. Then the cross tables are achieved as a basis for the association rules. The first achievement was micro companies would like all the collaborators to have innovation culture which is less important for the bigger companies. Some other samples of cross tables are given as below.

Table 1 gives the fact that majority of responders prefer for at least one collaborator to have design competence.

Table 1. Business Understanding and Design Competence

			DESIGN		
INNOV			All	At least one	Negli ge
Collaboration for others	Bus-Under	All	4	14	
		At least one	2	15	
Collaboration for innovation	Bus-Under	All	14	26	4
		At least one	7	13	2
		Neglig.	1	1	2

In the firms collaborating for innovation for 1 to 3 years for the success of innovation it is necessary understanding the market requirements by all the firms together with the well-developed innovation culture.

Table 2. Age-Innovation Culture and Market Requirements Relation

			INNO_CULT		
COLLAB_Age			All	At least one	Neglige
0	UNDERSTD_REQ	All	13	8	
		More than one	6	6	
		At least one	0	1	
<1 year	UNDERSTD_REQ	All	9	1	0
		More than one	3	1	1
1-3 years	UNDERSTD_REQ	All	13	1	
		More than one	1	1	
		At least one	4	2	
3-5 years	UNDERSTD_REQ	All	6	1	
		More than one	3	2	
		At least one	0	1	
>5	UNDERSTD_REQ	All	8	2	0
		More than one	3	2	2
		At least one	2	2	0

RESULTS

The Reliability analyzes have been done on the results of the questionnaires, the Cornbach's Alpha value is 0.602; this value is in the acceptable range.

In the logistic regression model, the determined significance level is 0.100. The values of attributes of innovation are neglected, since they are greater than 0.100. Innovation related criteria in this study have no significance based on firm size and collaborator types of these firms.
In other words, firms care the technological features but ignore the innovative attributes.
In the table below, the significance values are shown for the other 3 attributes and the results of analyzes for features of the questions. The significance values less than 0.100 are used here (Figure 5).

Significance Table				
		Clusters		
		Finance	*Technology*	*Management*
Properties of Firms				
Firm Size	Firm Size	-	-	-
	Employee Size	0.000	0.000	0.000
	Firm Age	0.007	0.003	0.045
	Collaborate for Innovation	0.111	0.068	0.399
	Collaboration Duration	0.019	0.019	0.027
Employee Size	Firm Size	0.000	0.000	0.000
	Employee Size	-	-	-
	Firm Age	0.152	0.193	0.074
	Collaborate for Innovation	0.497	0.441	0.083
	Collaboration Duration	0.560	0.513	0.013
Firm Age	Firm Size	0.019	0.000	0.000
	Employee Size	0.176	0.002	0.009
	Firm Age	-	-	-
	Collaborate for Innovation	0.203	0.002	0.030
	Collaboration Duration	0.000	0.000	0.000
Collaborate for Innovation	Firm Size	0.006	0.019	0.041
	Employee Size	0.054	0.076	0.084
	Firm Age	0.010	0.025	1.000
	Collaborate for Innovation	-	-	-
	Collaboration Duration	0.000	0.000	0.000
Collaboration Duration	Firm Size	0.099	0.001	0.000
	Employee Size	0.624	0.045	1.000
	Firm Age	0.000	0.000	0.043
	Collaborate for Innovation	0.000	0.000	0.000
	Collaboration Duration	-	-	-

Figure 5. Significance values for each clusters based on demographic properties

After the multinomial logistic regression analysis coefficients for all attributes are obtained. The statistically significant attributes are used. Coefficients of financially related criteria are shown in Figure 6.

SIZE		B	Std. Error	Wald	df	Sig.	Exp(B)	90% Confidence Interval for Exp(B) Lower Bound	Upper Bound
2	Intercept	-6,395	2,530	6,387	1	,011			
	[ComAge_A=0]	2,079	1,020	4,156	1	,041	7,994	1,494	42,771
	[ComAge_A=1]	0ª			0				
	[ComAge_B=0]	1,725	,758	5,186	1	,023	5,614	1,615	19,522
	[ComAge_B=1]	0ª			0				
	[ComAge_C=0]	,652	,680	,920	1	,338	1,920	,627	5,877
	[ComAge_C=1]	0ª			0				
	[ComAge_D=0]	-,752	,732	1,056	1	,304	,472	,142	1,571
	[ComAge_D=1]	0ª			0				
	[Cap_A=0]	,055	,611	,008	1	,928	1,057	,387	2,889
	[Cap_A=1]	0ª			0				
	[Cap_B=0]	,145	,718	,041	1	,840	1,156	,355	3,768
	[Cap_B=1]	0ª			0				
	[Innov_Op_A=0]	,184	,716	,066	1	,797	1,203	,370	3,906
	[Innov_Op_A=1]	0ª			0				
	[Innov_Op_B=0]	,741	,714	1,075	1	,300	2,098	,648	6,794
	[Innov_Op_B=1]	0ª			0				
	[Price_A=0]	-,504	,927	,296	1	,587	,604	,132	2,775
	[Price_A=1]	0ª			0				
	[Price_B=0]	,473	,973	,237	1	,627	1,606	,324	7,957
	[Price_B=1]	0ª			0				
	[Export_A=0]	2,347	,910	6,659	1	,010	10,459	2,342	46,706
	[Export_A=1]	0ª			0				
	[Export_B=0]	1,839	,792	5,389	1	,020	6,289	1,709	23,142
	[Export_B=1]	0ª			0				

Figure 6. Regression Coefficients & Significancies

The cross relation tables have been constructed to define the rules obtained from the model.

Finance Related Criteria :

RULE 1:

IF

(Firm Size = “Micro” AND Firm Age ≤ “1”)

THEN

(Innovation operation expenditure = “proportionally shared” AND Price = “important”)

MEANING:

- ✓ Preferences for the initiating micro SMEs emphasize the innovation operation expenditures according to the collaborator sharings and market value (price) of the innovated product (Figure 7).

Finance Related Criteria		Capital_More than Average of Sector (A)	Capital_Average of Sector (B)	Innovation Operation Expenditure_Balance (A)	Innovation Operation Expenditure_More is more (B)	Price_Important. (A)	Price_Only motivation (B)	Exportation_All collaborators (A)	Exportation_Only one (B)
Firm Size	Firm Age								
MICRO (1)	< 1 YEAR OLD (A)				■	■			
	1-3 YEARS OLD (B)				■				
	3-5 YEARS OLD ©								
	5-10 YEARS OLD (D)								
SMALL (2)	< 1 YEAR OLD (A)								
	1-3 YEARS OLD (B)			■		■			
	3-5 YEARS OLD ©								
	5-10 YEARS OLD (D)			■			■		
MEDIUM (3)	< 1 YEAR OLD (A)								
	1-3 YEARS OLD (B)								
	3-5 YEARS OLD ©						■	■	
	5-10 YEARS OLD (D)								

Figure 7. Cross – relations Table: Firm Size – Firm Age – Finance Related Criteria

RULE 2:

IF

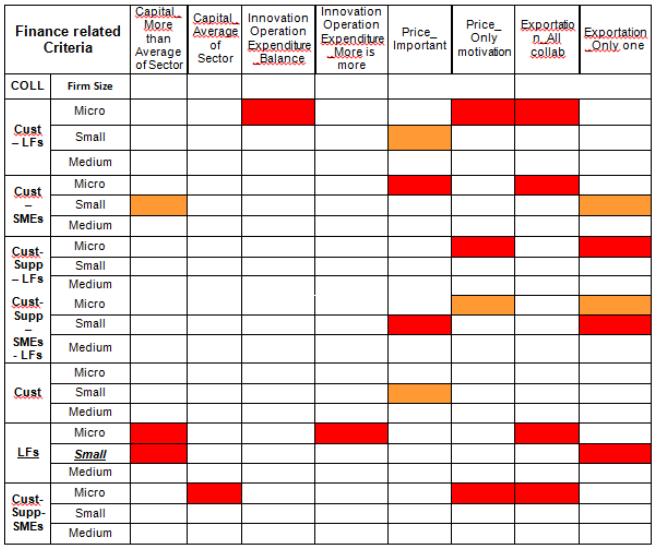
(Firm Size = “Small” AND Collaborate = “Large Firms”)

THEN

(Capital = “more than average” AND Exportation facilities = “only one firm”)

MEANING:

- ✓ Small SMEs emphasize that capital of the collaborators are to be more than the sector average and exportation facilities are done by only one collaborator for innovation (Figure 8).



Finance related Criteria		Capital_ More than Average of Sector	Capital_ Average of Sector	Innovation Operation Expenditure _Balance	Innovation Operation Expenditure _More is more	Price_ Important	Price_ Only motivation	Exportatio n_All collab	Exportation _Only one
COLL	Firm Size								
Cust – LFs	Micro								
	Small								
	Medium								
Cust – SMEs	Micro								
	Small								
	Medium								
Cust-Supp – LFs	Micro								
	Small								
	Medium								
Cust-Supp – SMEs - LFs	Micro								
	Small								
	Medium								
Cust	Micro								
	Small								
	Medium								
LFs	Micro								
	Small								
	Medium								
Cust-Supp-SMEs	Micro								
	Small								
	Medium								

Figure 8: Cross – relations Table: Collaboration – Firm Size – Finance Related Criteria

RULE 3:
IF
(Firm Age = "> 10 years old" AND Collaboration Duration = "3-5 years")
THEN
(Capital = "more than average" AND Price = "motivation")

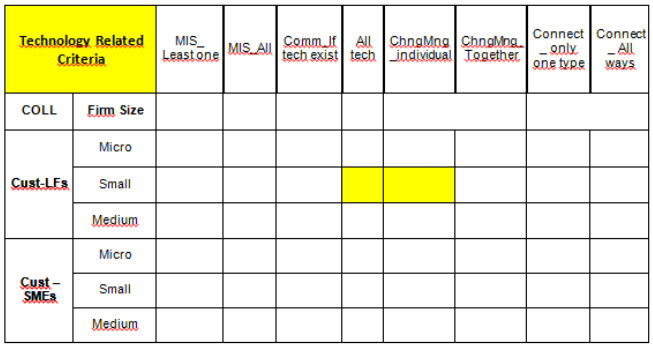
MEANING:

- ✓ 10 years (and more) old SMEs emphasize capital of the collaborators have to be more than the sector average whereas market value (price) of the innovated product is only a motivation.

Technology Related Criteria :

IF
(Firm Size = "Small" AND Firm Age = "3-5 years old")
THEN
(Connectivity = "All possible ways" AND Change Management = "All Collaborators")
MEANING:

- ✓ SMEs with age 5 to 10 years old emphasize change management is to be applied by all collaborators and use all possible connectivity possibilities.



Technology Related Criteria		MIS_ Least one	MIS_All	Comm_If tech exist	All tech	ChngMng _individual	ChngMng_ Together	Connect _ only one type	Connect _ All ways
COLL	Firm Size								
Cust-LFs	Micro								
	Small								
	Medium								
Cust – SMEs	Micro								
	Small								
	Medium								

Figure 9. Cross – relations Table: Collaboration – Firm Size – Technological Criteria

Figure 9 shows the cross – relations between collaborator type (customer – supplier – SMEs – large firms) and firms' size for the technology related criteria. Also this mentioned model makes sense statistically significant as a result of the logistic regression analysis.

RULE 2:
IF
(Firm Size = "Small" years old" AND Collaborate = "Customers" and "LFs")
THEN
(Communication technologies = "All opportunities" AND Change Management = "Individual")
MEANING:

- ✓ Small SMEs who are collaboraing with the customers which are large firms (LFs) emphasize usage of all communication technologies opportunities and change management is can be individual choice (Figure 9).

Management Related Criteria :

RULE 1:
IF
(Firm Size = "Medium" AND Firm Age = "5-10 years old")
THEN
(Professionalism = "Motivation" AND Organizational Structure = "Effective" AND Cooperation & Coordination = "All" AND Leadership = "Only one")
MEANING:

- ✓ Medium size SMEs of age 5-10 years prefer to collaborate with companies which have effective organizational structure with both cooperation and coordination attitude; in the collaboration a single leader is preferred and professionalism can be taken only as the motivator.

RULE 2:
IF
(Firm Size = "Micro" AND Collaborate = "Customers" and "SMEs")
THEN
(Professionalism = "All" AND Business Experience = "All" AND Leadership = "All")
MEANING:

- ✓ The micro SMEs collaborating with the customers emphasize professionalism, business experience for the problem solving and they prefer all the collaborators to have the leadership features.

CONCLUSION

This study investigates the most preferred conditions for a successful collaboration for innovative SMEs. SVM and SOM are used to construct the basis for creating the association rules. As the result of a survey in Turkey, there are hundreds of relations depicted in the analysis.

The achievements are interesting enough to show that the technology companies are confused in differentiating the technology and innovation concepts. It was interesting to observe micro and young companies not willing to collaborate with the big and overwhelming companies. Everybody asks for full communication technology, but only small SME with 5-10 years of experience ask for the collaborators to have effective organization and full professionalism.

The validation by logistic regression on the same data is in process. All the results achieved using logistics regression will be cross-validated with machine learning application results. Future survey will be aiming to improve the innovation concept of the technology firms in detail.

REFERENCES

[1] Kakabadse, N.K.; Kakabadse, A.; Ahmed, P.K.; Kouzmin, "The ASP phenomenon: an example of solution innovation that liberates organization from technology or captures it?," *Eur. J. Innov. Manag.*, vol. 7, no. 2, pp. 113–127, 2004.

[2] S. Bouncken, Ricarda B.; Kraus, "Innovation in knowledge-intensive industries: The double-edged sword of coopetition," *J. Bus. Res.*, vol. 66, no. 10, pp. 2060–2070, 2013.

[3] D. R. Gnyawali and B. R. Park, "Co-opetition and technological innovation in small and medium sized enterprizes A Multilevel Conceptual Model," *J. Small Bus. Manag.*, vol. 47, no. 3, pp. 308–330, 2009.

[4] R. Narula, "R&D collaboration by SMEs: New opportunities and limitations in the face of globalisation," *Technovation*, vol. 24, no. 2, pp. 153–161, 2004.

[5] R. Agrawal, "Parallel mining of association rules," *IEE Tran. Knowl. Data Eng.*, vol. 8, no. 6, pp. 962–969, 1996.

[6] G. J. J. Post, L. Hop, and J. E. van Aken, "Indicators for establishing SME product development networks," *J. Sci. Ind. Res.*, vol. 60, no. 3, pp. 264–276, 2001.

[7] R. Narula, "R&D collaboration by SMEs: New opportunities and limitations in the face of globalisation," *Technovation*, vol. 24, no. 2, pp. 153–161, 2004.

[8] D. R. Gnyawali and B. R. Park, "Co-opetition and technological innovation in small and medium sized enterprizes A Multilevel Conceptual Model," *J. Small Bus. Manag.*, vol. 47, no. 3, pp. 308–330, 2009.

[9] Z. Z. Z. Zheng, Z. L. Z. Lan, B. H. Park, and A. Geist, "System log pre-processing to improve failure prediction," *2009 IEEE/IFIP Int. Conf. Dependable Syst. Networks*, 2009.

[10] C. H. Wang, Y. C. Chin, and G. H. Tzeng, "Mining the R&D innovation performance processes for high-tech firms based on rough set theory," *Technovation*, vol. 30, pp. 447–458, 2010.

[11] Y. Suh and M.-S. Kim, "Effects of SME collaboration on R&D in the service sector in open innovation," *Innovation: Management, Policy & Practice*, vol. 14, no. 3. pp. 349–362, 2012.

[12] R. Swarnkar, A. K. Choudhary, J. A. Harding, B. P. Das, and R. I. Young, "A framework for collaboration moderator services to support knowledge based collaboration," *Journal of Intelligent Manufacturing*, vol. 23. pp. 2003–2023, 2012.

[13] S. Wiltsey Stirman, J. Kimberly, N. Cook, A. Calloway, F. Castro, and M. Charns, "The sustainability of new programs and innovations: a review of the empirical literature and recommendations for future research," *Implementation Science*, vol. 7. p. 17, 2012.

[14] R. H. Woodland and M. S. Hutton, "Evaluating Organizational Collaborations: Suggested Entry Points and Strategies," *American Journal of Evaluation*, vol. 33. pp. 366–383, 2012.

[15] H. H. Tsai, "Knowledge management vs. data mining: Research trend, forecast and citation approach," *Expert Syst. Appl.*, vol. 40, no. 8, pp. 3160–3173, 2013.

[16] I. B. Tae Hun Kim, Jae-Nam Lee, Jae Uk Chun, "Understanding the effect of knowledge management strategies on knowledge management performance: A contingency perspective," *Inf. Manag.*, vol. 51, pp. 398–416, 2014.

[17] J. Jackson, "Data mining: A conceptual overview," *Commun. Assoc. Inf. Syst.*, vol. 8, pp. 267–296, 2002.

[18] C.Cortes and V. Vapnik, "Support Vector Networks," *Mach. Learn.*, vol. 20, pp. 273-297, 1995.

[19] T. Finley and T. Joachims, "Supervised clustering with support vector machines," in *Proceedings of the 22nd International Conference on Machine learning (ICML)*, 2005, pp. 217–224.

[20] S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice-Hall, New Jersey, 1999.

[21] E. Alpaydın, Machine Learning, Massachusetts Institute of Technology, USA, 2004.

[22] V. Cherkassky, Y. Ma, Practical selection of SVM parameters and noise estimation for SVM regression, Neural Networks 17 (2004) 113–126.

[23] E. Leopold, M. May, and G. Paaß, "Data Mining and Text Mining for Science & Technology Research," in *Handbook of Quantitative Science and Technology Research - The Use of Publication and Patent Statistics in Studies of S&T Systems*, H. F. Moed, W. Glänzel, and U. Schmoch, Eds. Springer, 2004, pp. 187–213.

[24] Larose, D.T. 2005. Discovering Knowledge in Data: An Introduction to Data Mining. New Jersey: John Wiley & Sons, Inc.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1397–1406
DOI: 10.15439/2014F129
ACSIS, Vol. 2

Danger Theory-based Privacy Protection Model for Social Networks

Nai-Wei Lo[1] and Alexander Yohan[2]
Department of Information Management
Nat'l Taiwan Univ. of Sci. & Tech.
Taipei 106, Taiwan
nwlo@cs.ntust.edu.tw [1]
m10109803@mail.ntust.edu.tw [2]

***Abstract*—Privacy protection issues in Social Networking Sites (SNS) usually raise from insufficient user privacy control mechanisms offered by service providers, unauthorized usage of user's data by SNS, and lack of appropriate privacy protection schemes for user's data at the SNS servers. In this paper, we propose a privacy protection model based on danger theory concept to provide automatic detection and blocking of sensitive user information revealed in social communications. By utilizing the dynamic adaptability feature of danger theory, we show how a privacy protection model for SNS users can be built with system effectiveness and reasonable computing cost. A prototype based on the proposed model is constructed and evaluated. Our experiment results show that the proposed model achieves 88.9% detection and blocking rate in average for user-sensitive data revealed by the services of SNS.**

I. Introduction

NOWADAYS, people tend to use Social Networking Sites (SNS) to keep personal connections with others. According to Ho et al. [1], SNS is a website that provides a virtual community for people with similar interests in particular subject, or just to "hangout" together. Based on the definition of SNS, people can use SNS services to share information, thoughts and feelings, chat with other users, play online games with other users, and even promote their own businesses to other users [2]–[5]. The rapid growth and huge amount of service usage of SNS show that SNS has taken a significant role on communication media and culture among people in modern societies. For instance, Facebook Message is a social service provided by Facebook that allows people to chat online or message with each other offline.

During chatting or messaging sessions of SNS services, people may consciously or unconsciously input sensitive user information into their exchanging messages. Facebook provides a set of rules to protect its user's privacy [2], and it allows its users to determine who can access and see information shared by individual user. However, these rules cannot protect sensitive information shown within shared messages. When it comes to user privacy protection, it all depends on individual user itself to carefully avoid inputting user-sensitive data during the usage of SNS services.

This work was supported by Taiwan Information Security Center (TWISC) and National Science Council, Taiwan, under Grant no. NSC 102-2218-E-011-013

In the past few years, several mechanisms had been proposed by researchers to protect individual privacy for different occasions. User anonymization [6]–[10] and data encryption [11]–[14] are two major investigation directions.

User anonymization schemes as described in [6]–[10] protect a user from being identified from a set of data records by replacing user-related data within these records with system-generated strings before outputting these data for people to use. However, Beye et al. [3] explained that this kind of mechanisms still cannot defend against re-identification threat. Moreover, these mechanisms are not suitable to apply directly onto real-time chatting or messaging services since these mechanisms are only suitable for large dataset which contains thousands of different users' information. In addition, users have to identify themselves with a set of unique personal attributes first before they use any service offered by SNS.

Data encryption schemes as described in [11]–[14] protect communication privacy between users by encrypting communicating data or messages. Both communicating parties have to get corresponding keys in advance to encrypt and decrypt messages transmitting between them. Inevitably these mechanisms have key distribution/management problem to be resolved in practice. In SNS services, a real time message is usually shared or broadcast to multiple users/friends. Therefore, it is not easy to do group key management for SNS services provided that data encryption schemes are adopted directly. As the friend list of individual user will dynamically change from time to time, group key management for individual user will become an annoying routine job if the complexity of key management is endurable. In SNS environment, it is desirable to have a dynamic information protection mechanism that meets individual user's need. Since not all information shared by using SNS services is user-sensitive, a user privacy protection mechanism should be able to filter and determine which information is sensitive and needs to be protected.

Danger theory has been investigated in the field of Artificial Immune System [15] for its built-in ability to adapt dynamic changes automatically. Some researches using danger theory to construct application systems such as virus detection, network intrusion detection, and message filtering; have been conducted and shown its effectiveness.

978-83-60810-58-3/$25.00

In this paper we propose a privacy protection model for SNS Messaging System based on danger theory concept. Messages that does not contain sensitive information are defined as healthy cells. In contrary, messages that contains sensitive information are defined as injured cells. Danger signal defines what kind of dangers should be detected by the system. Danger signal in our model is the signal sent out by SNS messaging system when user-sensitive information inside a message is detected. Antigens are defined as the collection of user-related information. Based on our antigen's definition, we define the antibodies as a set of rules that regulate user-related information and determine what information is allowed to be shared to other users. Binary string is adopted to represent antigens and antibodies. Specific semantics of each bit within a binary string format are defined to indicate user-related data items and rules, respectively. A prototype was built based on the proposed model and performance evaluation was conducted accordingly. Based on the experiment results, the average accuracy rate for the proposed privacy protection model to correctly detect and protect user-sensitive information among shared messages is 88.9%.

II. Literature Review

A. Privacy Issues on Social Networking Sites

There are several user-related data that must exist and store in a SNS according to Beye et al. [3] such as profiles, connections, login credentials, messages, multimedia, groups, tags, preferences/rating/interests, and behavioral information. In this paper, we focus on one of the most popular services provided by SNS, i.e., online messaging service. This kind of service allows a user to exchange data/messages online or offline with another user or a group of users in SNS.

Since social networking sites such as Facebook usually collect and store all user-related data as stated by Beye et al., it raises two types of privacy issues. User-related privacy issues are generally caused by limited or lack of privacy control functionality support from service providers of SNS for their users [2], [3]. When SNS services provide a convenient platform for users to freely share information, individual user's sensitive information may be revealed by a user itself and freely accessed by other SNS users or even anonymous attackers. Another privacy issue is inability to hide user-sensitive information from other parties (friends or a specific group of users) [16], [17] because SNS vendors did not provide suitable privacy protection mechanisms. Another issue in this type category is user-sensitive information leakage caused by other users. When someone posted sensitive information related to a user of some SNS, it can harm the privacy of the indicated user.

As user's sensitive information can also be used to make profits for SNS service providers or other third-party companies that gain user information from SNS, the second type of privacy issues is often involved by SNS vendors and other companies cooperated with SNS vendors. According to Smith et al. [18], SNS users have no control over their published information on social network sites. In addition, users often do not know what SNS companies will do with their published data/messages. This kind of user privacy concern affects trust relationship between users and SNS service providers. Data retention on SNS is an example of the second type issues, in which all information that a user has ever been posted on the site is often impossible to be removed. Another example of privacy issues in this type category is unauthorized access to user data done by employees of SNS. Beye et al. [3], S. Mahmood [17], and D. J. Weitzner [19] indicated that most cases of privacy issue are related to sell user information to another party, such as an advertising company. Since SNS users cannot remove posted messages in SNS, those valuable information related with some users can be sold to other user-hunting companies such as advertising providers or insurance companies.

B. Privacy Protection Techniques

1. Privacy protection at service provider side

To solve privacy issues at service provider side, it is necessary for service providers to develop a privacy protection mechanism. One technique used to protect data privacy is to reduce the possibility that a user is identified based on data collected by a service provider. Several techniques have been proposed by researchers in the past to work on this issue, such as l-diversity [6], (α, k)-anonymization [7], and t-Closeness [8].

According to Machanavajjhala et al. [6], each row of data is composed from three different types of attributes: key attributes, quasi-identifiers, and sensitive attributes. Generally in published dataset, the value of a key attribute is in encrypted form to protect individual user's privacy. Aside from key attributes, quasi-identifiers might also be released in partial-encrypted form or in plain-text form. As for sensitive attributes, it is always released in plain-text form without any modification.

Based on the attribute structure for data records proposed by Machanavajjhala et al., user anonymization techniques developed in [6]–[8] can anonymize user-related information in published dataset. However, these mechanisms applying for huge dataset are not suitable to be applied directly in SNS environment, especially in real-time chatting or messaging services.

2. Privacy protection at the user side

Data encryption schemes were introduced to protect user's privacy in [11]–[14]. Privacy protection techniques based on encryption algorithms can be applied not only by SNS service providers but also by users of SNS. Since SNS users might be aware of privacy protection concern in SNS, they could take the initiative to protect their own privacy. In [11], Koch et al. presented a way to secure information shared among Facebook users by developing a third-party browser plug-in for Mozilla Firefox.

Since the plug-in utilizes cryptographic mechanisms, a user have to provide some information to invoke the plug-in.

The required data are targeted SNS URLs, usernames in the targeted sites, cryptographic algorithms, and their corresponding keys. Once the plug-in is activated, all messages entered by the user within targeted SNS web pages will be encrypted. To display the original messages posted in SNS, a user needs to present the corresponding key to the plug-in.

The mechanism in [11] has key distribution/management problem in practice. Since a real time message in SNS services is usually shared or broadcast to multiple users/friends, therefore group key management in SNS services is not an easy thing to do. Given that individual user's friend list is dynamically changed from time to time, it makes group key management become an annoying routine job. A dynamic user privacy protection should be able to manage this group key distribution/management problem. A good privacy protection mechanism should be able to filter sensitive information and protect it, because not all information shared using SNS services is user-sensitive.

C. Danger Theory

The original danger theory concept proposed by Polly Matzinger [20], [21] is a novel explanation on how the human body's immune system works. It is also an adaptive algorithm in the field of Artificial Immune System that generally used to perform virus detection, network intrusion, and message filtering. According to Lin et al. [15], this theory supersedes traditional self – non-self model, where the traditional model is more focusing on coping with any danger possibilities that may come from individual itself or outside of individual. The danger theory offers two advantages: the ability to prevent dangers in the future and the ability to defend against currently identified dangers.

There are several biology terms used in danger theory:

1. Tissue: a collection of cells in an organized form; multiple tissues can be organized to form an organ.
2. Cell: the basic structural, functional and biological unit of all known living organisms. In danger theory, there are two types of cell: normal cell and injured cell. A cell that does not cause harm to the corresponding organism is known as a normal cell. A cell that harms the corresponding organism is known as an injured cell.
3. Lymphocyte: any one of three types of white blood cell in a vertebrate's immune system. These three types of white blood cell are natural killer cell, T-cell, and B-cell.
4. B-cell: a type of lymphocyte in adaptive immune system. B-cell can be distinguished from other type of lymphocytes because it can bind to a specific antigen.
5. Antigen: any substance or incident that provokes an adaptive immune response in the body of an organism.
6. Antibody: Y-shape protein produced by plasma cells. Antibody is used by the immune system to identify and neutralize foreign objects such as bacteria and viruses.
7. Danger signal: an alarm signal sent out by a cell that is in distress or by an injured cell. The form of danger signal is varied in each immune system. For example, Lu et al. in [22] defined their danger signals based on the spreading and the damaging characteristics of mobile phone virus.

Fig. 1 shows an immune response according to the danger theory concept. A cell that is in distress sends out an alarm signal, known as the danger signal, whereupon antigens in the neighborhood are captured by *Antigen Presenting Cells* (APC), which then travel to local lymph nodes and present these antigens to lymphocytes. Essentially, the injured cell will establish a danger zone around itself. B-cell, one type of lymphocyte, will produce antibodies. Antibodies that can neutralize the antigens within the danger zone will perform clonal expansion process. Those antibodies that cannot neutralize the antigens or are located far away from the injured cell will not be stimulated and performed the clonal process.

Based on [23]–[25], the danger theory model can be

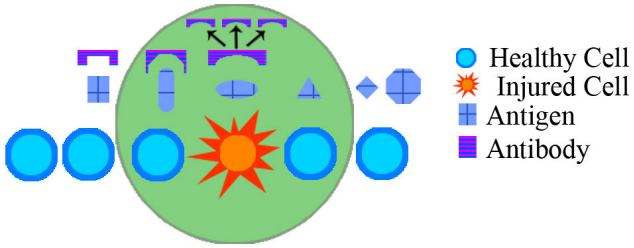


Fig 1. Danger theory model

viewed as an extension of Two-Signal model. In Two-Signal model, the two signals are antigen recognition (signal one) and co-stimulation (signal two). Co-stimulation signal is used to indicate the corresponding antigen is a dangerous one.

According to U. Aickelin and S. Cayzer [25], lymphocyte behaviors in danger theory are determined based on three laws:

- Law 1: A lymphocyte will be activated if the system receives both signal one and signal two altogether. If the system only receives signal one without signal two, then an activated lymphocyte will die. If the system only receives signal two without signal one, then it ignores the signal.
- Law 2: A lymphocyte only accepts signal two from Antigen Presenting Cells. Any cells can issue signal one to the system. Notice that an experienced T-cell or B-cell can act as Antigen Presenting Cells.
- Law 3: After the activation of a lymphocyte, the lymphocyte will revert to the resting state after a short time.

III. Proposed Privacy Protection Model

One of the most important privacy protection issues in SNS is the unintentional leakage of user-sensitive information caused by individual user's own carelessness. The importance of privacy protection on the data/messages shared by SNS users is not only for users' own interest, but also for service providers of SNS by offering privacy-aware SNS

TABLE I.
DEFINITIONS OF DANGER THEORY TERMINOLOGY FOR AISMPP

Danger Theory	AISMPP
Tissue	A message sent by a user to the other user(s).
Healthy cell	A message that does not contain any sensitive information related to individual user.
Injured cell	A message that contains any kind of sensitive information related to individual user.
Antigen (AG)	Collection of sensitive information related to individual user.
Antibody (AB)	A set of rules that regulates sensitive information, related to individual user, to decide which information is allowed to be shared to other(s).
Lymphocyte	The decision center.
Danger signal	A signal sent out by SNS messaging system indicating detected sensitive information inside one message and the user's response according to the given signal.
Signal one (danger signal)	A signal sent out by the SNS messaging system each time it detects user-sensitive information in one message.
Signal two (danger signal)	A signal used to indicate user's response or action toward the notification email sent by the system.
Danger zone	A status (state) indicator for a suspected message which might reveal user-sensitive information.
APCs	Messages received by users with alarm indication.

services to gain more users. To resolve this important user privacy issue in SNS environments, we propose the *Artificial Immune System Model for Privacy Protection* (AISMPP), which is derived from the concept of *Danger Theory* proposed by Polly Matzinger [20], [21]. To investigate detailed design of AISMPP, we select the most popular service among SNS functionalities, i.e., the messaging service, as the objective for AISMPP. The architecture of this system model is shown as Fig. 2.

In AISMPP design, there are six main components: SNS messaging service system, users database, user privacy settings, decision center, general rule repository, and antigen-antibody (AG-AB) database. As seen in Fig. 2, a user utilizes the messaging service offered by SNS to socialize with other users.

Each SNS has its own users and user privacy settings database. The privacy settings database contains a collection of rules and settings. These rules help the user to configure what data item in a user profile can be viewed by others, a black user list for each user, and a user data sharing list for third-party services (e.g. search engine services).

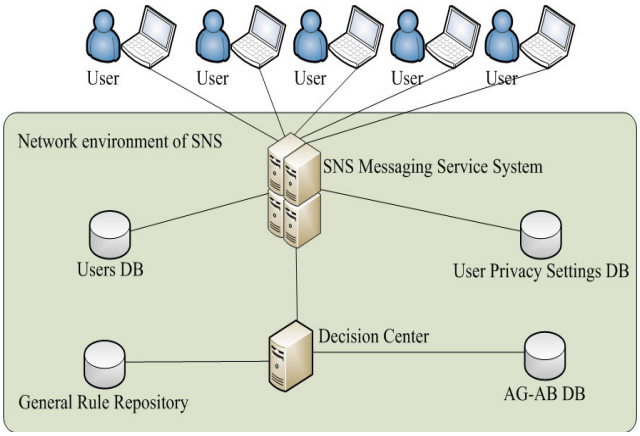


Fig 2. AISMPP architecture design

The decision center is responsible for processing danger signals, building danger zones, receiving antigens, generating and distributing antibodies. The decision center communicates with general rule repository and AG-AB database. The general rule repository describes all default actions for messages containing user-sensitive information. The AG-AB database stores all antigens and antibodies generated by the system.

The corresponding element definitions for AISMPP based on the danger theory are described in Table I.

Based on rules settings in [26] and [27], and our observation on E-Commerce websites and social networking websites; we propose 22 personal data items that are usually collected by SNS services. Thus in our design, an antigen is defined as a 22-bits binary string, where each bit represents one of 22 user's personal data items.

In Fig. 3 the AISMPP data flows are depicted. We describe each one of the flow processes as below:

1. AISMPP data flow processes start when user *A* sends a message *M* to user *B* using SNS's chatting/messaging service. Both user *A* and user *B* are participants in this chatting session.
2. Within the message SNS messaging system searches for user-sensitive information related with user *A*, user *B*, and their SNS friends list. If the system detects user-sensitive information in the message, then it will create an appropriate antigen based on the detected information.
3. After creating the appropriate antigen, the system will check the user privacy settings database for the privacy-affected user (or one of its friends) whether the disclosed user information is allowed to be shared or not. In our scenario, we assume that every user disallows any of its information to be shared by others.

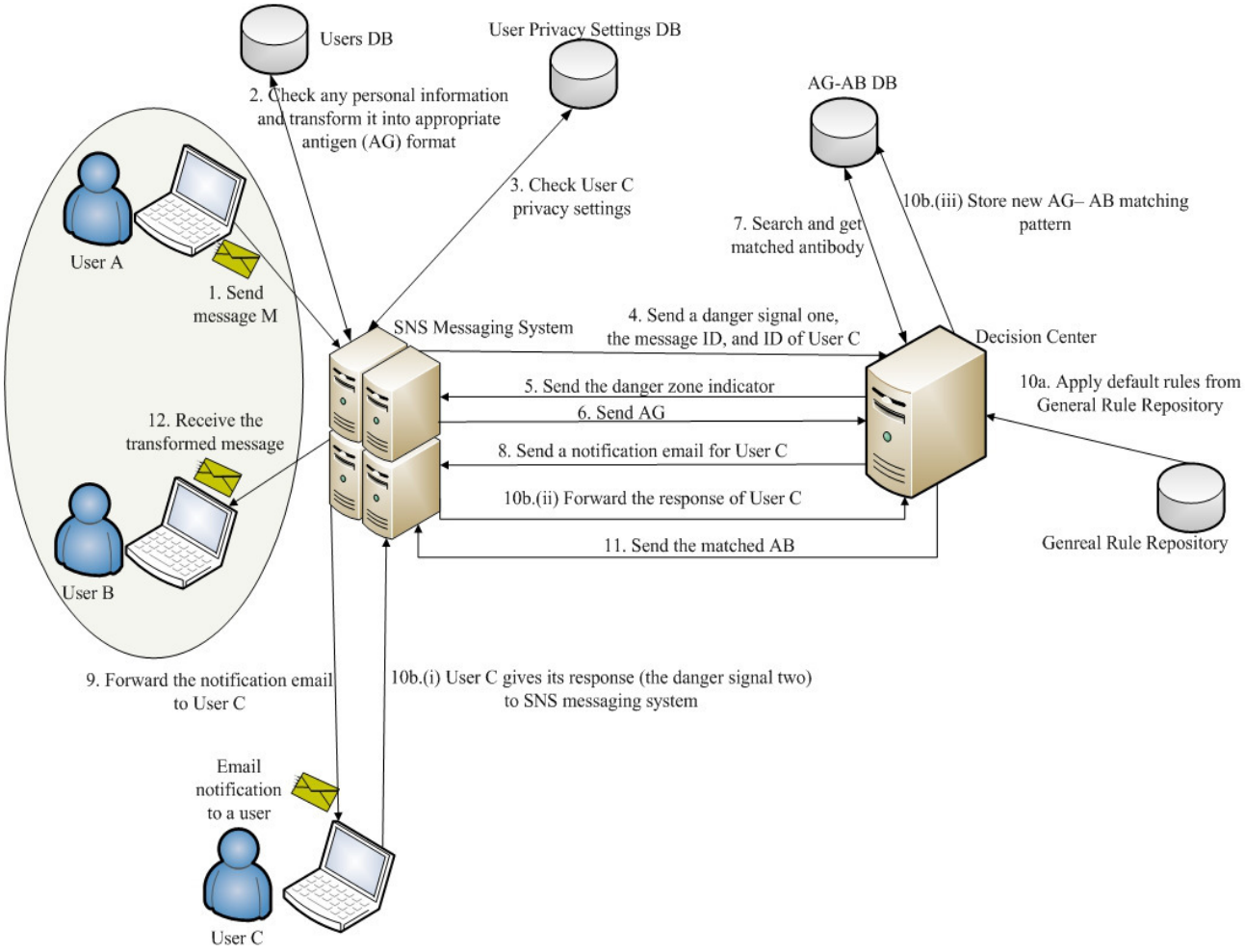


Fig 3. AISMPP data flow diagram

4. Every time the system detects any user-sensitive information in a message, the system will generate and send a danger signal (signal one) to the decision center along with the suspected message ID and the user ID of which user privacy has been breached.
5. The decision center will create a danger zone indicator on the suspected message based on the message ID and send the indicator to SNS messaging system.
6. After receiving the danger zone indicator from the decision center, the SNS messaging system sends the suspected antigen data to the decision center.
7. The decision center then check the antibody database to search the matched antibody for the received antigen.
8. If there is no matched antibody found on the antibody database, the decision center will request the SNS messaging system to forward a notification email to the message-affected user (assume it is user *C*) whose privacy is breached. The email notifies user *C* whether its sensitive information displayed in a message can be shared or not. If the decision center finds a matched antibody, then it will use this antibody and proceed to Step 11.
9. SNS messaging system will forward the notification email to User C.
10. Step 10 is divided into two cases based on how fast user *C* gives its response after the notification email sent by SNS messaging system:
    a. If there is no response from user *C* to SNS messaging system within a given period of time or in case that the decision center does not receive any response of user *C* forwarded by SNS system, then the decision center will apply default conservative rules stored in general rule repository for this suspected message.
    b. (i) If User C gives its response (the danger signal two) in time to SNS messaging system. (ii) SNS messaging system forwards the response of user *C* to the decision center. (iii) Based on the user response, the decision center will create a new antigen-antibody matching pattern and store it into the AG-AB database.
11. The decision center sends the matched antibody to the SNS messaging system. This antibody will be used by SNS messaging system to transform the message sent by user *A*.

12. After transforming the message, SNS messaging system will relay the message to user *B*.

IV. Prototype Design

Based on the data flow and functionality design shown in Section 3, an AISMPP prototype is developed. The AISMPP prototype consists of two stages. In the initialization and training stage, the prototype is initialized with default users list, default antigens, default antibodies, and initial message log based on real Facebook user data. In the system execution stage, the prototype will dynamically identify user-input messages with user-sensitive information and transform these message contents before they are displayed to designated users to preserve user privacy.

Fig. 4 shows the pseudocode of initialization and training stage. In addition to initialize the users list, default antigens are generated in the form of 22-bit binary string for each antigen pattern. If our prototype detects any specific user-sensitive information in a message, then the corresponding

```
PROCEDURE Init_Training_Model()
  INIT USERS ← LOAD (initial users list)
  INIT AG_DB ← LOAD (initial antigen
    data)
  INIT AB_DB ← LOAD (initial antibody
    data)
  INIT L ← LOAD (initial message log
  data)
  FOREACH (person P ∈ USERS)
    FOREACH (message M ∈ L)
      IF (message M contains sensitive
        information)
        new_ag ← create an antigen
          record from the found
          sensitive information
        new_w ← detected sensitive
          words
        new_ab ← create new antibody
          based on new_ag
        IF (new_ag ∈ AG_DB)
          UPDATE the number of
            occurrence of detected
            words based on new_w and
            promote the antigen new_ag
            to have longer life time.
        ELSE
          ADD new_ag to AG_DB
        ENDIF
        IF (new_ab ∈ AB_DB)
          UPDATE the number of usage
            of the antibody new_ab
        ELSE
          ADD new_ab to AB_DB
        ENDIF
      ENDIF
    ENDFOREACH
  ENDFOREACH
ENDPROCEDURE
```

Fig 4. Initialization and training stage of AISMPP

binary bit that represents one specific user-sensitive data item will be set to 1, otherwise 0. A corresponding antibody pattern for each antigen is also generated, in which a 22-bit binary string is used to indicate whether the data item in the corresponding antigen should be encrypted; the corresponding bit of the antibody is set to 1, if the data item in the antigen needs to be encrypted.

The user profile in our prototype contains user's real name, user's email(s), user's birthday date, user's address (city, region, and country), user's phone number(s), and other private information, e.g. religious and political views. According to [26] and [27], we define all items in user profile as user-sensitive information.

The prototype also loads friend information of users from real Facebook user data. The friend information getting from Facebook is limited to the full user name, email addresses, and phone numbers.

Some historical messages in plain-text form are stored in the message log during the initialization and training stage. These historical messages are used as the training data for our prototype. All historical messages will be processed in the initialization and training stage to make our prototype ready for regular operation. Each time the prototype receives an input message, the system will search the message for any user-sensitive information. If there is no user-sensitive information found in the message, then the message will be forwarded directly to the recipient without any transformation. Otherwise, the message-related data will be recorded, i.e., the receivers, the sender, and the message itself.

After the suspected message data has been recorded, a temporary antigen pattern is created and the suspected words in the message are recorded. In case the created antigen exists at antigen database, our prototype will update the number of occurrence of the words at the antigen database, which cause the system to generate the corresponding antigen. Otherwise, the newly created antigen will be added into the antigen database. Furthermore, our prototype will also create new antibody pattern based on the newly created antigen.

In Fig. 5 we show regular operation process of AISMPP at system execution stage. Two additional processes, user feedback processing and user message transformation, are included in this stage in comparison with the process for initialization and training stage.

In Fig. 6 the user feedback process assigns a flag indicator along with the user-affected message based on the user's response. Notice that if the user does not give any response after receiving notification email from SNS messaging system, then the prototype will assume the message possesses user-sensitive information and apply the message transformation process to protect user-sensitive information inside the message.

User message transformation process in Fig. 7 will transform all the detected words, which are user-sensitive information, within a message into encrypted ones. Considering robust security, SHA-256 algorithm is adopted for data encryption operation in our prototype.

```
PROCEDURE Execution_Process()
  LOOP
    M ← a new message inputted by user
    U ← a user ID that generates the
      message M
    FOREACH (person P ∈ friends of U)
      IF (M contains sensitive
        information)
        IF (person P responds to
          system notification)
          SET the flag of applying
            default rules F = FALSE
        ELSE
          SET flag F = TRUE
        ENDIF
        DF = Process_Feedback(M,
          default flag F)
        new_ag ← create an antigen
          record from the found
          sensitive information
        new_w ← collection of detected
          sensitive words
        new_ab ← create new antibody
          based on new_ag
        IF (new_ag ∈ AG_DB)
          UPDATE the number of
            occurrence of detected
            words based on new_w and
            promote the antigen
            new_ag to have longer
            life time.
        ELSE
          ADD new_ag to AG_DB
        ENDIF
        IF (new_ab ∈ AB_DB)
          UPDATE the number of usage
            of the antibody new_ab
        ELSE
          ADD new_ab to AB_DB
        ENDIF
        IF (DF)
          M = Transform_Message(M)
        ENDIF
      ENDIF
    ENDFOREACH
    SEND M to user
  ENDLOOP
ENDPROCEDURE
```

Fig 5. System execution stage of AISMPP

V. Implementation and Experiments

The prototype is developed on Windows 7 platform within a PC hardware of Intel Core i5 3.1 GHz CPU and 4 GB RAM. For prototype initialization, user profiles are generated based on the 274 friend contacts of one real Facebook user account and 100 user contacts generated by the Web service of generatedata.com. The historical message log from the same Facebook user account is loaded for the training stage. In initialization stage, 6 antigen patterns are defined in which 7 user data items including name, email, social security number, phone number, passport number, credit card number and credit card expiration date are used and

```
PROCEDURE Transform_Message(message M)
  M' = M
  FOREACH (get each word string W in M)
    IF (W is user-sensitive data)
      W' = ENCRYPT(W)
    ELSE
      W' = W
    ENDIF
    M' = REPLACE(W,W')
  ENDFOREACH
  RETURN new message M'
ENDPROCEDURE
```

Fig 7. The function to transform a user's message in AISMPP

combined to generate these patterns. In addition, 6 corresponding antibody patterns are generated based on the assumption that any revealed user data item should be blocked (by data transformation). For the system execution stage, extra 200 friend contacts from another Facebook user account and 300 new user contacts generated by the Web service of generatedata.com are loaded into our prototype. In addition, 5000 newly generated messages using the Web service of RandomTextGenerator.com combined with the historical message log from the second Facebook user account are imported to evaluate the effectiveness of our prototype.

Fig. 8 shows the time distribution for processing 5000 messages within 1000 experiments. The processing time includes the time for message scanning process and the time for detecting user-sensitive information among 5000 messages. The higher processing time during the first 30 experiments is caused by adding more antigen and corresponding antibody patterns based on the self-adaptive capability of danger theory model as shown in Fig. 9. Notice that the processing time gradually reaches stable condition around 25 seconds during our experiments.

Fig. 9 shows the distribution of total number of antigen patterns within 1000 experiments. After completing the training stage of our prototype, the total number of antigen is around 260. In the first couple of experiments, it can be seen that the total number of antigen increases quickly and reaches beyond 500. The reason is the patterns of newly generated messages for our experiments are different with the message patterns applied in the training stage. As the number of antigen increases, the message processing time is

```
PROCEDURE Process_Feedback(message M,
  flag F)
  IF (!F && user consider the sensitive
    information in message M is alright
    to be shared)
    ASSIGN the safe flag DF=FALSE to
      message M
  ELSE
    ASSIGN the danger flag DF=TRUE to
      message M
  ENDIF
  RETURN DF
ENDPROCEDURE
```

Fig 6. The function to process user feedback in AISMPP

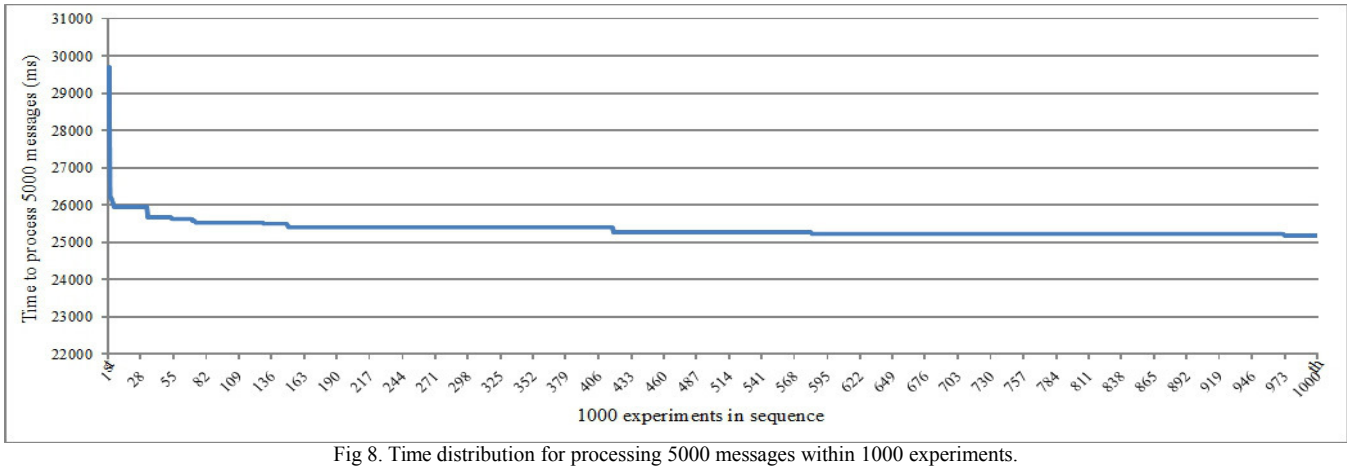


Fig 8. Time distribution for processing 5000 messages within 1000 experiments.

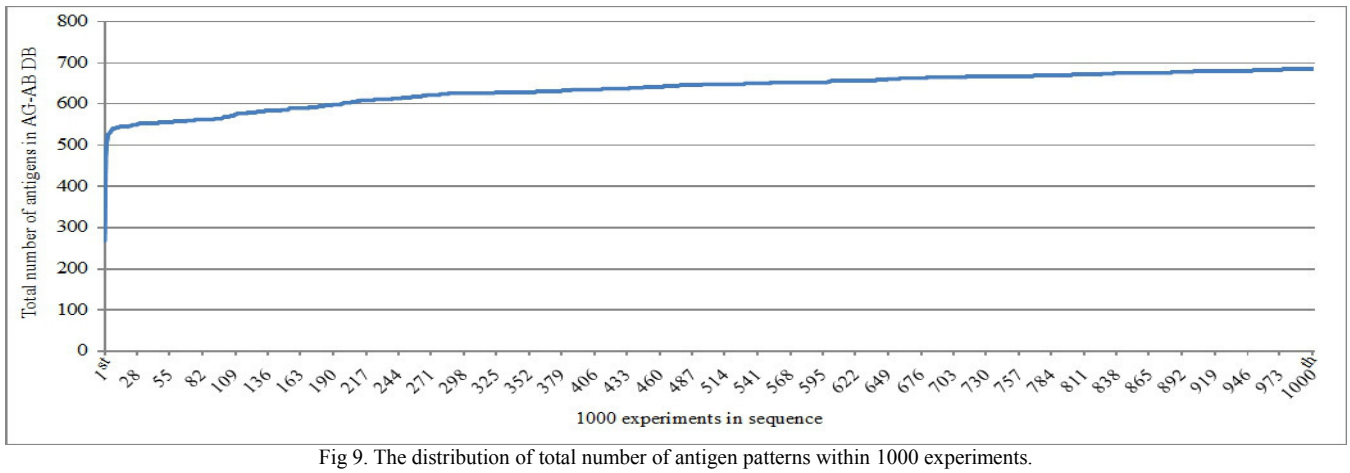


Fig 9. The distribution of total number of antigen patterns within 1000 experiments.

gradually reduced based on the observation of both Fig. 8 and Fig. 9.

Fig. 10 shows the time distribution for processing a message from the first message to the 5000th message at the 50th experiment. In Fig. 10 there are many messages been processed within 10ms. The reason is existing antigens have recognized this message pattern. Therefore, it does not require much time for the prototype to process these messages. For messages with new pattern, the prototype requires more time to process and generate new antigens and antibodies if necessary.

Fig. 11 shows the average time for our prototype to process one message in the first 100 experiments. For example, the average time to process one message in the 51st experiment is 5.21ms as shown in Fig. 11. Based on Fig. 11, we can know that the average time to process one message is ranged between 5ms and 10ms. We think the processing time is acceptable for a SNS messaging service.

We define a true positive (TP) value as our prototype correctly predicting one user-sensitive data shown in a message. A true negative (TN) value is defined as our prototype correctly predicting no user-sensitive data shown in a message. A false positive (FP) value is defined as our prototype wrongly predicting one user-sensitive data shown in a message. A false negative (FN) value is defined as our prototype wrongly predicting no user-sensitive data shown in a message. The precision rate is defined as TP / (TP + FP). The recall rate is defined as TP / (TP + FN). The accuracy rate is defined as (TP + TN) / (TP + TN + FP + FN). The average precision rate of our prototype is around 91.7% and the average recall rate is around 96.7%. The average accuracy rate of the prototype to correctly detect user-sensitive information in messages is around 88.9%.

VI. Conclusion

Lacking of automatic user privacy control mechanisms and privacy protection schemes is one of the most concerning issues in Social Networking Sites (SNS). In this paper, a privacy protection model based on danger theory is proposed. Several danger theory components are re-defined to fit into SNS environment and binary string format is adopted to represent antigens and antibodies used in our privacy protection model. Specific semantics of each bit within a binary string format are defined to indicate user-related data items and rules, respectively. Based on these components, an automatic adaptive immune system for user-sensitive information protection during online chatting/messaging is designed.

A prototype of our design was built based on the proposed model and performance evaluation was conducted accordingly. Based on the experiment results, the average accuracy rate for the proposed privacy protection model to correctly

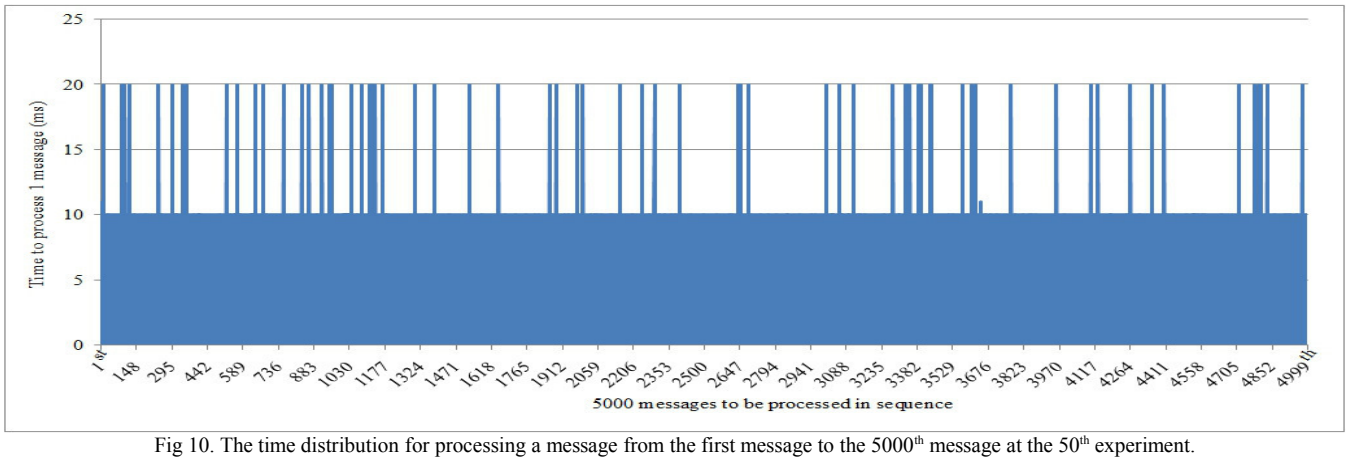


Fig 10. The time distribution for processing a message from the first message to the 5000th message at the 50th experiment.

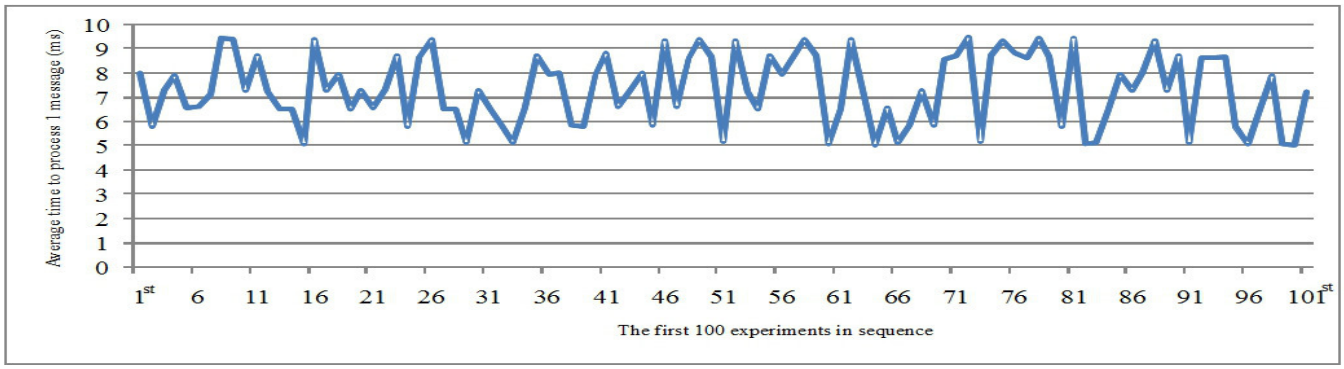


Fig 11. The average time for our prototype to process one message in the first 100 experiments.

detect and protect user-sensitive information among shared/broadcast messages is 88.9%. In addition, we found that the system processing time for each shared message is reduced with the increase of the number of recognized anti-gen patterns.

References

[1] A. Ho, A. Maiga, and E. Aimeur, "Privacy protection issues in social networking sites," presented at the IEEE/ACS International Conference on Computer Systems and Applications, 2009, pp. 271–278, http://dx.doi.org/10.1109/AICCSA.2009.5069336

[2] N. Talukder, M. Ouzzani, A. K. Elmagarmid, H. Elmeleegy, and M. Yakout, "Privometer: Privacy protection in social networks," presented at the IEEE 26th International Conference on Data Engineering Workshops (ICDEW), 2010, pp. 266–269, http://dx.doi.org/10.1109/ICDEW.2010.5452715

[3] M. Beye, A. J. P. Jeckmans, Z. Erkin, P. Hartel, R. Lagendijk, and Q. Tang, "Privacy in Online Social Networks," in *Computational Social Networks*, A. Abraham, Ed. Springer London, 2012, pp. 87–113.

[4] M. S. Choi, H. W. Kim, Y. H. Kim, K. H. Chung, and K. S. Ahn, "Private Information Protection System with Web-Crawler," presented at the IEEE International Conference on Wireless and Mobile Computing, Networking and Communications, 2008, pp. 672–677, http://dx.doi.org/10.1109/WiMob.2008.63

[5] Z. Chi, S. Jinyuan, Z. Xiaoyan, and F. Yuguang, "Privacy and security for online social networks: challenges and opportunities," *IEEE Network*, vol. 24, pp. 13–18, 2010, http://dx.doi.org/10.1109/MNET.2010.5510913

[6] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," in *Proceedings of the 22nd International Conference on Data Engineering*, 2006, p. 24, http://dx.doi.org/10.1109/ICDE.2006.1

[7] R. C. W. Wong, J. Li, A. W. C. Fu, and K. Wang, "(α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, Philadelphia, PA, USA, 2006, pp. 754–759, http://dx.doi.org/10.1145/1150402.1150499

[8] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," presented at the IEEE 23rd International Conference on Data Engineering, 2007, pp. 106–115, http://dx.doi.org/10.1109/ICDE.2007.367856

[9] P. Jurczyk and L. Xiong, "Privacy-preserving data publishing for horizontally partitioned databases," in *Proceedings of the 17th ACM conference on Information and knowledge management*, Napa Valley, California, USA, 2008, pp. 1321–1322, http://dx.doi.org/10.1145/1458082.1458257

[10] X. Jin, N. Zhang, and G. Das, "Algorithm-safe privacy-preserving data publishing," in *Proceedings of the 13th International Conference on Extending Database Technology*, Lausanne, Switzerland, 2010, pp. 633–644, http://dx.doi.org/10.1145/1739041.1739116

[11] R. Koch, D. Holzapfel, and G. D. Rodosek, "Data control in social networks," presented at the 5th International Conference on Network and System Security (NSS), 2011, pp. 274–279, http://dx.doi.org/10.1109/ICNSS.2011.6060014

[12] A. P. A. G. Deshmukh and R. Qureshi, "Transparent Data Encryption - Solution for Security of Database Contents," *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 3, pp. 25–28, Mar. 2011. Available: http://arxiv.org/abs/1303.0418

[13] Xukai Zou, Peng Liu, and J. Y. Chen, "Personal genome privacy protection with feature-based hierarchical dual-stage encryption," presented at the IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS), 2011, pp. 178–181, http://dx.doi.org/10.1109/GENSiPS.2011.6169474

[14] S. Jahid, P. Mittal, and N. Borisov, "EASiER: Encryption-based Access Control in Social Networks with Efficient Revocation," in *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, New York, NY, USA, 2011, pp. 411–415, http://dx.doi.org/10.1145/1966913.1966970

[15] L. Lin, L. Yiwen, Y. He, and Y. Chao, "Danger Theory: A new approach in big data analysis," presented at the International Conference on Automatic Control and Artificial Intelligence (ACAI), 2012, pp. 739–742, http://dx.doi.org/10.1049/cp.2012.1083

[16] A. Johnston and S. Wilson, "Privacy Compliance Risks for Facebook," *IEEE Technology and Society Magazine*, vol. 31, pp. 59–64, 2012, http://dx.doi.org/10.1109/MTS.2012.2185731

[17] S. Mahmood, "New Privacy Threats for Facebook and Twitter Users," presented at the 7th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2012, pp. 164–169, http://dx.doi.org/10.1109/3PGCIC.2012.46

[18] M. Smith, C. Szongott, B. Henne, and G. von Voigt, "Big data privacy issues in public social media," presented at the 6th IEEE International Conference on Digital Ecosystems Technologies (DEST), 2012, pp. 1–6, http://dx.doi.org/10.1109/DEST.2012.6227909

[19] D. J. Weitzner, "Google, Profiling, and Privacy," *IEEE Internet Computing*, vol. 11, pp. 95–96, c3, 2007, http://dx.doi.org/10.1109/MIC.2007.129

[20] P. Matzinger, "Tolerance, Danger, and the Extended Family," *Annual Review of Immunology*, vol. 12, no. 1, pp. 991–1045, Apr. 1994, http://dx.doi.org/10.1146/annurev.iy.12.040194.005015

[21] P. Matzinger, "The Danger Model: A Renewed Sense of Self," *Science*, vol. 296, no. 5566, pp. 301–305, Apr. 2002. Available: http://www.sciencemag.org/content/296/5566/301.abstract

[22] T. Lu, K. Zheng, R. Fu, Y. Liu, B. Wu, and S. Guo, "A Danger Theory Based Mobile Virus Detection Model and Its Application in Inhibiting Virus," *Journal of Networks*, vol. 7, pp. 1227–1232, 2012, http://dx.doi.org/10.4304/jnw.7.8.1227-1232

[23] M. Read, P. S. Andrews, and J. Timmis, "An Introduction to Artificial Immune Systems," in *Handbook of Natural Computing*, G. Rozenberg, T. Bäck, and J. N. Kok, Eds. Springer Berlin Heidelberg, 2012, pp. 1575–1597.

[24] E. Hart and J. Timmis, "Application areas of AIS: The past, the present and the future," *Applied Soft Computing*, vol. 8, pp. 191–201, 2008, http://dx.doi.org/10.1016/j.asoc.2006.12.004

[25] U. Aickelin and S. Cayzer, "The Danger Theory and Its Application to Artificial Immune Systems," in *1st International Conference on Artificial Immune Systems (ICARIS)*, Canterbury, UK., 2002, pp. 141–148.

[26] E. McCallister, T. Grance, Scarfone, and NIST U. S. Department of Commerce, *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)*. 2010.

[27] M. E. Callahan and U. S. Department of Homeland Security, *Handbook for Safeguarding Sensitive Personally Identifiable Information*. 2012.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1407–1414
DOI: 10.15439/2014F161
ACSIS, Vol. 2

Knowledge Extraction from professional e-mails

Nada Matta, Hassan Atifi, François Rauscher
ICD/Tech-CICO, Université de Technologie de Troyes
12 rue Marie Curie, BP. 2060, 10010 Troyes Cedex, France
nada.matta@utt.fr, Hassan.atifi@utt.fr, Fracois.Rauscher@utt.fr

***Abstract*—Some professional e-mails contain knowledge about how actor face problem in order to realize projects. This type of knowledge is produced in cooperative activity. Representing project knowledge leads to structure link between coordination, cooperative decision-making and communication. The main objective of our work is to extract knowledge from daily work. So the main questions of our research are:**

- **Can we extract knowledge from professional e-mails?**
- **If so, which type of knowledge can be represented?**
- **How to link this knowledge to project memory?**

We present in this paper our first work in this aim. Our hypothesis is tested on a software development application.

***Index Terms*—Knowledge Engineering, Knowledge Management, Project memory, Traceability, Professional e-mails, Pragmatics analysis.**

I. Introduction

CURRENTLY, Designers use knowledge learned from past projects in order to deal with new ones. They reuse design rationale memory to face new problems. Knowledge Management provides techniques to enhance learning from the past [5]. Their approaches aim at making explicit the problem solving process in an organization. Their techniques are inherited mainly from knowledge engineering. So, we find in these approaches in one hand, models representing tasks, manipulated concepts and problem solving strategies, and in the other hand, methods to extract and represent knowledge. We note for instance MASK [7], [14] and REX [11] methods. These methods are used mainly to extract expertise knowledge and allow defining profession memories.

But, design projects involve several actors from different fields. These actors produce knowledge when interacting together and take collaborative decisions. So, it is important to also tackle this type in knowledge, which is generally volatile.

We deal, in our approach with this type of knowledge, called Project memory [13]. Project memory must represent organizational and cooperative dimension of knowledge. Current techniques used in Knowledge management, based on expert interviews are not adapted to extract these dimensions of knowledge. To tackle knowledge produced in collaborative activity, we need techniques that help to extract knowledge from daily work. In this paper, we present a technique that help to extract knowledge from professional e-mails. The presented approach allows structuring extracted concepts and linking them to the project context. We use pragmatics analysis and knowledge engineering techniques for this aim.

II. Project Memory

A project memory is generally described as "the history of a project and the experience gained during the realization of a project" [13]. It must consider mainly (Fig. 1.):

- The project organization: different participants, their competences, their organization in sub-teams, the tasks, which are assigned to each participant, etc.
- The reference frames (rules, methods, laws, ...) used in the various stages of the project.
- The realization of the project: the potential problem solving, the evaluation of the solutions as well as the management of the incidents met.
- The decision making process: the negotiation strategy, which guides the making of the decisions as well as the results of the decisions.

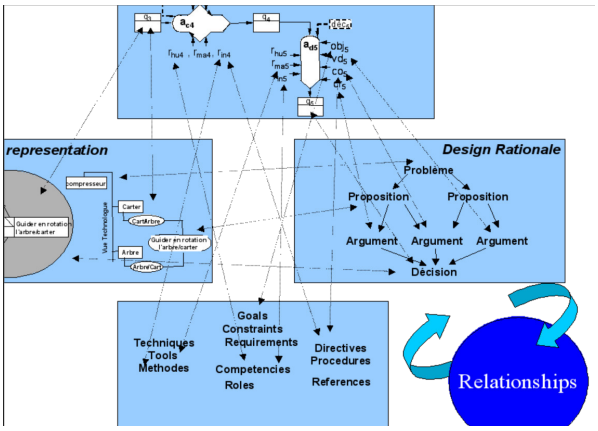


Fig. 1. Project memory

Often, there are interdependence relations among the various elements of a project memory. Through the analysis of these relations, it is possible to make explicit and relevance of the knowledge used in the realization of the project. The traceability of this type of memory can be guided by design rationale studies and by knowledge engineering techniques.

III. PRAGMATICS ANALYSIS

The act of request has been extensively studied in the field of theoretical linguistics (Searle 1969), intercultural and inter-language pragmatics [2], NLP community on automated speech act identification in emails [3], [11] etc. However, as pointed by *Rachele De Felice et al. [4]* there is very little work concerned with data other than spoken language and few researches seem to fully respond to requirements of being sufficiently general, non-domain specific, and easily related to traditional speech acts. In addition, few researchers have focused their research requests in business written discourse (workplace email communication). Lambert et al 2010 try to create tools that assist email users to identify and manage requests contained in incoming and outgoing email. Atifi et al. [1] analyze email effectiveness from the professional's point of view by mixing two kinds of analysis: a content analysis of interviews of professionals and a pragmatic and conversational analysis of emails. *Rachele De Felice et al. [4]* propose a global classification scheme for annotating speech acts in a business email corpus based on traditional speech act theory described by Austin and Searle [15].

.

IV. RELATED WORK ON E-MAILS ANALYSIS

Several approaches study how to analyze e-mails as a specific discourse. We note for instance, tagging work [17], in which Yelati presents techniques that help to identify topics in e-mails, or the use of zoning segmentation in [10]. Other works use natural language processing in order to identify messages concerning tasks and commitment [8]. They parse verbs and sentences in order to identify tasks and they track messages between senders and receivers.

Even there is lot of work on pragmatics, which study dialogue and distinguish techniques in order to identify speech intention (Patient/doctor dialogue analysis [8]), coding dialogue scheme [Core et al, 1997], etc. Pragmatics analysis of e-mails uses only some of these methods like ngrams analysis by Carvalho in [16], Verbal Response Mode scheme by Lampert in [10]or a custom coding scheme like De Felice [4].

Techniques studying e-mails, often do not consider the context of discussions, which is important to identify speech intention. We deal with our work with professional e-mails, extracting from projects. So, we mix pragmatics analysis and topic parsing and we link this type of analysis to project context (skill and role of messages senders and receivers, project phases, and deliverables, etc.) in order to keep track of speech intention. As pragmatics analysis shows, there is not only one grid to analyze different types of speech intention. In project memory, we look for problem solving, design rationale, coordination, etc. In this study, we focus on problem solving and we build an analysis grid for this purpose.

V. PROJECT KNOWLEDGE EXTRACTION FROM E-MAILS

The main objective of our work is to extract knowledge from daily work. So the main questions of our research are:

- Can we extract knowledge from professional e-mails?
- If so, which type of knowledge can be represented?
- How to link this knowledge to project memory?

To answer these questions, we analyze professional e-mails related to projects. In last studies, we identify a structure to analysis coordination messages [12]. Based on pragmatics analysis, we defined a grid to structure coordination messages based on the main act to do (inform, request, describe, etc.) and the objects of coordination (task, role, product, etc.). In this paper, we will go ahead and define an approach that helps to extract knowledge from professional e-mails. So, we identify firstly step by step how to isolate important messages and how to analyze them. Knowledge from e-mails, as knowledge produced in daily work, cannot be very structured. It is related closely to context. In our work, we focus on knowledge produced during project realization. We will show in our method how information from project organization help in e-mails knowledge extraction.

A. Classification of e-mails

Firstly, we have to identify important messages (Fig. 2). For that, we have to gather messages in subjects. Then, we can identify the volume of messages related to each subject. Then we analyze only messages that heave more then 4 answers; we believe that knowledge can be extracted based on interaction. Finally, we link the messages to be analyzed to project phases.

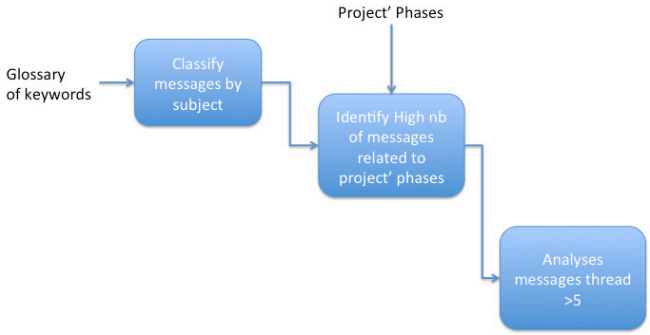


Fig. 2. First e-mails analysis

B. Messages analysis

For each message thread (message and answers), we identify (Fig. 3) :

- Information to be linked to organization:
  - Authors, To whom, In Copy
- Information about phases:
  - Date and hour of messages and answers
- Information about product:
  - Topic and joined files
- Information about message intention:
  - Main speech act and intention of message

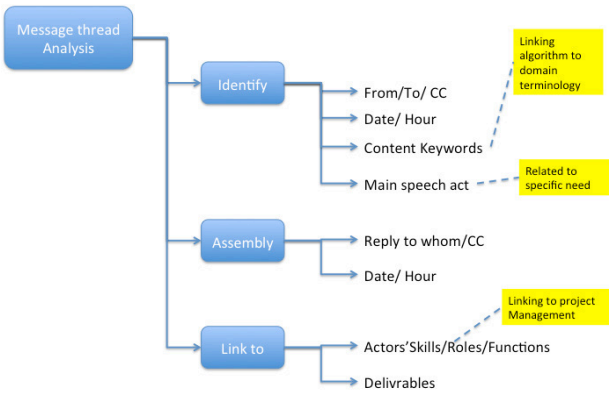


Fig. 3. Analysis of messages

By linking messages to project organization, we help in making sense of interactions between actors. In fact, the role and skill of messages' senders and receivers help to analyze the role of the message in problem solving and the nature of the content (solution answering a problem, proposition discussions, coordination messages, etc.). In the same way, linking messages to phases help to identify main problems to deal in each phase of the same type of projects.

As first work, we focus our speech act analysis on problem solving by identifying request and solution. So, we identify first speech acts that help to localize a request in a message (**Erreur ! Source du renvoi introuvable.**). Then, we study the organization of related messages thread in order to identify the solution proposed (if it exists) to the request. Our analysis is based first on pragmatics in order to characterize request speech act, and that by identifying request verbs and forms. In the present study we limited our research to the analysis of the act of requesting in problem solving sequences.

From a pragmatic point of view, a request is a directive speech act whose purpose is to get the hearer to do something in circumstances in which it is not obvious that he/she will perform the action in the normal course of events [15]. By introducing a request, the speaker believes that the hearer is able to perform an action. Request strategies are divided into two types according to the level of interpretation (on the part of the hearer) needed to understand the utterance as a request. The two types of requests include direct request and indirect requests. The request can be emphasized either projecting to: 1- the speaker (Can I do X?) or 2- the hearer (Can you do X?). A direct request may be use an imperative, a performativity, obligations and want or need statements.
An Indirect request may use query questions about ability, willingness, and capacity etc. of the hearer to do the action or use statements about the willingness (desire) of the speaker to see the hearer doing x. At last, for us, a grammatical utterance corresponds to only one speech act as in TABLE1.

TABLE 1.
GRID OF REQUEST SPEECH ACT

Request Form	Linguistic form	Examples
Direct request	Imperative	Do x
	Performative	I am asking you to do x.
	Want or Need statements.	I need/want you to do x
	Obligation statements	You have to do x
Indirect request	Query questions about ability of the hearer to do X	Can you do x? Could you do x?
	Query questions about Willingness of the Hearer to do X	Would you like to do x?
	Statements about the willingness (desire) of the speaker	I would like if you ca do X I would appreciate if you can do X

Then, we complete our analysis by from one side identifying answers verbs and from another side, linking answers to actors' role and skills and also joining files. The date of answers can be an indicator of several elements in the organizations: engagement, difficulty of time spending of solution, stress and multi-responsibilities, etc. We aim at analyzing in the future the frequency of answers.

VI. EXAMPLE

A. *Example description*

INFOPRO Business Publishing Company asked a software Company to develop a workflow tool that helps journalists to edit their articles and to follow the modification of the journal. The period of the project was more than one year. Nearly all negotiations and discussions were through e-mails. In this project, the actors were:

- SRA: an editing responsible (skill: law and management, Role: Contractor)
- JBJ: Information System Manager (Skill: Information system, Role: Contractor)
- FX: Information System Developer (Skill: Software Engineering, Role: Development manager)
- CV: Prototyping (Skill: Human Machine Interface, Role: User Interface Modeling)
- RT: Information System Developer (Skill: Software Engineering, Role: Sub-contractor)

Principles phases of the project were (TABLE 2):

TABLE 2.
PHASES OF THE PROJECT

	2009											2010									
	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct
XML Import from existing Database																					
Document DataBase specification and development																					
Workflow link specification and development																					
User Interface specification																					
Export to magazine and website																					
Web service specification and development																					
Application test																					

B. *E-mails Analysis*

As first step, we identify messages topics based on e-mails subjects. In our project, we identify main discussions topics based on keywords:

- XML : structuration, tag, tree, xsd, dtd, schema
- BDD : Data base, table, editing part, code part
- Interface workflow : UI, Workflow, User Interface, login, user management,
- Code : Insurance Code, Legifrance, auto code, vehicle code, mutu code, chapter, article
- Document : new collection, construction, document
- Export paper : Indesign, layout, mapping, tag indd, indd template
- Export site: export web, web tag, web format, dtd web
- Export Author: word, author, xslt author
- Services: update legi, Legifrance update, FTP
- Word : macro word, addin, web service, word 2007, wordlink

Based on these topics we use a Lucene based algorithm, to compute distance and similarity between words in order to identify main topics of messages (boosting email subject importance compared to email body (Fig. 4). It is to be noted that some preprocessing occurred before on raw email body to remove duplicated answers):

$$score(q,d) = coord\text{-}factor(q,d) \cdot query\text{-}boost(q) \cdot \frac{V(q) \cdot V(d)}{|V(q)|} \cdot doc\text{-}len\text{-}norm(d) \cdot doc\text{-}boost(d)$$

Fig. 5 shows first step of analysis of these messages; in which we show senders and receivers and their skills, topics of messages and date of messages.

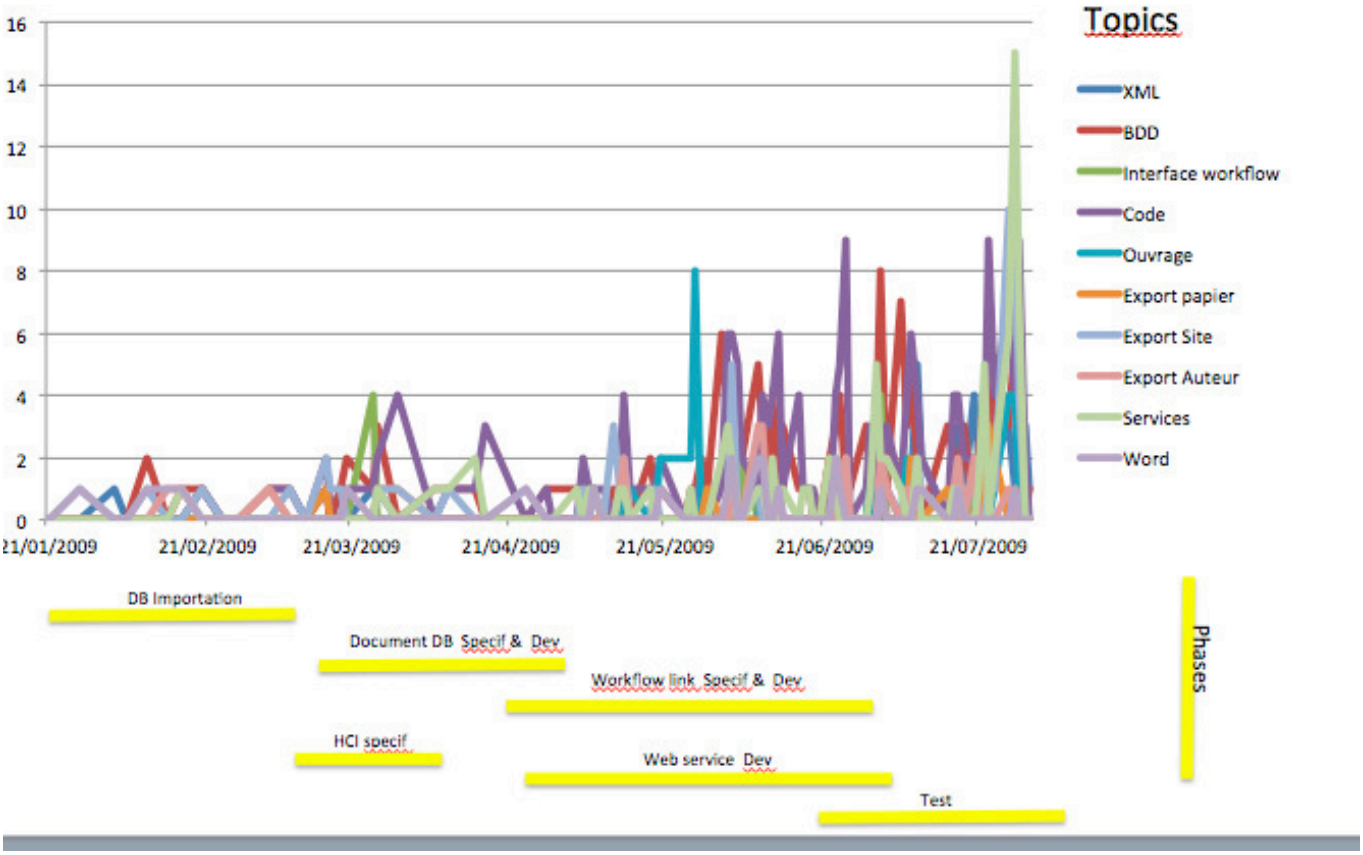


Fig. 4. Topics analysis

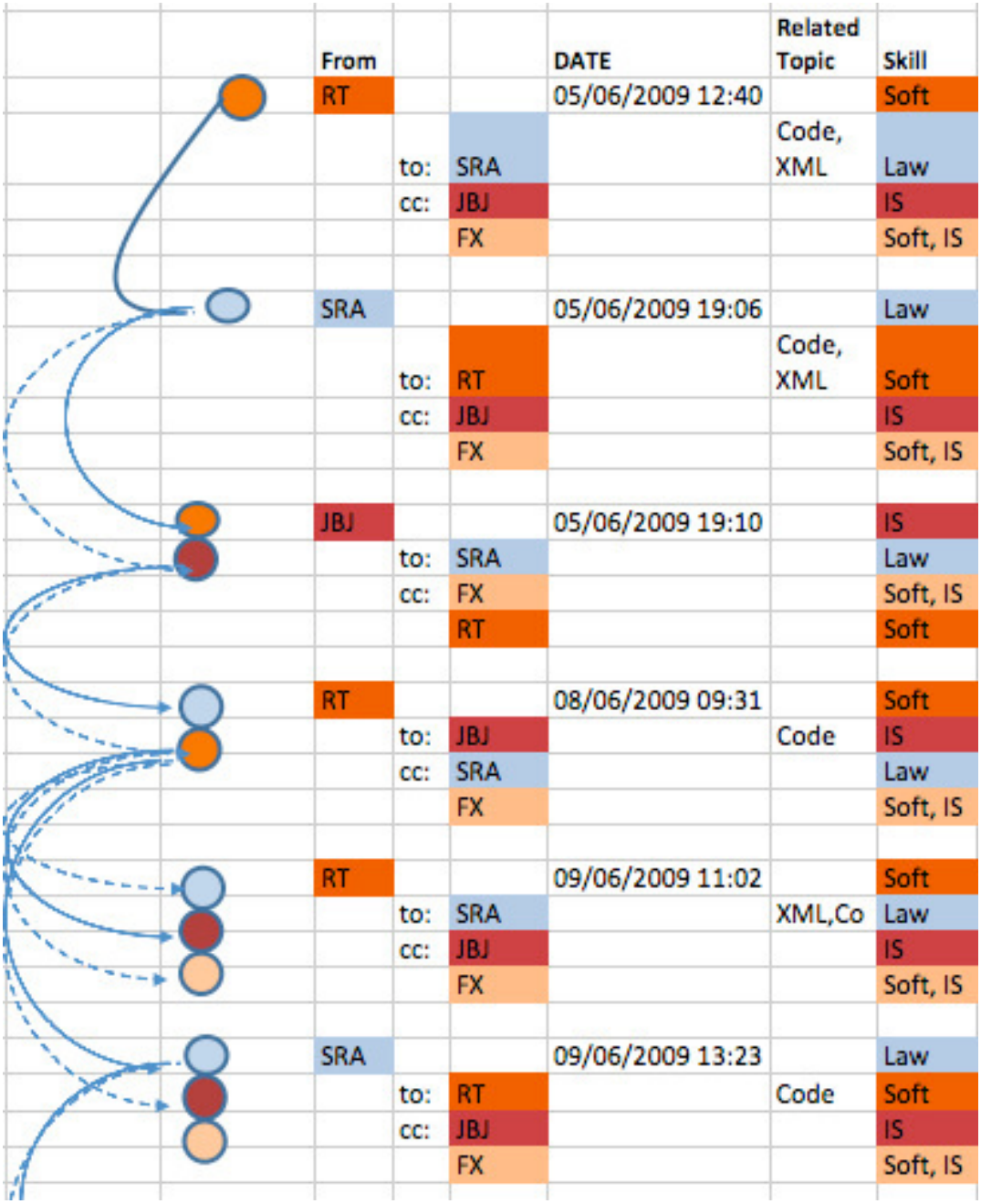


Fig. 5. First analysis of messages: representing of senders/Recievers/Copy, date and actors role and skill

To analyze messages text, we use pragmatics in order to identify problem and solution discussions. For that, we identify Request messages based on Request speech acts. Then, we identify related answers messages. In these messages we look for senders skill and joined files. So, we identify for the "Annexes" topic, in which there are 23 messages, related topics are XML and code. Messages were during 12 days from 5th until 17th June. They concern workflow development phase. Based on the Request-Answer grid and role actors, we analyze messages, in order to identify problem-solving intentions. So, we identify for instance, the problem Insurance Text extraction. SRA; the editing responsible (contractor) asks FX to extract Insurance text in a good format. When FX; the Information System Developer (Development manager) answer him, we suppose that as an answer, based on the role

of sender of message and the main topic. We consider also joined files as part of this answer. Fig. 6 shows this example.

From			Date	Sentence elements	Related Topic	Function
SRA			2009-06-05 12:40:46.	I put in "Bold", what **I need**:		Request
	to:	FX		1- *Inssurances*		
	cc:	JBJ		2- Text without tags Texte in XML files	Code	
		CV		3- Tag Pb : Text outside tag in XML	XML, Code	
		RT		4- Tag Pb <b> is opened and not closed, as same as, tag is badly imbircated		
FX			2009-06-05 19:06:34.			Answer
	to:	SRA		1- *Inssurances*		
	cc:	JBJ		I propose to convert: Xpress format in XML	XML	
		CV		Beware, the text will contain a lot of error blanc, "enter" and image	Code	
		RT		I can transform it on enriched XML	XML	
				It containts a lot of reference, so we have to compose with links		

Fig. 6. Example of messages analysis

VII. Conclusion

The aim of our study is to identify knowledge from daily work. In this paper, we show that it is possible to study professional e-mails for this aim. We consider e-mails as specific discourse. So we use pragmatics generally used to analyze discourse and to categorize it to identify knowledge from professional e-mails. Our hypothesis is can we identify a grid as guide to analyze professional e-mails? If so, can the result be relevant as project knowledge?

Based on this hypothesis, we know that pragmatics intention must be based to context. So, we consider the project context from different aspect: organization and environment. We believe that this context is very helpful to clarify ambiguity of sentence analysis. We show in the example how sender/receiver role can identify problem-solving answer. Adding this analysis to the identification of keywords of messages, as topics can be a first step, towards a structuring of knowledge: Problem related to a topic, possible answers.
We will continue to validate this work on other type of projects. This work can open to identify other grid analysis like: engagement of actors, design-rationale, coordination [12], etc.

Finally, this study is a part of our work on project memory: Keeping track and structuring knowledge in daily work realization of project. We developed techniques to extract knowledge from project meetings [6] and to identify occurrences in order to identify concepts in project memory.

References

[1] Atifi H., Gauducheau N., Marcoccia M.2011.The Effectiveness of Professional Emails: Representations and Communicative Practices , in proceedings of 13th Conference of the International Association for Dialogue Analysis, Dialogue and Representation, Montréal.

[2] Blum Kulka, Shoshana, Juliane House, and Gabriele Kasper (eds.) 1989. Cross-cultural pragmatics: Requests and apologies. Norwood : Ablex Publishing.

[3] Carvalho, Vitor and William Cohen. 2005. On the collective classification of email "speech acts". Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 345–352. New York: Association for Computing Machinery.

[4] De Felice, R; Darby, J; Fisher, A; Peplow, D; (2013) A classification scheme for annotating speech acts in a business email corpus. ICAME Journal , 37 71 – 105

[5] Dieng R. , Corby O., Giboin A. and Ribière M., Methods and Tools for Corporate Knowledge Management, in Proc. of KAW'98, Banff, Canada. 1998.

[6] Ducellier G., Matta N., Charlot Y., Tribouillois F., "Traceability and structuring of cooperative Knowledge in design using PLM", Knowledge Management and collaboration Special Issue of International Journal of Knowledge Management Research and Practices, Vol.11, No.1, 2013, pp 53-61.

[7] Ermine J.L., La gestion de connaissances, J.-L. Ermine.- Hermès sciences publications, 2002.

[8] Kalia K.A., Identifying Business Tasks and Commitments from Email and Chat Conversations, tech. report, HP Labs, 2013

[9] Lampert, Andrew, Robert Dale and Cecile Paris. 2010. Detecting emails containing requests for action. Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), 984–992. Association for Computational Linguistics.

[10] Lampert, A., Dale, R., & Paris, C. (2006). Classifying speech acts using Verbal Response Modes. In Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006), 34-31

[11] Malvache P., Prieur P., Mastering Corporate Experience with the REX Method, Proceedings of ISMICK'93,

International Synopsium on Management of industrial and corporate knowledge, Compiegne, October, 1993.

[12] Matta N., Atifi H., Sediri M. Sagdal M. , Analysis of interactions on coordination for design projects, IEEE proceedings of the 5th International Conference on Signal-Image Technology and Internet based Systems, Kula Lumpur, December, 2010.

[13] Matta, N., Ribière, M., Corby, O., Lewkowicz, M., et Zacklad, M. Project Memory in Design, Industrial Knowledge Management - A Micro Level Approach. Springer-Verlag : Rajkumar Roy, 2000.

[14] Matta N., Ermine J-L., Gérard Aubertin, Jean-Yves Trivin, Knowledge Capitalization with a knowledge engineering approach: the MASK method, In Knowledge Management and Organizational Memories, Dieng-Kuntz R., Matta N.(Eds.), Kluwer Academic Publishers, 2002.

[15] Searle, J.R. (1969). Speech Acts. An Essay in the Philosophy of Language. Cambridge: Cambridge University Press.

[16] Vitor R. Carvalho and William W. Cohen. 2006. Improving "email speech acts" analysis via n-gram selection. In *Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech* (ACTS '09). Association for Computational Linguistics, Stroudsburg, PA, USA, 35-41

[17] Yelati, S.; Sangal, R., "Novel Approach for Tagging of Discourse Segments in Help-Desk E-Mails," Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on , vol.3, no., pp.369,372, 22-27 Aug. 2011

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1415–1420
DOI: 10.15439/2014F193
ACSIS, Vol. 2

Knowledge Portal for Exclusion Process Services

Krzysztof Hauke
Wroclaw University of Economics
Komandorska 118/120,
53-345 Wrocław, Poland
Email: krzysztof.hauke@ue.wroc.pl

Mieczysław L. Owoc
Wroclaw University of Economics
Komandorska 118/120,
53-345 Wrocław, Poland
Email: mieczyslaw.owoc@ue.wroc.pl

Maciej Pondel
Wroclaw University of Economics
Komandorska 118/120,
53-345 Wrocław, Poland
Email: maciej.pondel@ue.wroc.pl

Abstract—Exclusion phenomenon in common understanding denotes processes in which some group of people or individuals are permanently blocked from resources (mostly considered as social exclusion). For sure such sort of phenomena is observed as unwanted not only from "outsiders" but also from local and global society point of view. The exclusion (and its antonym inclusion) phenomenon can be investigated including many aspects: starting from identifying exclusion as a process, multidimensional research in this area up to solutions available in the domain.
In order to be successful in overtaking this phenomenon groups and institution involved in this process should be supported by ICT solutions. The paper consists of five parts which gradually present context of the problem and proposed solutions. After short introduction concerning research background the discussed concept of exclusion processes is presented. In the main section real examples of exclusion processes are investigated which allows to define assumptions of knowledge portal architecture and some examples of performed services. It creates opportunities for formulation final conclusions about the necessity and usability the developed platform.

I. Introduction

Nowadays people, especially in the era of information society, know more and more about a man, groups and the whole nations, about their needs and levels of satisfaction of the human beings alone or of human associations. If so, natural tendency - present in democracy - to treat all citizens equally becomes very important. Any form of discrimination some group of people is against the democratic order and societies and governments try to counteract with this unwanted phenomenon.

The roots of the exclusion phenomena "discovering" (or better reflexion on human sense of equality and justice) can be found in the discourse in France in the mid 1970s (see: N. Rawal review of social inclusion and exclusion - [11]). A bit later H. Silver (see: [1]) formulated three paradigms of social exclusion: solidarity (stressing social dimension of human interactions), specialization (discovering exclusion as a form of discrimination) and monopoly (interpreting exclusion as a consequence of the existing group monopolies). Anyway in older and later approaches to the phenomena research many aspects of social exclusion and inclusion were analysed.

In order to be successful, a problem of exclusion should be investigated, reasons of its occurrence should be discovered and solutions for inclusion should be proposed. According to EU policy the poverty and exclusion problems are very important and responses could be projects prepared in the Europe 2020 strategy (Societal challenges section in Horizon 2020 programme, see: [14]).

Two examples should be stressed as promising solutions in the discussed area:

- GSDRC – Applied Knowledge Services devoted to maintaining knowledge about exclusion phenomena (see: [12])
- Exclusion-Inclusion Suburbs – prepared for knowledge services essential in city environments (see: [13]).

In both cases presented solutions are limited to selected phases or areas of exclusion phenomena. Therefore lack of common platform developed for the whole community seems to be obvious.

The paper is managed as follows. In the next section theoretical background of the investigated phenomena is described including nature of exclusion and inclusion phenomena, reasons of its occurrence is discussed and multidimensional characteristics of investigation is stressed. An essence of knowledge portals developed for modern society is presented in the subsequent section with focus on society needs and functionality such portals, offered architectures and applications useful for different segments of the society. The most innovative part of the research is presented in the main section of the paper devoted to concepts of the knowledge portal addressed to exclusives covering: assumptions, architecture and examples of supported tasks. The paper ends conclusion remarks and future research.

II. Exclusion Processes as Research Challenges

No doubts, exclusion as a phenomenon seems to be very important and difficult problems to solve in modern society. There are many contexts in which exclusionary processes can occur including different objects, or time- and territory-aspects. At least two approaches should be taken into account actors-oriented and capability-oriented.

In accordance of the first one the critical thing is: relationships between "actors" essential for the exclusion. It is very important in understanding the "exclusion" idea as a concept. According to R Saith: [2] "Social exclusion is a

socially constructed concept, and can depend on an idea of *what is considered 'normal'*". So here crucial topic of understanding discussed phenomenon is definition of normality which, in turn, depends on standard living, hierarchy of values, assumed criteria of society organization and the like what finally can be identified with "actors".

On the other hand A.Sen (see: [3]) keeps that an essence of social exclusion relates to 'functionings' and 'capabilities' concepts. Functionings denotes things important in leading life (health, education, cultural life etc.) while capabilities concerns individual combination of different functionings specific for human-beings or some group of people. So, social exclusion relies on inability of achieve certain 'functionings' or difficulties with reaching the goals which leads to deprivation and poverty - unwanted states in of any society.

Multidimensional character of exclusion has been stressed by the following authors: De Haan [4], Bhalla and Lapeyre [6], Burchardt et. al. [5] and Fisher [7]. Potential dimensions cover the different aspects: un/employment, markets (so difficulties with access to goods and services), neglecting of political laws and social relationships. Therefore in research conducted in this domain all aspects of exclusion processes should be examined not only individually but also from more general point of view. There are many intersections between the mentioned dimensions, for example: unemployment has the strong impact on poverty, poverty in turn causes limited access to services and products available on a market in local and global range and so on.

From individual as well as from society point of views the mentioned manifestations of exclusion processes lead to segregation of many sorts, sense of social inequality and conflicts in broader perspectives. Monitoring and investigation of the discussed phenomenon need to be performed continuously and be supported by information and communication technologies or better by specialized knowledge portal(s).

The described phenomenon basically relates to social exclusion which should be separated from voluntary exclusion – see Barry [8]. Exclusion of this short is a specific one and not always is regarded as unwanted process. On the contrary, as intentionally prepared and performed activities cannot be integrated with social exclusion.

The process considered as the solution to neutralize exclusion effects is called social inclusion. The social inclusion can be defined as ..."a process of improving the terms on which people take part in society" - see World Bank definition [15]. In sociology social inclusion means the provision of certain rights to all individuals and groups in society, such as employment, adequate housing, health care, education and training, etc. – see Collins Dictionary [12]. Of course social inclusion – as the process of organizing social life – sometimes is problematic or even inequitable – see Hickey and du Toit [9]. The concept of 'adverse corporation' brings better results because of its implementation in particular contexts – see [9].

Social inclusion as the process of counteraction negative results of exclusion should be supported in many ways by information technologies. Social equality should be enforced by access to information and knowledge available in computer networks; M. Warschauer discussed many aspects of usability and consequences of technology and social inclusion intersection see [10]. In the next part importance of knowledge portals as natural source of information for modern society is presented.

III. KNOWLEDGE PORTAL FOR THE MODERN SOCIETY

Portal - a website presenting the overview and systematic form the most important and best - developed articles, as well as other content related to the topic. It presents readers with the resources available on the subject in a more accessible and attractive than categories, indirectly encouraging the active involvement in the development of the content.

Web Portal (or Internet Portal) - an online information services extended to a variety of Internet functions, available from a single Internet address (compare [17]). Typically, the portal contains information of interest to a wide audience. As an example, you can specify the content of the portal: forum news, weather, web directory, chat, forums and mechanisms for information retrieval in the same or in external Internet sites (search engines). In addition, portals may offer free services such as email, web hosting.

Knowledge Portal (our interpretation) - an online service that includes a generally reliable information about a specific fragment of reality and can be used in the further development of the issues or bind it with another issue. Knowledge expressed in many forms including its own theories is a crucial component of the discussed portals. The classic approach takes into account the following elements:

- beliefs, judgments logically,
- justification belief is justified,
- veracity, the belief is true.

Examples of category knowledge portals:

- social sciences and humanities, such as astronomy, biology, chemistry, genetics, medical science, zoology,
- society, such as: anarchism, atheism, bible, biographies, philately, Hinduism, Judaism, Catholicism, the saints, the religious,
- geography, for example: individual continents, different countries, different cities,
- national, for example, Poland, Germany, USA,
- culture - fiction, film, comics, art, games, anime,
- sports, for example, the Olympic Games, check, ski jumping, rallies,
- technique, such as: architecture, electronics, energy, computers, mobile phones, websites, army,
- social sciences and humanities, such as philosophy, history, psychology, sociology, foreign languages,

The intentions of the portal is to encourage users to set the address as the portal home page in a Web browser, and treat

it as a gateway to the Internet. There is a tendency as a synonym for the treatment of portal website.

One of the oldest knowledge portals is presented in Fig. 1 (see: [19]). Main functionality of the portal concerns to co-operation with experts, e-learning offerings, Case Law Directory apart of typical FAQ (frequently asked questions) capabilities.

Fig. 1. An example of RTI Knowledge Portal

Another example of knowledge portal is depicted in Fig. 2 (available at [20]). The goal of its solution is to convince to company products as well as to offer complete courses devoted to accelerate user skills.

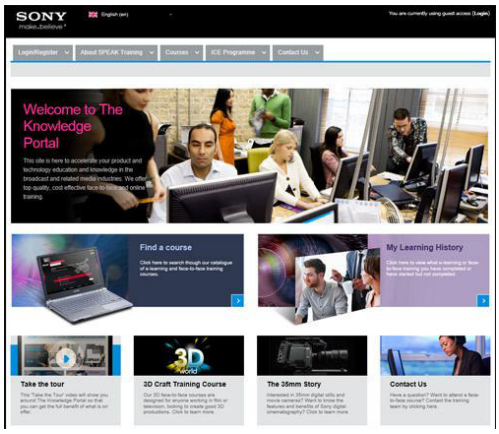


Fig. 2. Working with SONY Knowledge Portal

A useful tool for strengthening a network of people who are excluded can be a portal of knowledge. It serves not only as a tool to convey information, but above all to enrich and exchange their knowledge resources. Knowledge portal provides access to knowledge resources that are owned by all of the entities forming the network issues excluded people. The first impulse to gain specific knowledge resources is to determine your needs. Information goes to the portal, which locates the appropriate resource, a category of knowledge transfer process. Then the knowledge portal allows supplement domain knowledge. In this way, the knowledge portal becomes a compendium of knowledge on the particular matter. Site work requires not only knowledge of information technology, but appropriate organizational structure and knowledge management strategy. The proper functioning of the portal of knowledge brokers play an important role. They should be in the portals of knowledge as individuals or institutions performing oversight knowledge development in the portal. The main task of the broker is:

- assessment of the needs of the knowledge,
- location of resources,
- supervision of the transfer,
- assessment of the degree of absorption and utilization of knowledge.

All operators should be interested in the proper functioning of the portal of knowledge. Knowledge that remains in the resources the organization is the basis for the acquisition of new resources. We have in this case to deal with a spiral of knowledge that all time develops, through which actors are able to develop. Basic knowledge resources of the company are based on five categories of knowledge (compare [18]):

- codified knowledge,
- explicit knowledge,
- knowledge protected,
- tacit knowledge,
- latent knowledge.

The portal should provide exploration of tacit knowledge and latent. These two categories can contribute significantly to the development of the organization, and thus to achieve a competitive advantage. Portals should not only be sauce for the implementation of business functions typically associated with competitiveness, achieving measurable financial results. They should also be used for indirect actions that will achieve the objectives of a typical business. Different types of social organizations also need such solutions. A multitude of knowledge, which allows to solve problems of social institutions requires her to organize and codify the first stage of creating the portal. The next step is to complement this knowledge different solutions at the local or individual. The establishment and operation of such a portal will allow institutions to function more efficiently to deal with the problems of excluded people. Perhaps those who are excluded will return to society. However, their efficiency and effectiveness will be determined only by which time after testing.

IV. A Concept of Knowledge Portal for Exclusives

Considering development of dedicated knowledge portal at least the following aspects should be taken into account: purpose and audience, technology and tools useful for the creation and maintenance stages. Initially our proposal of the portal is expressed in two parts: assumptions and architecture.

Assumptions

We assume the following groups of people are the target users of the proposed portal:

- Policy makers – represented by politicians and officials who are responsible for legal solutions, law creation and adjustments.
- Social workers employed in:
  - Governmental / local governmental units.
  - NGO's – Non-governmental organizations.
- Institutions interested or engaged in the exclusion problem like universities/ scientific organisations and their researchers, media and journalists,
- Independent entities who are interested in the exclusion problem
- Commercial enterprises who develop programs / solutions targeted to help excluded groups

By using another perspective we can divide the users into:

- Corporate users representing authority organisations
- Individuals representing basically themselves that consist of:
  - Authorities that should be verified
  - Regular users

The main basis of the portal assume that every published content is available for all portal users. There will be no confidential matters stored in the portal so there is no need to build sophisticated information protection module.

No confidentiality assumption determines the simplicity of permissions model in the portal. The administrators of the portal will be able to configure it's modules and manage the permissions for every defined role, but there is no need to limit the access to the specific content.

We perceive that if we build a dedicated authentication module with separated credential management it will become a serious pain point for the users who will be forced to create and remember yet another user login and password. That is why we would like to integrate with as many as possible authentication providers that can exchange the data with the portal. Such approach will allow the users to use in our portal the same credentials as they are using in their enterprise systems or social solutions. Our portal will be integrated with:

- LDAP solutions,
- Facebook/ twitter / google accounts.

Our authorisation module architecture approach will allow organisational users to authenticate in our portal with the corporate login and password or even include the portal into the corporate SSO (single sign-on) solution. Individual users will be able to access the portal with the private facebook account and sign in automatically.

Of course we can use the mixed mode of authentication which means that corporate users can use also their private social accounts if their prefer or if the integration with the corporate LDAP will be impossible.

The corporate users of the portal should be verified by:

- Portal administrators,
- Organisation representatives.

In case of the individual users that should be verified – this task will be done by portal administrators.

One of the most important assumptions regards the functional offerings of the portal. It is not designed to work with specific excluded people and solve their particular problems. The main aim of the portal is to connect, enhance the interchange of ideas, knowledge and experience between the experts engaged in exclusion problem. It is designated to inspire all the stakeholders to solve the general problem of exclusion and provide them relevant information and knowledge. That is why the portal will not consist workflow functionality/ application processing capabilities but it is focused on information and knowledge tools processing, ideas management, communication platform and e-learning modules.

Architecture

We propose the portal to be built in traditional three layers architecture presented in Fig. 3:

- **Database layer** – built as a database storing the business and process objects. In this layer all documents, multimedia files, learning objects that has no relational structure will be stored. This layer will also provide the services of reporting and integration with other systems.
  Database layer will be supported also with Knowledge Base. The knowledge to the Knowledge Base will be supplied by:
  - Experts (portal users)
  - Knowledge exploration module that will be operated by portal power users with data exploration skills.
- **Application layer** that will be responsible for whole business logic. It will provide:
  - the basic portal functionality,
  - the communication services that will give the users possibility for on-line bilateral communication, teleconference services, off-line communication,
  - e-learning services with all capabilities for hosting and providing on-line courses,
- **Presentation layer** built in portal technologies (which means that all the portal features will be accessible by web browser). We must take into account that for some users the portal will not be a tool in which they will work every day. That is why from user perspective it is crucial to provide them notifications of all important event happening in portal as new available content, tasks for users, activities expected to be done by users and others. Such notifications will be provided by:
  - Automatically generated emails,
  - Mobile application push notifications,
  - Newsfeed published by social portal as Twitter or Facebook.

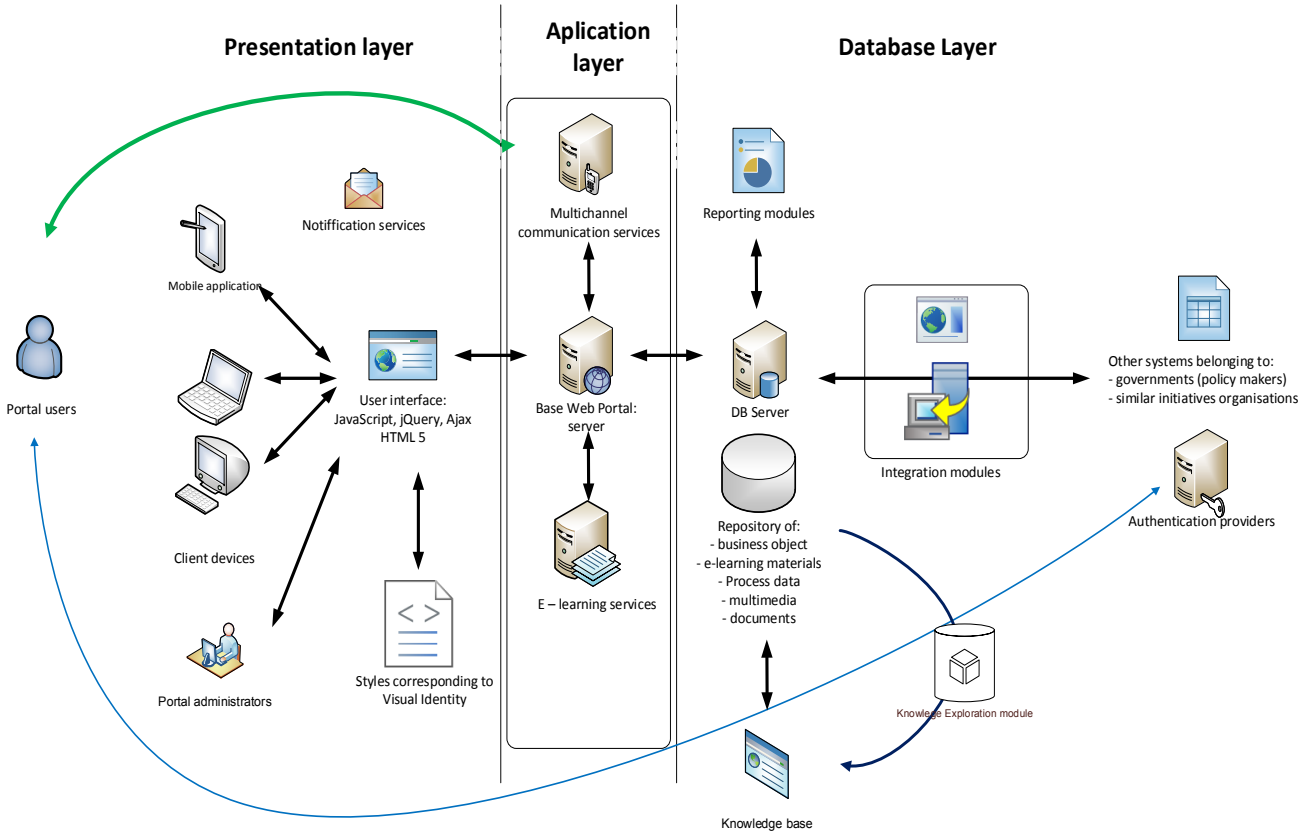


Fig. 3. Architecture of Knowledge Portal for Exclusives

Presentation layer will provide end users access to the information and knowledge stored in the database layer with regard to: users permissions and users preferences (content should be targeted to end user's needs and requirements).

V. Conclusions and Further Research

One of the most important problems of the modern society at the local, national and global levels is exclusion processes. The next findings can be formulated from the research:

- phenomena exclusion and inclusion in society are complex and multidimensional. Therefore investigation of such processes is difficult also because of differentiation of components, approaches and dimensions,
- information and communication technologies must support all processes belonging to registration and all services typical for exclusion and inclusion phenomena in modern society,
- the best solution for the discussed problems is specialized knowledge portal – proposed in the paper as a form of an initial version. At the beginning architecture will be developed for some narrow application and systematically extended covering new areas and levels of supporting.

Further investigations can be devoted to more deeply analysis of the problem in order to fulfill requirements of knowledge portal users. For example specification of knowledge bases and courses essential at local and/national level of administration.

References

[1] Silver, H., 1994, "Social Exclusion and Social Solidarity: Three Paradigms", International Labour Review, Volume 133, Numbers 5-6, pp. 531-578.

[2] Saith, R., 2007, Social Exclusion: The Concept and Application to Developing Countries [in:] Stewart, F., Saith, R. and Harriss-White, B., (eds.), Defining Poverty in the Developing World, Palgrave, pp. 75-90.

[3] Sen, A., 2000, Social Exclusion: Concept, Application, and Scrutiny, Asian Development Bank.

[4] De Haan, A., 1999, Social Exclusion: Towards an Holistic Understanding of Deprivation, Department for International Development, London.

[5] Burchardt T., Le Grand J., and Piachaud D., 2002, Introduction, [in:] Hills, J., Le Grand, J. and Piachaud, D., Understanding Social Exclusion, Oxford University Press, Oxford.

[6] Bhalla, A. and Lapeyre, F., 1997, "Social Exclusion: Towards an Analytical and Operational Framework", Development and Change, Volume 28, pp. 413-433.

[7] Fischer, A., 2011, Reconceiving Social Exclusion, BWPI Working Paper 146, Brooks World Poverty Institute, Manchester.

[8] Barry, B., 1998, Social Exclusion, Social Isolation and Distribution of Income, Centre for Analysis of Social Exclusion, London School of Economics, London.

[9] Hickey, S. and du Toit, A., 2007, Adverse Incorporation, Social Exclusion and Chronic Poverty, Working Paper 81, Chronic Poverty Research Centre, University of Manchester.

[10] Warschauer M., 2004, *Technology and Social Inclusion. Rethinking the Digital Divide*. The MIT Press, Massachusetts Institute of Technology.
[11] Rawal N., 2008, "Social Inclusion and Exclusion: A Review", Dhaulagiri Journal of Sociology and Anthropology Vol.2. pp. 161-180.
[12] Social Exclusion portal: http://www.gsdrc.org/go/topic-guides/social-exclusion/definitions-and-different-understandings-of-social-exclusion [2014-04-23].
[13] Exclusion-Inclusion in Suburb: http://www.hioa.no/eng/About-HiOA/Centre-for-Welfare-and-Labour-Research/NOVA/NOVA-Projects/Prosjekter-migrasjon-og-transnasjonalitet/Avsluttede-prosjekter/2011/Exclusion-and-inclusion-in-the-suburb/(language)/eng-GB [2014-04-23].
[14] H2020 sections: http://ec.europa.eu/programmes/horizon2020/en/h2020-sections [2014-04-23].
[15] World Bank website: http://www.worldbank.org/en/topic/socialdevelopment/brief/social-inclusion [2014-04-23].
[16] Collins Dictionary website: http://www.collinsdictionary.com/dictionary/english/social-inclusion [2014-04-23].
[17] Internet portal http://en.wikipedia.org/wiki/Internet_portal [2014-04-23].
[18] Knowledge concept http://en.wikipedia.org/wiki/Knowledge [2014-04-23].
[19] RTI Knowledge Portal http://rti.img.kerala.gov.in/RTI/index.jsp [2014-04-23].
[20] Sony Knowledge Portal https://training.sony-europe.com/ [2014-04-23].

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1421–1430
DOI: 10.15439/2014F163
ACSIS, Vol. 2

Data Warehouse as a Source of Knowledge Acquisition. An Empirical Study

Moh'd Alsqour
Wroclaw University of Economics
Komandorska 118/120,
53-345 Wrocław, Poland
Email: mohsqour@wp.pl

Mieczysław L. Owoc
Wroclaw University of Economics
Komandorska 118/120,
53-345 Wrocław, Poland
Email: mieczyslaw.owoc@ue.wroc.pl

Abdulrahman S. Ahmed
Wroclaw University of Economics
Komandorska 118/120,
53-345 Wrocław, Poland
Email: aatukow29@yahoo.com

Abstract—**The main objective of conducting this research is to investigate empirically the role of data warehouse (DW) in enhancing decision-making or, more precisely, the researchers surveyed the top management of Jordanian firms regarding the role and importance of DW in improving the effectiveness of decision-making. It is believed and assumed that meaningful and significant information can be acquired from DW which provides valuable knowledge to support business process and decision-making.**

A mail questionnaire survey was regarded as the appropriate method for gathering data. The questionnaire was developed based on the findings from related literature and other related research questionnaires. All the firms (277), which were listed on Amman Stock Exchange (ASE), at the time of data collection, were selected over half answered positively. The researchers arrived at scores of significant and remarkable results regarding DW and its role in enhancing the process of decision-making. The survey's findings showed that the percentage of implementing DW in the Jordanian firms involved is 35%. In general, the respondents had a positive attitude towards the implementation of DW.

I. Introduction

Today's organizations face a very hard time, largely as a result of competition, globalization, automation and scarcity of resources. As the business environment is changing. Companies rely more and more on changing technology. At the same time, those companies are likewise evolving. In this changing environment, companies are much more eager in getting immediate and accurate information to make better decisions. Successfully supporting managerial decision-making has become critically dependent upon the availability of integrated and high quality information organized and presented to managers in a timely and easily understood manner [1].

Nowadays there is a need for more advanced and accurate systems. Today's systems are more likely to produce more accurate information than previous systems. Traditional systems, indubitably, are more likely to produce less accurate data and information which lead to bad and incorrect decisions. However, the situation will be different if the firms implement data warehouse (DW) technology, which has emerged as a key source and powerful tool for delivering and accessing information for decision-makers ([2]; [3]; [4]; [1]). Since the 1990s [5], DWs have been an essential information technology (IT) strategy component for large and medium-sized global organizations [5].

Large organizations are facing significant challenges in maintaining an integrated view of their business [6]. In an ever more complex and competitive world, the complexity of the organizational context and the management task involving decision-making and assessment of information has increased [7]. Decision-making in those environments involves large data volumes and includes a wide variety of decision tasks [8]. In such environments it is important to assure decision-makers of the quality of data they use ([13; [8]). DW systems are perceived to be important tools in the modeling of that complexity, however, reports of high failure of DW systems are common ([7]; [5]; [9]). [10] claim that DW hastens the process of retrieving information needed for decision-making.

Despite the recognition of data warehousing as an important area of practice and research, there is little empirical research [1] about implementation of DW in general ([11]; [9]; [1]).

The changing business environment has an impact on the nature of decisions and decision-making drivers [12]. Timely and informed decision-making is becoming crucial for the long-term success of businesses [13]. [14] claims that business decisions must be made with speed and accuracy if organizations are to remain competitive. There is quasi-consent that DW provides more detailed and accurate information for decision-makers to improve their decisions. Considering the usefulness of DW there has been little research in Jordan on DW, i.e. it has been comparatively less investigated in Jordan. Therefore, the main focus of this study is on the advantages of DW as a provider of information to the process of decision-making. It investigates mainly the relation between decision-making, the need for information and the employment of DW in Jordanian firms.

In addition, this study investigates how top management of Jordanian firms perceives the effectiveness (usefulness) of DW as a source of reliable and accurate information for decision-making. Although many studies to DW have been published, they have been concerned with technical issues. However, it has been understood recently that the information systems (IS) failure is due to psychological, environmental, organizational issues etc. rather than

This work was not supported by any organization

technological issues, hence individual differences must be addressed [15]. Overall, there is a scarcity of empirical studies that examine the data warehousing success within an integrative model [9].

II. AIMS OF THE STUDY

In this paper, a field study of DW and its role in easing and enhancing the process of strategic decision-making among top managers in Jordanian firms were investigated. Therefore, the main aim of this study is to investigate the role, which is played by DW, in decision-making.

In realizing this aim, the researchers believe that the following matters in particular deserve careful investigation largely due to their close connection with the aim and its achievement:

1. To investigate the relationship between the implementation of DW by Jordanian firms and the effectiveness of strategic decision-making (assuming that there is a positive association between the implementation of DW and the degree of decision performance).

It has been ascertained that DW is superior to traditional database and improves the process of decision-making. Therefore, one of the study's aims is to investigate empirically whether or not the DW provides better and more accurate information. It is anticipated in this study that the enhancement of decision-making's process might drive Jordanian firms to employ DW. Thus, another aim of this study is to identify the Jordanian firms' reasons for implementing DW.

2. To identify the grounds of implementing DW by Jordanian firms.

A study of a large number of data warehousing practitioners and experts by [16] showed that the implementation of DW was motivated more by internal pressures than external. A majority of the respondents said that the need was information related. [17] are of the opinion that improving access to information and delivering better and more accurate information are motivations for using DW.

III. THE STUDY'S QUESTIONS AND HYPOTHESES

The hypotheses are formulated in the light of the study theoretical and conceptual framework and partially based on related previous studies. It is assumed that the outputs of DW, such as data and information, have more positive effect on the process of decision-making by comparison with traditional database systems. This assumption leads the researchers to question whether there is a direct positive relationship between improving the process of decision-making and the implementation of DW? In other words, does the DW provide relevant, reliable and sufficient information for decision-makers to take and achieve sound and effective decisions? This question leads to hypothesis 1.

Hypothesis 1: There is a strong association between employing DW in Jordanian firms and improving the process of decision-making (soundness and effectiveness of decisions).

In addition to the principal objective of this study, the researchers attempt to achieve multiple aims through this study, including the grounds, which drove Jordanian firms, of implementing DW. This aim led to question what the Jordanian firms' grounds of implementing DW are. This question, in turn, led the researchers to formulate hypothesis 2.

Hypothesis 2: The dearth of reliable information for taking decisions, the highly competitive environment, the need for more accurate, reliable, relevant and timely information, the changes in manufacturing technology, techniques and processes, the unreliability of existing systems of decision support and inability of the existing systems of decision support to provide reliable, useful and relevant information to the process of decision-making are the major grounds of implementing DW in Jordanian firms.

IV. THE NOVELTY VALUE OF THE STUDY

This study was applied in Jordan, which is one of the developing countries in the Middle East. As Jordan's firms are not in isolation from the rest of the world, they are also influenced by the current competitive environment. The implementation of advanced and recent innovations, such as DW, is essential for the firms in developing economies, such as Jordan. Therefore, investigating their implementation in Jordanian firms is well worth considering. However, a few empirical studies, which had been identified on Jordanian firms, prompted the researchers to examine whether this innovation has been successfully implemented in Jordan. Because of the need for further investigation on this topic, considering its importance and the shortage of the empirical research, the researchers have every reason to investigate this issue which has not been given the proper attention in Jordanian environment.

Despite the exaltation and adulation of DW, There is a need for evidences that the implementation of DW improves the quality and accessibility to information. Therefore, this study aims to practically investigate whether or not the implementation of DW improves the quality and accessibility to information and enhances the process of decision-making. [18] drew attention to the role and importance of information accessibility. The authors claim that the information accessibility is a precursor of information quality-it has a significant impact on the information's usage, and consequently is an indicator of the DW's success in storing and processing information.

In addition to information's quality, previous literature and studies on DW have emphasized the importance of accessing information. The accessibility to information and their quality are crucial to the success of their use. It has been claimed that the use/ implementation of DW improves the quality and accessibility to information, and consequently leads to sound decisions. In other words, it leads to more fact-based decisions. DW has, according to

DW literature, the ability to store a vast amount of data in a usable and appropriate form for the decision-makers' needs and uses. Although a wide range of primary and secondary sources has emphasized the importance and role of information quality and accessibility in enhancing the process of decision-making, little empirical research has been conducted so far. Such claims need to be tested empirically. It is essential, therefore, that the researchers investigate whether or not DW provides easy access to data and information, frequent, accessible and timely reports and more accurate, useful, reliable, complete and relevant information to decision-makers.

Based on extensive literature review, the researcher has identified that firms are often unsuccessful due to a lack of appropriate information or more precisely their inability to get the right information to the right person at the right time. The availability of apposite information to decision-making helps managers in taking reliable decisions, which improve the firm's performance. For this reason, this study is one of the few empirical studies (e.g. [4]), which attempts to examine the effect of DW on decision effectiveness. In addition, previous research has not empirically tested its effectiveness in DSSs contexts in Jordanian firms "to the researchers' knowledge".

Additionally, the researchers have not found and are completely unaware of any empirical studies regarding the implementation of DW in Jordan. Therefore, it is hoped that the findings of this study give the readers and those who are interested in these issues practical insights into DW's field in Jordan. To some extent, it is one of the academic contributions. It contributes to our understanding of the DW in general and in Jordan in particular, and may form a basis and motivation for future research in the important fields. It is also believed that the final outcome of this paper adds up to the improvement and development in DSS, such as DW, by helping their users and developers to be more aware of the data and information's quality.

V. RESEARCH METHODOLOGY

The writing of this paper is passed in different phases. In the initial phases of the study, renowned journals, publications, conferences proceedings and books were reviewed. In addition to these sources, the findings of numerous empirical studies were researched and analyzed. Based on the literature review the study' hypotheses were formulated. As the researchers previously pointed out, the sample of the study comprises all the 277 firms, which are listed on ASE at the time of the data collection. Thus, a questionnaire was found to be the best instrument for collecting the data in this study. During the next phases, therefore, a survey questionnaire was conducted with the top managers of Jordanian firms. The questionnaire, which is used in this study, is based on previous studies and the researchers' assessments and discretion and adapted to suit the objectives and requirements of the study. The primary objective of the questionnaire is to amass appropriate data and responses from the potential respondents for testing the significance of hypotheses.

In order to answer the study's questions, fulfill its aims and find out whether or not there is a positive relationship between independent and dependent variables, the study's hypotheses were statistically tested. The data, which were collected, is very quantitative in nature. Therefore, in the final phases of the study, the data were statistically analyzed by employing Statistical Package for the Social Sciences (SPSS) in order that proper descriptive and inferential statistics to analyze the results and draw conclusions can be reached, including means, frequencies, standard deviation, t-test, f-test and chi-square.

VI. LITERATURE REVIEW

The concept of data warehousing has evolved out of the need for easy access to a structured store of quality data that can be used for decision-making [19 p. 5]. Organizations have vast amounts of data but have found it increasingly difficult to access it and make use of it. This is because it is in many different formats, exists on many different platforms, and resides in many different file and database structures, and, as a result, organizations have had to write and maintain perhaps hundreds of programs that are used to extract, prepare, and consolidate data for use by many different applications for analysis and reporting [19 p. 5].

As an attempt to solve the problem, DWs were introduced. DWs have become the focal point for decision support in organizations today [20] and emerged as a key platform for the integrated management of decision support data in organizations [3]. [19 p. 5] claim that the data warehousing offers a better approach. Data warehousing implements the process to access heterogeneous data sources; clean, filter, and transform the data; and store the data in a structure that is easy to access, understand, and use. The data is then used for query, reporting, and data analysis [19 p. 5]. [21] also claim that the data warehousing has emerged as an effective mechanism for converting data into useful information.

DW systems offer efficient access to integrated and historical data from heterogeneous sources to support managers in their planning and decision-making [22]. [23] also claim that data warehousing provides an infrastructure that enables businesses to extract, cleanse, and store vast amounts of data. According to [19 p. 1], businesses of all sizes and in different industries, as well as government agencies can realize significant benefits by implementing a DW. Most medium to large organizations, according to [24], operate DWs.

It has been claimed that the main DW's role is to support decision-making. However, the role of DWs has been broadened. [25] state that DW provides information from external data sources for decision-making. DW has the potential to create radical changes to existing business processes and is often viewed within the context of business process reengineering [11]. Accordingly, [26] claims that

DW gives business' users the ability to analyze data. [27] also claim that DWs enable organizations to exploit decision-making.

According to [28], DWs provide the basis for management reports and decision support. They added that the purpose of DWs is to take that vast amount of data from many internal and external sources and present them in meaningful formats for making better decisions. In support of the above mentioned claims participants to a study by [3] agreed that the Return on Investment (ROI) for the DW was well justified through considerable gains in productivity and enhanced quality of customer service. Moreover, in an independent detailed study of 62 organizations worldwide [34], the major findings of International Data Corporation (IDC) based upon 62 case studies of organizations that have successful DWs in use are an average three-year ROI of 401percent was realized by organizations building DWs. Although this study is primarily focused on quantitative information, there are several qualitative benefits [34], such as providing standardized, clean and value-added data to create information from disparate sources. In addition, the DW makes the data available across corporate organizations and provides the needed information quickly.

The DW is developed in order to support the integration of external data sources [29] for the purpose of advanced data analysis. [40 p. 35] argues that a DW produces tangible impacts to the quality of day-to-day business transactions. Previous research on DW has produced some encouraging findings about its benefits and indicated that a DW can offer several benefits to an organization [11], such as enabling effective decision support; ensuring data integrity, accuracy, security, and availability; easing the setting and enforcing of standards, facilitating data sharing, and improving customer service [35]. [30] presented time savings for data suppliers and for users, more and better information, better decisions, improvement of business processes, and support for the accomplishment of strategic business objectives as benefits from data warehousing.

Furthermore, [2], who examined data warehousing at the Housing and Development Board (HDB) in Singapore, found that the main benefits of the DW, which were developed by HDB, are enabling the users to have access to consistent and reliable data in a timely fashion which facilitated forecasting and planning efforts and improved decision-making. In addition, a study by [31] revealed that DW appears to be used more to improve the flow of information in an organization than to change the way the organization does business. The authors found that more and better data is the greatest realized benefit from DW. Moreover, a study by [32] identified time savings, new and better information, and improved decision-making as benefits of DW.

[33] conducted an explanatory case study at a financial services organization to investigate how DW provides decision support to individual decision-makers. The results showed that the organizations successfully automated the retrieval and input of data for front-end users. [40 p. 33-34], who interviewed people from seven companies, found that the benefits of implementing DW were improving asset management, reducing customer support costs, auditing billing practices, terminating unprofitable product, reducing staff requirements and running the business. [28], who described the DW implementation at Blue Cross and Blue Shield of North Carolina (BCBSNC), claim that the DW had resulted in many organizational benefits, including better data analysis and time savings for users. [34], who looked at the DW of Egypt's Cabinet Information and Decision Support Center, found that the DW provides a lot of benefits to the users, including ease of access to the information, fast and more consistent reports, support the decision-makers and integrating the data from various sources.

Additionally, [4], who conducted a laboratory experiment in 2006, found that the implementation and use of DW improves the DSS users' decision performance, by which he means improving the quality of the DSS by adding a DW can improve information availability and quality and enhance DSS users' decision performance. In conclusion, the study showed that DW can have a positive impact on decision-making.

[35], who described an example of implementing DW in medical institutions, found that the DWs provide the users access to important information. [36], who conducted a survey to find out how DW assists decision-making process in healthcare, found that the DW provides better accessibility to data, integrated disparate data sources and improved decision-making. [37] found that all companies, which are studied, recognized some benefits such as cost reduction, reach-out to other markets, increase in sales, time saving in amount and preparation of reports and more effective decision-making based on the obtained information. [38], who conducted two case studies on American Airlines and Hallmark Cards, found the easy to use, speedy information retrieval, more information, better quality information, improved productivity, and better decisions as benefits of DW. [39], who examined the implementation of DW in public security, found that the DW is very important in improving the comprehensive ability of leadership and decision-making. In addition, it quickly and efficiently integrates heterogeneous data sources.

Previous literature on DW, such as [31] and [22], claims that the DW does not create value by itself; the value comes from the use of the data in the DW. [22] claim that improved decision-making results from the better information available in the DW. By making the right information available at the right time to the right decision-makers in the right manner, DWs empower the users with the ability to make the right decisions [40]. [31] also claim that this use can result in numerous benefits, including more and better information, improved user ability to produce information and reduced effort by developers to produce information. [41] also maintain that DWs have tremendous potential to present information. The greatest potential benefits of the

DW occur when it is used to redesign business processes and to support strategic business objectives [30].

[42] also identified many different measures of success; these include benefits such as data accuracy, useful information, accurate information, ease of use, user satisfaction, time to make decision and increased revenue. However, [43] claims that despite clear evidences that many DW projects have resulted in interesting business benefits, there are also many examples of cost and schedule overruns and dissatisfaction regarding the results from these projects. [44] argue that DW is one of the key developments in the IS field and has plentiful benefits. In addition, [45] indicates that the introduction of a new IS into an organization should deliver multiple benefits.

Since the early 1990s, DWs have become the technology of choice for building data management infrastructures [5] and been investigated and implemented around the world in many areas and by many researchers, authors, and scholars [46]. According to [47 p. 13], the early successful implementation of DW dates back to mid-1980s at ABN AMRO Bank (Netherlands). The author claims that the end-user's needs were the key feature behind the implementation. As a result, those requirements were modeled rather broadly, and all available data was stored in the DW. In fact, in the first few years of general use, its usage had grown at an annual rate of 50%, and by 1995 the DW had supported some 3,000 end-users. [47 p. 17] also mentioned that a study of 62 DW projects, which was conducted in 1996, showed an average return on investment (ROI) of 321% for these enterprise-wide implementation in an average payback period of 2.73 years.

In addition, [48], who investigated whether lodging companies are involved with DW technology through a sample of twelve large lodging corporations, found that the most of hotel corporations in the study were using their DWs to support market analysis. However, [13], who conducted a survey on a large Australian public organization, found that 60% of the users were with limited or no usage (or were anticipating the use in the future) of the DW. The data also helped the users to make informed decisions and the data, which was retrieved from the DW, was also presented to the senior management and other strategically oriented sections in the form of reports i.e. annual and quarterly reports.

Similarly, [11], who surveyed DW's managers and data suppliers from 111 organizations in different regions of the United States (US), also found that all companies had operational DW and nearly all of them considered that their initiative is successful. In other words, 26% of the respondents considered it runaway success. Another survey by Forrester showed that 878 IT decision-makers in the US enterprises were somewhat satisfied with the accessibility and quality of customer information; 82% of these respondents are satisfied or very satisfied [49]. [27] also found that about 55% of 196 respondents firms (107) in two major states in the US had already adopted and used DWs.

Similarly, [23 p.192-196) found that 60% of the respondents consider the functionality of their DW below expectations. 40% of the dissatisfied group was actually still using it. This means that 80% of the respondents were still using their internal DW, which is a good indication of the overall degree of satisfaction.

In addition to those studies, [50], who reported the results of a survey which was conducted at The Data Warehousing Institute (TDWI) World Conference in New Orleans 2003, found that 45% of the respondents had been already in production with their current DW or implementing a second or third release.

In Taiwan, [51], who conducted a survey on Taiwanese banks, found that 53.33% of 30 respondent banks had DW functions or capabilities, 31.25% (5) out of 16 respondent banks had already implemented the DW and 68.75% (11) were still in the process of development to implement DW. 85.71% (12) of 14 banks which had not implemented the DW were evaluating the possibility and/or potential of adopting DW. Another study in Asian countries was conducted by [52]. [52], who surveyed 115 users of DW in four Korean financial companies, also found that the participated companies had been using their data warehousing systems for approximately two years. In addition, the findings also showed that all the respondents were end-users of DWs in their companies, using the systems mainly for financial analyses. The majority of data warehousing users (approximately 70–75%) use the DWs regularly, i.e. every day.

The 2007's IBM Data Warehousing Satisfaction Survey showed that 56% of those questioned were very successful with their DW (200 end-user enterprises were participated), however, 43% of the respondents acknowledged the need for improvement, as there are still a number of business and technical challenges confronting the enterprises which make use of DWs according to the survey [53]. The success of DW's implementation, according to [53], is growing as 56% of the respondents were very successful and satisfied. [54], who conducted a survey on 84 users of DW, found that the majority of respondents (73%) in the surveyed firm were successful in obtaining and accessing the needed data and information from the DW, only two respondents indicated that they were not at all successful, while 56% indicated somewhat successful and 13% indicated very successful. In addition to these results, 67% and 33% of the respondents rated the importance of the obtained information in performing their job better as vital and somewhat important respectively.

Numerous studies have also shown successful implementation of DW. For example, [55] used a case study and conducted a series of interviews at Continental Airlines. The results showed that the organization has realized an enviable level of DW maturity and significant cumulative benefits. [56], who surveyed 244 members of TDWI, found that 51.2% of the respondents had at least one DW application and 30.3% were still in development stage.

17.2% were still in planning stage and 1.2% of respondents had no any efforts made in their organizations to implement the DW. This result showed that more than 80% of respondents had implemented DW.

According to [9], there is a scarcity of empirical studies that examine the DW success. In the study population, i.e. Jordan, there is not a shred of empirical evidence that the DW has been investigated or showed the degree of DW implementation, to the best of the researchers' knowledge. However, the literature review was limited to materials which have been published in English language only– evidence in languages other than English could be possible. Therefore, this paper investigates the extent to which Jordanian firms implemented DW and the reasons for implementing DW. It is aimed at providing empirical evidence, thereby extending the body of research regarding the implementation of DSSs in general and DW in particular.

VII. RESULTS, FINDINGS AND DISCUSSION

As mentioned earlier, the study's sample consists of the Jordanian firms' top management. [57 p. 142) claim that although the large populations are referred to as ideal populations, sampling every person in these populations is not realistic or doable. They believe that time, money, and other restrictions make it impossible for the average survey researcher to reach all members of an ideal population. Therefore, the researchers have to forgo these grand expectations and select a smaller or realistic population. This argument forms the basis for this study's population, therefore, the study's population, which is selected, is only the Jordanian firms that are listed on ASE. It is believed that this category is more likely to use the outputs of DW for strategic decision-making, i.e. the data, which is obtained from these respondents, fulfill the purpose of this study. Therefore, the potential respondents were solicited for their opinions. The covering letter was addressed to the top management of each firm. The questionnaire, along with a self-addressed stamped envelope (SASE) and a covering letter were posted to the potential respondents. The potential respondents were requested to return the completed questionnaires in the enclosed SASE.

As the main aim of this study is to investigate the role of DW in strategic decision-making, the choice of the target population (Jordanian firms which are listed on ASE) and potential respondents (top management) is made on the basis that those firms have sufficient resources to implement such expensive and time-consuming systems and the respondents have sufficient knowledge about their firms and take the strategic decisions. In other words, they are the main consumers of the DW reports. In addition, the companies' Guide by ASE is the only listing that specifically covers all sectors and industries in Jordan. This directory lists the names, titles, and the general information about the listed companies. Top management of Jordanian firms, which are listed on ASE, constitutes the population of this study. A total of 277 questionnaires were sent out by post along with a covering letter. The potential respondents were assured that all data, which would be provided by them, would be treated by complete confidentiality.

The results of the study's sample analysis are shown in the table below (Table 1). In this table, the sample results are broken down by responses.

TABLE 1.
THE STUDY'S POPULATION ANALYSIS

The state of the survey's responses	Number	Percentage %
Usable responses	120	43.3
Unusable responses	20	7.2
Non-response	137	49.5
Total number of responses	277	100

Source: The figures are based on the responses to the survey.

As can be seen from these results, 140 completed questionnaires were returned to the researcher's address with a response rate of 50.5%. According to these figures, usable questionnaires accounts for 43.3% of the total sample. 20 of the questionnaires were discarded as unreliable, i.e. there were many essential questions missing from the questionnaires. To sum up, all the firms (277), which were listed on ASE at the time of data collection, were selected. 140 filled questionnaires were returned generating 50.5 %. This response rate somehow on average comparison to many similar studies such as ([58]; [59]). [58], who studied 350 Jordanian companies, had a response rate of 30%. [59], who used a self-administered questionnaire and targeted similar population in Jordan, had a response rate less than 50%.

The reliability, internal consistency and validity of the Likert scale questions are assessed by using Cronbach's alpha. It was found that Cronbach's alpha is the most popular method for assessing the reliability of scales. It has been used by many researchers, including [60]-[63]. Cronbach's alpha determines the internal consistency of the items in a survey instrument (questionnaire) to assess its reliability [64]. Table 2 shows Cronbach's alpha for all Likert-scale questions. In addition, the table demonstrates the mean, SD, sum of item variances (V) and standard error (SE) for the rating scale questions.

TABLE 2.
THE STATISTICAL ANALYSIS OF RELIABILITY AND VALIDITY

Questions	Items	Mean	SD	V	Cronbach's alpha	SE
Q 11	6	4.80	1.64	2.70	0.711111	1.0
Q 14	12	4.20	5.21	27.2	0.998288	1.6

Source: The figures are based on the study's questionnaire.

As can be seen from these figures, Cronbach's alpha coefficient is more than 0.9 for question 14. Overall, the results of the statistical computations show high reliability of

both questions, as evidenced by the high value of the Cronbach's alpha, i.e. more than 0.7.

The survey's results showed that 80% of respondents were males and 20% were females. In addition, the survey results were broken down by age. The 51-55 age bracket is the highest among the others (33.3%). According to these figures, the 56-60 age bracket accounts for 29.3% of the respondents. From the data, which were obtained, it is apparent that only 5.8% of the respondents are below 40 years old.

The results also showed that almost half (56) of the 120 respondents surveyed have a bachelor degree. 32.5% of the respondents are postgraduate, i.e. have master degrees, considerably more than those who hold PhD (18.3%). Moreover, one-third of the respondents (40) have from 11 to 14 years of experience. Similarly, almost a third of the respondents (29.2%) have more than14 years of experience, considerably higher than those who have from 3 to 6 years (8.3%) and less than 3 years (1.7%).

The survey results were broken down by industry group. The results showed that the financial services industry is the highest among the others (14.2%). According to these figures, the insurance industry accounts for 11.7% of the industry groups. The survey's figures also show that 9.2% of the firms are within the banking industry, considerably higher than the 0.8% of firms within the glass and ceramic, textile, leather and clothing and utilities and energy industries and 2.5% within the pharmaceutical and medical, electric and educational services industries.

Almost a quarter (30) of the 120 firms involved in the survey has from 100 to 500 employees. The figures also show that 20% of the firms are in the range 501 to 900 employees, considerably more than those who are in the ranges 901 to 1300 (14.2%), 1301 to 1700 (10.8%), 1701 to 2100 (9.2%), 2101 to 2500 (8.3%) and more than 2500 (5%). In addition to these results, almost a quarter of the firms have a capital from 51 to 100 million dollar. The survey also found that 16.7% of the firms are in the 10 to 50 million capital range. A small proportion (4.2%) of the firms is in the range of 1501 to 2000 million. Another and last characteristic to be discussed is the annual revenues of the firms. The results showed that the highest percentage (23.3%) of firms involved has annual revenue from 11 to 50 million. The results also show that 20% of the firms involved are in the 51 to 100 annual revenue range.

This result is virtually identical to the 1 to 10 revenue range (18.3%). the survey also found 5 (4.3%) of the 120 firms involved in the survey have more than 1000 million, considerably lower than the rest of the categories. Based on these figures, it can be concluded that the rate of implementing DW in Jordanian firms is very low and less than other countries. For example, [27] found that about 55% of 196 respondents firms (107) in two major states in the continental US had already adopted and used DWs. [56], who surveyed 244 members of TDWI, found that 51.2% of the respondents had at least one DW application. However, the results of this study are similar to some previous studies' results. For example, [51], who conducted a survey on 16 Taiwanese banks, found the rate of implementing DW among those banks is 31.25%.

The 42 respondents, who their firms implemented DW, were solicited for their opinions regarding the reasons behind the plan (decision) to implement a DW. The likely and expected reasons, which were measured on a five-point scale ranging from 1 (unimportant) to 5 (extremely important), are shown in Table (3). The table shows the mean and standard deviation (SD) for all the select reasons.

As can be seen from these results, the need for more accurate, reliable, relevant and timely information (mean= 4.28) were the most important reason to implement DW. According to these figures, the existing systems of decision support have not provided reliable, useful and relevant information to the process of decision-making (mean=4.24) were the next second important reason to implement DW. Other important reasons include the existing systems of decision support have not been reliable (mean= 4.09) and there is a dearth of reliable information for taking decisions (mean=4.07). From the data in the table, it is apparent that the highly competitive environment created the need to replace the existing systems of decision support and changes in manufacturing technology, techniques and processes created the need to replace the existing systems of decision support mean= 2.59 and 2.47 respectively were the least important reasons to implement DW.

The researchers believe that some of the keys to measure success are satisfaction and approval. This is why it is essential that DW (if it is to be a success) must satisfy the needs (requirements) of its users. Similarly, the users need to be satisfied of the need for DW.

TABLE 3.
THE REASONS FOR IMPLEMENTING DW

The reasons for implementing DW	Number	Minimum	Maximum	Mean	SD
There is a dearth of reliable information for taking decisions	42	2	5	4.07	0.97
The highly competitive environment created the need to replace the existing systems of decision support	42	1	4	2.59	1.11
The need for more accurate, reliable, relevant and timely information	42	3	5	4.28	0.81
Changes in manufacturing technology, techniques and processes created the need to replace the existing systems of decision support	42	1	5	2.47	1.19
The existing systems of decision support have not been reliable	42	2	5	4.09	1.01
The existing systems of decision support have not provided reliable, useful and relevant information to the process of decision-making	42	2	5	4.24	0.93

Source: The figures are based on the responses to questionnaire.

To measure the success of implementing DW, the respondents, who their firms implemented DW (42 respondents), were solicited for their opinions regarding the derived satisfactions from implementing DW. A seven-point scale from 1 (strongly disagree) to 7 (strongly agree) was used to measure these responses.

Table 4 shows that the great majority of respondents, since the least mean is 5.43, are very satisfied with the derived benefits of implementing DW. As can be seen from these results,most of the respondents believe that DW is useful in making better business decisions (mean=6.69).

TABLE 4.
THE DEGREE OF SATISFACTION FROM IMPLEMENTING DW

Satisfaction and approval of implementing DW	Number	Minimum	Maximum	Mean	SD
The information in your firm's DW is sufficient for taking sound decisions	42	4	7	6.14	0.89
You are content with the benefits of your firm's DW	42	4	7	6.26	0.85
The DW's information meets the requirements of your task	42	5	7	6.17	0.86
DW is a reliable source of information	42	4	7	6.21	0.95
DW's information is beneficial to different areas of decisions	42	4	7	6.19	0.83
DW's information is used in different areas of decisions	42	5	7	6.14	0.81
The DW offers user-friendly query capability to decision-makers	42	3	7	5.83	1.10
DW is a reliable information system	42	4	7	6.02	1.04
DW is a user-friendly information system	42	3	7	5.43	1.32
DW is useful in making better business decisions	42	5	7	6.69	0.66
DW is useful for decision maker's task	42	5	7	6.62	0.62
DW can output information much more quickly	42	5	7	6.33	0.72
The benefits of implementing DW exceed its cost	42	4	7	6.14	0.89

Source: The figures are based on the responses to questionnaire.

Moreover, the figures show clearly that the involved respondents were very satisfied on the grounds that DW is useful for decision maker's task (mean=6.62), DW can output information much more quickly (mean=6.33) and if the users content with the benefits of their firm's DW (mean=6.26).Other high responses include DW's information met the requirements of the respondents' task (mean=6.17), DW's information is beneficial to different areas of decisions(mean=6.19), DW is a reliable source of information (mean=6.21), the benefits of implementing DW exceed its cost (mean=6.14) and the DW's information is sufficient for taking sound decisions (mean=6.14). According to these figures, DW is a user-friendly information system (mean=5.43) and The DW offers user-friendly query capability to decision-makers (mean= 5.83) were the least responses.

Based on these figures, it can be concluded that the implementation of DW (in Jordanian firms) is very useful and have a great positive effect on the process of decision-making. These results are similar to some previous studies' results on DW, for example, [13] and [11].

For the purpose of this study, two hypotheses were formulated. The mean, standard deviation (SD), degrees of freedom (DF), standard error (SE), one sample t-test, p-value and one sample Chi-square test were computed. This assumption led to formulating hypothesis H1. Hypothesis H1: There is a strong association between employing DW in Jordanian firms and improving the process of decision-making. From the data, it was apparent that there is strong evidence of a link between implementing the DW in Jordanian firms and soundness and effectiveness of decisions. According to the statistical analysis, the highest p-value is approximately 0.34%, i.e. 0.0034 and since all the p-values of the variable are less than 0.05 and 0.1 level of significance, the null hypothesis (there is no strong association between employing DW in Jordanian firms and improving the process of decision-making) is rejected. As a result the alternative hypothesis (there is a strong association between employing DW in Jordanian firms and improving the process of decision-making) is accepted.

In addition, the same statistical analysis techniques are used to test hypothesis H2. The results showed that all the variables without exception are the major reasons of implementing DW in Jordanian firms. According to these figures, the highest p-value is approximately 0.28%, i.e. 0.0028 and since all the p-values of the variable are less than 0.05 and 0.1 levels of significance, the null hypothesis, i.e. the variables are not the major grounds of implementing DW in Jordanian firms is rejected. As a result the alternative hypothesis (the variables are the major grounds of implementing DW in Jordanian firms) is accepted.

VIII. CONCLUSION

The evidences, which are obtained by analyzing the data from the questionnaires, reveal exceptionally remarkable facts, first of all and to some extent, the percentage of implementing DW is high (35%) by comparison with other countries. one consequence of implementing DW was the great role of DW's information in enhancing and facilitating the process of decision-making. The results showed that the Jordanian firms benefited greatly from implementing DW. The results also revealed that the DW is a fruitful source of information. Moreover, the implementation of DW proved to be a success through helping decision-makers in taking fact-based decisions. Furthermore, the DW was lauded by the

users for the successful use of its information as a basis for decision-making.

This study has humbly contributed to the field of scientific research in general and the field of decision support systems (DSS) in particular in many ways, first of all, the studies on the implementation of DW were nearly all in developed countries. This study was applied in Jordan, which is one of the developing countries in the Middle East. Therefore, the results of this study made a humble contribution to the existing knowledge in the field of implementing DW worldwide in general and in Jordan in particular. Second, there is a need for evidences that the DW improves the quality and accessibility to information. Therefore, this study practically investigated whether or not the implementation of DW improves the quality and accessibility to information and facilitates the decision-makers' tasks. Lastly, based on extensive literature review, the researchers have identified that firms are often unsuccessful due to a lack of appropriate information. For this reason, this study is one of the few empirical studies which have attempted to examine the effect of DW on decision effectiveness. In addition, previous research has not empirically tested its effectiveness in DSS contexts in Jordanian firms "to the researchers' knowledge".

Despite the usefulness and positive contributions of the study's results, these results should be treated and interpreted with caution. In fact, the study's sample included only the Jordanian firms which are listed on ASE. As a consequence, this might severely restrict the generalization of the results. It is believed that the results of this study might have been dissimilar, if all Jordanian firms have been surveyed. Therefore, prospective researchers are recommended to broaden the scope of their investigation to include all Jordanian firms.

REFERENCES

[1] R. L. Hayen, C. D. Rutashobya, and D. E. Vetter, An investigation of the factors affecting data warehousing success, Issues in Information Systems, Vol. 8, No. 2, 2007, pp. 547-53.

[2] J. Ang, and T.S.H. Teo, Management issues in data warehousing: insights from the Housing and Development Board, Decision Support Systems, 29, 2000, pp. 11–20.

[3] B. Shin, An exploratory investigation of system success factors in data warehousing, Journal of the Association for Information Systems, Vol. 4, 2003, pp. 141–170.

[4] Y.-T. Park, An empirical investigation of the effects of data warehousing on decision performance, Information & Management, 43, 2006, pp. 51–61.

[5] D. Mukherjee, and D. D'Souza, Think phased implementation for successful data warehousing, information systems management, Spring 2003, pp. 82-90.

[6] W. Eckerson, Evolution of Data Warehousing: The Trend toward Analytical Applications, Journal of Data Warehousing, Vol. 25, No. 1, 2003, pp. 1-8.

[7] D. Briggs, A Critical Review of Literature on Data Warehouse Systems Success/Failure, Journal f Data Warehousing, Vol. 49, No. 3, 2002, pp. 1 – 20.

[8] G. Shankaranarayanan, and Y. Cai, Supporting data quality management in decision-making, Decision Support Systems 42, 2006, pp. 302–317.

[9] A. Rudra, and E. Yeo, Issues in User Perceptions of Data Quality and Satisfaction in Using a Data Warehouse - An Australian Experience, Proceedings of the 33rd Hawaii International Conference on System Sciences, IEEE 2000, pp. 1-7.

[10] F. Hegazy, and K. Ghorab, The impact of system support on adoption & diffusion of data warehousing success, 2003, http://www.hicbusiness.org/biz2003proceedings, accessed 31/08/2011.

[11] A. Aljanabi, A. Alhamami, and B. Alhadidi, Query Dispatching Tool Supporting Fast Access to Data Warehouse, The International Arab Journal of Information Technology, Vol. 10, No. 3, May 2013, pp. 269-275.

[12] B. H. Wixom, and H. J. Watson, An empirical investigation of the factors affecting data warehousing success, MIS Quarterly, Vol. 25, No. 1, 2001, pp. 17–41.

[13] E. Gimzauskiene, and L. Valanciene, Efficiency of Performance Measurement System: The Perspective of Decision Making, economics and management, 15, 2010, pp. 917-923.

[14] S. A. Mansouri, D. Gallear, and M. H. Askariazad, Decision support for build-to-order supply chain management through multiobjective optimization, International Journal of Production Economics, 135, 2012, pp. 24–36.

[15] J. P. McKenna, Moving Toward Real-Time Data Warehousing, business intelligence Journal, Vol. 16, No. 3, 2011, pp. 14-19.

[16] N. Au, E. W. T. Ngai, and T. C. E. Cheng Extending the Understanding of End User Information Systems Satisfaction Formation: An Equitable Needs Fulfillment Model Approach, MIS Quarterly, Vol. 32, Issue 1, 2008, pp. 43-66.

[17] N. Rasmussen, P. S. Goldy, and P. O. Solli Financial Business Intelligence Trends, Technology, Software Selection, and Implementation, John Wiley and Sons, Inc., New York, 2002.

[18] H. J. Watson, and B. J. Haley, Data warehousing: A framework and survey of current practices, Journal of Data Warehousing, Vol. 2, No. 1, 1997, pp.10-17.

[19] S. Gatziu, and A. Vavouras, Data Warehousing: Concepts and Mechanisms, Informatik, Informatique 1, 1999.

[20] N. F. Doherty, and G. Doig, The role of enhanced information accessibility in realizing the benefits from data warehousing investments, Journal of Organizational Transformation and Social Change, Vol. 8, No, 2, 2011, pp. 163-182.

[21] C.Ballard, D.Herreman, D. Schau, R.Bell, E.Kim, and A. Valencic, Data Modeling Techniques for Data Warehousing, International Business Machines Corporation(IBM Corp), 1st edition,1998.

[22] H. J. Watson, C. Fuller and T. Ariyachandra, Data warehouse governance: best practices at blue cross and blue shield of North Carolina, Decision Support Systems archive, Vol. 38, Issue 3, December 2004, pp. 435 – 450.

[23] I. Ahmad, and S. Azhar, "Data Warehousing in Construction: From Conception to Application," Proceedings of the First International Conference on Construction in the Twenty First Century, Miami, Florida, USA, April 2002.

[24] B.List, R.Bruckner, K. Machaczek, and J. Schiefer, A Comparison of Data Warehouse Development Methodologies - Case Study of the Process Warehouse, DEXA, Munich, 2002.

[25] H. R. Nemati, D. M.Steiger, L. S.Iyer, and R. T. Herschel, Knowledge warehouse: an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing, Decision Support Systems, Vol. 33, Issue 2, June 2002, pp. 143–161.

[26] M. V.Mannino, S. N. Hong, and I. J. Choi, Efficiency evaluation of data warehouse operations, Decision Support Systems, Vol. 44, No. 4, 2008, pp. 883-898.

[27] B.Bçbel, J.Eder, C.Koncilia, T.Morzy, and R. Wrembel, Creation and management of versions in multiversion data warehouse, Proceedings of the 2004 ACM symposium on Applied computing, SAC 2004, March 14-17, Nicosia, Cyprus, pp. 717-723.

[28] T. Brown, Data Warehouse Implementation with the SAS System, SAS Institute Inc., Dallas, TX, 1996, http://www2.sas.com/proceedings/sugi22/DATAWARE/PAPER132.PDF.

[29] K. R Ramamurthy, A. Sen, and A. P. Sinha, An empirical investigation of the key determinants of data warehouse adoption, Decision Support Systems, 44, 2008, pp. 817–841.

[30] S.Nilakanta, K.Scheibe, and A. Rai, Dimensional issues in agricultural data warehouse designs, computers and electronics in agriculture, 60, 2008, pp. 263–278.

[31] J. Chmiel T. Morzy, and R. Wrembel, Multiversion join index for multiversion data warehouse, Information and Software T echnology archive, Vol. 51, Issue 1, January 2009, pp. 98-108.
[32] R. Hackathorn, Current Practices in Active Data Warehousing, Bolder Technology, Inc., 2002.
[33] H. Watson, and B. Haley, Managerial Considerations, In Communications of the ACM, Vol. 41, No. 9, September 1998, pp. 32-37.
[34] H. J.Watson, D.Goodhue, and B.H. Wixom, The benefits of data warehousing: why some organizations realize exceptional payoffs, Information & Management, 2001 (a), pp. 1–12.
[35] H. Watson, T. Ariyachandra, and Jr, R. J. Matyska, Data Warehousing Stages of Growth, Information Systems Management, Vol. 18, Issue 3, June2001 (b),pp. 42–50.
[36] J. D. Wells, and T. J. Hess, Understanding decision-making in data warehousing and related decision support systems: An Explanatory Study of Customer Relationship Management Application, Information Resources Management Journal, Vol. 15, No. 4, October-December 2002, pp. 16-32.
[37] H. A. Abdel Hafez, and S. Kamel, Web Based Data Warehouse in the Egyptian Cabinet Information and Decision Support Center, Decision Support in an Uncertain and Complex World: The IFIP T C8/WG8.3 International Conference, 2004, pp. 402-409.
[38] D. L. Rubin, and T . S. Desser, A Data Warehouse for Integrating Radiologic and Pathologic Data, Journal of the American College of Radiology, Vol. 5, No. 3, March 2008, pp. 210-217.
[39] P. K. Mawilmada, Impact of a data warehouse model for improved decision-making process in healthcare. Masters by Research thesis, Queensland University of T echnology, October 2011.
[40] Á. Ojeda-Castro, M. Ramaswamy, A. Rivera-Collazo, and A. Jumah, Critical Factors For Successful Implementation Of Data Warehouses, Issues in Information Systems, Vol. 12, No. 1, 2011, pp. 88-96.
[41] R. Alshboul, Data Warehouse Explorative Study, Applied Mathematical Sciences, Vol. 6, No. 61, 2012, pp. 3015– 3024.
[42] L. Shen, S. Liu, S. Chen, and X. Wang, The Application Research of OLAP in Police Intelligence Decision System, Procedia Engineering 29, 2012, pp. 397 – 402.
[43] K.Shams, and M. Farishta, Data warehousing: toward knowledge management, Topics in Health Information Management, Vol. 21, No. 3, February 2001, pp. 24-32.
[44] T. Chenoweth, K.Corral, and H. Demirkan, Seven Key Interventions for data warehouse success, Communications of the ACM, Vol. 49, No. 1, January 2006, pp. 115-119.
[45] W.H. DeLone, and E.R. McLean, Information systems success: the quest for the dependent variable, Information Systems Research, Vol. 3, No. 1, 1992, pp. 60–95.
[46] R. L. Kumar, Justifying Data Warehousing Investments, in Data Warehousing and Web Engineering, Shirley Becke (Ed.), 2002, pp. 100-102.
[47] H. J. Watson, J.G.Gerard, L.E. Gonzalez, M.E. Haywood, and D. Fenton, Data warehousing failures: case studies and findings, Journal of Data Warehousing, Vol. 4, No. 1, Spring 1999, pp. 44– 55.
[48] D. Sammon, F. Adam, and F. Carton, Benefit Realisation through ERP: The Re-Emergence of Data Warehousing, Electronic Journal of Information Systems Evaluation, Vol. 6, Issue 2, 2003, pp. 155-16.
[49] M. D. Aguila, and E. Felber, Data Warehouses and Evidence-Based Dental Insurance Benefits, Journal of Evidence Based Dental Practice, Vol. 4, Issue 1, 2004, pp. 113-119.
[50] Devlin, B. Data Warehouse from Architecture to Implementation, Addison Wesley Longman, Inc., 1997.
[51] R. K. Griffin, Data warehousing, Cornell Hotel and Restaurant Administration Quarterly, Vol. 39, No. 4, 1998, pp. 28–40.
[52] N.Wilkoff, T .Pohlmann, R. Hudson, and N. Lambert, The State Of Technology Adoption, Business Technographics North America, May 5 2004, Forrester Research, Inc.
[53] L. Agosta, Hub-and- Spoke Architecture Favored, DM Review, Vol. 15, Issue 3, March 2005, pp. 14-63.
[54] H.-G. Hwang, C.-Y. Ku, D. C.Yen, and C.-C. Cheng, Critical factors influencing the adoption of data warehouse technology: a study of the banking industry in T aiwan, Decision Support Systems, 37, 2004, pp. 1–21.
[55] S. Hong, P. Katerattanakul, S.-K. Hong, and Q. Cao, Usage and perceived impact of data warehouses: a study in Korean financial companies, International Journal of Information Technology & Decision Making, Vol. 5, No. 2, 2006, pp. 297–315.
[56] L.Agosta, M. Andrews, and M. Ritzmann, The Data Warehouse Satisfaction Survey, Part 1: The Number One Complaint About Data Warehousing, Information Management Special Reports, October 2 2007.
[57] K. L. Merritt, User Satisfaction In Data warehousing: An Empirical Investigation Of Salient Variables, Issues in Information Systems, Vol. 9, No. 2, 2008, pp. 500-508.
[58] B. H. Wixom, H. J. Watson, A. M. Reynolds, and J. A. Hoffer, Continental Airlines Continues to Soar with Business Intelligence, Information Systems Management, 25, 2008, pp. 102–112.
[59] A. Almabhouh, A. R.Saleh, and A. Azizah, Examining the Influence of Relationship Quality on Data Warehouse Success, International Journal of Modeling and Optimization, Vol. 1, No. 5, December 2011, pp. 402-409.
[60] T. Ramakrishnan, , M. C. Jones, and A. Sidorova, Factors influencing business intelligence (BI) data collection strategies: An empirical investigation, Decision Support Systems, 52, 2012, pp. 486–496.
[61] M. Saban, and Z. Efeoğlu, An Examination of the Effects of Information Technology on Managerial Accounting in the T urkish Iron and Steel Industry , International Journal of Business and Social Science, Vol. 3, No. 12, Special Issue – June 2012, pp. 105-117.
[62] A. Ebimobowei, and B. Binaebi, Analysis of Factors Influencing Activity-Based Costing Applications in the Hospitality Industry in Yenagoa, Nigeria, Asian Journal of Business Management, Vol. 5, No. 3, 2013, pp. 284-290.
[63] A. S. Hardan, and T. M. Shatnawi, Impact of Applying the ABC on Improving the Financial Performance in T elecom Companies, International Journal of Business and Management, Vol. 8, No. 12, 2013, pp.48-61.
[64] D. George, and P. Mallery, SPSS for Windows Step -by-Step: A Simple Guide and Reference, 14.0 update, 7th Edition 2006, Allyn & Bacon.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1431–1440
DOI: 10.15439/2014F280
ACSIS, Vol. 2

Knowledge Sharing in Distributed Agile Projects: Techniques, Strategies and Challenges

Mohammad Abdur Razzak
School of Software Engineering
Daffodil International University-Bangladesh
abdur.razzak@ieee.org

Rajib Ahmed
Ztorm AB
Stockholm, Sweden
l.rajibahmed@gmail.com

***Abstract*—Knowledge management (KM) is essential for success in global software development. Software organizations are now managing knowledge in innovative ways to increase productivity. In agile software development, collaboration and coordination depend on the communication, which is the key to success. To maintain effective collaboration and coordination in distributed agile projects, practitioners need to adopt different types of knowledge sharing techniques and strategies. There are also few studies that focus on knowledge sharing in distributed agile projects. This research identified the knowledge sharing techniques and strategies applied by the practitioners in distributed agile projects. Challenges faced by the practitioners during knowledge sharing in distributed agile projects are also identified and discussed.**

***Index Terms*—Global software development, knowledge management, knowledge sharing, distributed, agile.**

I. Introduction

Software engineering is a knowledge intensive area. This forces software organizations to manage their knowledge and later use it in smarter, innovative ways to solve problems [46]. It helps software development organizations to acquire and maintain a competitive advantage. KM is crucial for success in global software development [41].

Global software development can be described as "software work which is attempted in different geographical locations across the national boundaries in a coordinated fashion, to involve synchronous and asynchronous interaction" [45]. Software developers work with knowledge and are dependent on each other's work. In global software development this synchronization is dependent on KM. Some studies have identified that knowledge sharing is difficult in distributed agile project due to the lack of face-to-face communication between team members [7,23]. In the agile software development collaboration and coordination depends on communication, which is crucial to successful software development [50]. One of the major objectives of KM is to improve productivity through effective knowledge sharing and transfer [24]. So, the success of agile projects relies on effective knowledge sharing among teams.

This research focuses on exploring knowledge sharing in distributed agile projects. More specifically, this research attempts to identify knowledge sharing techniques, strategies and practices that take place between locally and globally distributed agile teams, and the challenges faced by the practitioners in a distributed agile environment. We are driven by the following research questions:

RQ1: How do team members contribute to knowledge creation in a distributed agile project?

RQ2: How do team members share knowledge in a distributed agile project?

RQ3: What are the challenges faced by the practitioners when sharing knowledge in a distributed agile project?

The rest of the paper is organized as follows. Section II describes a theoretical background of knowledge management in a distributed agile projects is described. In section III we present the research methodology applied and this is followed by validity threats in section IV. A series of semi-structured interviews are described in section V. Results of the different findings are presented in section VI. Discussion of the empirical studies is provided in section VII. Finally, section VIII concludes the paper with a summary of the major findings and future work.

II. Related Work

Software development is considered to be a complex, knowledge intensive and rapidly changing activity, where a number of individuals, teams and organizations are involved infulfilling common goals, interests and responsibilities [13,33]. Technological and strategic knowledge helps developers to communicate; so it is essential to keep the knowledge stored in the organization for the future reuse. Davenport and Prusak [14] define it as "a method that simplifies the process of sharing, distributing, creating, capturing and understanding the company's knowledge". As the size of the organization grows rapidly, it becomes harder to find out where the knowledge resides. Research shows that if the companies manage their knowledge in a better way, they can increase quality, and decrease the time and development costs [44]. To improve the organizational performance, it is important to manage knowledge in a structured way which will help to convey the right knowledge to the right people at the right time. O'Dell and Grayson [37] discussed that, knowledge management is not a vital methodology; it is a framework, a management

978-83-60810-58-3/$25.00

mind-set which is based on past experiences and the creation of new wheels for exchanging knowledge.

To foster dynamic knowledge sharing, improve productivity and coordination in software development teams, agile approaches were introduced. Agile teams share knowledge through several practices [10]: pair programming, release and sprint planning, customer collaboration, cross-functional teams, daily scrum meetings and project retrospectives. But, the authors [10] argue that, these practices are team-oriented and rely on face-to-face interaction between team members. These practices do not facilitate knowledge sharing in distributed agile teams but are effective for collocated and small teams. In one study Dorairaj *et al.* [16] reported that in distributed agile project, team members practice sprint planning, daily scrums, sprint reviews and project retrospective meetings. Distributed agile team members share knowledge through effective use of knowledge management tools like *Wiki*, pair-programming and video-conferencing.

Michael Earl [17] has classified knowledge management (KM) into three categories: technocratic, economic and behavioral. Earl also divided these three categories into seven schools, Technocratic: *Systems, Cartographic and Engineering*, Economic: *Commercial* and Behavioral: *Organizational, Spatial and Strategic*. Both *codification* [20] strategy and *systems school* practice depend on the technology which applies Nonaka's [34] *externalization* conversion technique to convert tacit knowledge into explicit knowledge. Research shows that the technocratic school is closely related with traditional software development and those who are developing software through traditional approaches they are probably benefiting from the technocratic schools [15]. On the other hand, behavioral schools are more related with the agile approaches and agile teams are more benefit more from the behavioral school. A survey in traditional and agile companies shows that agile companies seem to be more satisfied with their knowledge management approaches compared to traditional companies [5]. In agile software development, knowledge sharing happens through the interaction. Developers share knowledge by working together and through close interaction with customers; and more specifically, pair programming, extreme programming, daily scrum meetings, and sprint retrospectives in Scrum. In traditional software development, knowledge management relied primarily on explicit knowledge but in the agile software development KM relies on tacit knowledge [32]. In agile software development, information radiators and collocating teams are related with the spatial school [5].

In traditional software development, knowledge stored explicitly in the documentation, but in the agile development methodology the knowledge is tacit [24]. Extracting tacit knowledge to create explicit knowledge is one of the greatest challenges of knowledge organization [36]. Due to the absence of explicit knowledge in the agile software development, experts need to spend much of their time on repeatedly answering the same questions, knowledge is lost when experienced developers leave project, there is less support for re-usability and there is less contribution to organizational knowledge [24]. In the agile collocated development, informal communication is the key enabler for knowledge sharing but when an agile project is distributed, informal communication and knowledge sharing is a challenge due to low communication bandwidth as well as social and cultural distance [26]. Due to spatial, temporal and cultural factors, communication also becomes aggravated in the distributed settings [21]. Several studies [8,23] also point out that, knowledge sharing in the distributed agile projects is difficult due the challenges in communication, especially face-to-face interaction between team members in different geographical locations. To address these problems, we investigated how shared knowledge creation and transfer activities performed in the distributed agile projects. Along with that, we also investigated what challenges are faced by the practitioners when sharing knowledge among globally distributed agile team members.

III. Research Methodology

Because this research addresses an issue *"How can we retain the benefits that agile practices provide with respect to KM in distributed agile projects"* which is rather under-investigated, this study takes an explorative approach. Exploratory research helps to find out what is happening, seeking new insights and gathering ideas [27,43]. In some qualitative research, data collected through observation or interviews are exploratory in nature. So, extensive interviews are helpful to handle this type of situation [47]. This type of exploratory research was also helpful in achieving our goal through analysis of similarities and differences among the cases [12]. The primary focus of this study was to discover the knowledge sharing activities in distributed agile projects in order to identify techniques, strategies and challenges.

A. Sampling

The selection criteria for these interviewees were based on the kind of company they work at, the experience of the company in distributed agile development (more than 2 years), interviewee role in the distributed team as well as in the company, project duration and project distribution. The participants of this research were project managers, team leaders, software architects, line managers, senior software developers, system developers and Scrum masters in different countries involved in distributed agile projects, located in different countries i.e. Sweden, Norway, Germany, Ukraine, China, India, Bangladesh, USA, and Latvia. To get the rounded perspective of this research phenomenon we included different roles from the agile team.

B. Data Collection

There are three types of interview techniques namely structured, semi-structured and unstructured [18]. Due to the qualitative nature of this study we used semi-structured interviews for conducting a series of interviews in software industries involved in distributed agile projects. According to Robson [42], an in-depth semi-structured interview is helpful in *finding out what is happening and seeking new insights*. *Seventeen*

semi-structured interviews were conducted from *seven* teams in order to identify how practitioners are creating, storing and sharing knowledge related to software development among geographically distributed agile teams. These semi-structured interviews were a combination of both open and focused questions. It helps both interviewer and interviewee to discuss a topic in more details. Before the interviews started, we discussed about overall goal of this research to interviewee. The interview questions were *descriptive* and with the base questions there were follow up questions asked based on the discussion. We were concern about some key terms: *shared knowledge creation, knowledge transfer, strategies and challenges* which later helped us for data analysis and those terms which also evolve with interview questions. We conducted seventeen semi-structured interviews from six different companies. The selected companies are involved with software product development, have different organizational settings and structure and are located in different countries. The duration of these interviews averaged 60 minutes and the interview sessions were tape recorded. Among the seventeen semi-structured interviews, nine were conducted through Skype and eight were face-to-face, depending on distance between interviewer and interviewee.

C. Analysis and Synthesis

In qualitative research, data analysis is the most difficult and crucial aspect due to raw data sets. According to Basit [2], raw data can not help the reader to understand the social world or the participants view unless such data is systematically analyzed. To organize collected data we adopted *thematic analysis* [9] technique during analysis. Thematic analysis is used to identify, analyze and report patterns or themes within data. It minimally organizes and describes data set in detail. In thematic analysis a theme captures data with relation to research questions and represents them in a pattern within the data set [9]. This analysis is performed through a process which maintain six phases to establish meaningful patterns of the data set. Braun and Clarke [9] provides an outline through the six phases of analysis. These phases are: familiarization with data, generating initial codes, searching for themes among code, reviewing themes, defining and naming themes, and producing the final report.

In the *first* stage, we transcribed all the collected interview data into written form in order to conduct a thematic analysis. It helped us to identify possible themes, patterns and to develop potential codes [19]. *Second* phase started with initial codes from the extracted data. There are different types of Coding techniques suggested in different studies such as; *open, axial, selective, descriptive/topic and pattern/analytic* [29,40,48]. In our case, we applied open coding technique and went through all transcribed textual data by highlighting sections of the selected codes. That also helped us to relate coded data with research theme and research questions. In *third* stage, we analyzed broader level of theme rather than codes that helps to sort different codes into potential themes [9].As Braun and Clarke suggested coding as many potential themes/patterns as possible because initially some themes seems to be insignificant, but later they may be important in the analysis process. Later, mind mapping tools were used to represent them into theme-piles. This stage gave us a sense of the significance of individual themes. Stage *four* is reviewing themes. In this stage we identified irrelevant (not enough or diverse) data with relate to different themes and broken down into separate themes. After refining all themes we identified "essence" of each theme and different aspects of the data each theme captures in stage *five*. At the end, in stage *six*, we provided extract data with relate to research questions and present some dialog that connected with different themes in support of results and discussion sections.

IV. Validity Threats

To handle validity threats it is important to identify all possible factors that might affect the accuracy or dependability of the results.

A. Internal Validity

Internal validity for qualitative research mostly relates to the researchers biasness and interpretation of data [6].For finding a similar knowledge level for our interviewees, we went on interviewee profiles on *Linkedin* and their years of experience. After finding out the basic information, the interviewer sent a formal email to the interviewee with an invitation letter about becoming involved with this research. To mitigate the threat of following our own bias, interview questions were designed to have a majority of open ended questions. Every interview started with a similar introduction and some clarification questions. Then the recorded interview was transcribed immediately afterward to reduce the risk of missing some information. Furthermore, researchers sent an interview report to the interviewee in order to check whether interview data was correctly transcribed and to confirm the content indicated participants thoughts, viewpoints, feelings and experiences. In qualitative research it is important to understand the interviewee's inner meaning words. To maintain reliability during data analysis we used a thematic, qualitative data analysis technique, that helped to identify, analyze and report themes within data. The extracted data from the transcribed data was checked twice for any discrepancy by two researchers.

B. External Validity

External validity threat is more applicable to research that are quantitative and which tries to generalize outcome of the research. However, our findings can be generalized only for the agile software development teams which are involved in the development of a shared project from distributed locations.

V. Results

In this section, we describe different findings (techniques, strategies and challenges) from the *seven* cases, that promote effective knowledge creation and sharing activities in distributed agile projects.

TABLE I: Overview of distributed Agile projects

Projects	*Project Distribution*	*Team Size*	*Team Types*	*Agile Position/Roles*	*No. of Interviewees*
Alpha	Sweden-Germany	6-7	Dispersed	Team Leader Developer	2
Beta	Norway-Bangladesh	5-6	Dispersed	Project Manager Developer	2
Gamma	USA-Bangladesh	12-16	Distributed	Head of Engineering Senior Developer Developer	3
Delta	Sweden-Bangladesh	16-18	Dispersed	Software Architect Developer	2
Epsilon	Latvia- Ukraine	11-15	Distributed	Project Manager Developer	2
Zeta	Sweden-China	26-35	Distributed	Line Manager Software Developers System Developer	4
Eta	Sweden-India	45-55	Hybrid	System Developer; Scrum Master	2

TABLE II: Knowledge creation techniques: Locally and Globally

Techniques	α	β	γ	δ	ε	ζ	η
Pair programming	L,G	L,G	L,G	L,G	—	L	L,G
Customer collaboration	L,G	—	L,G	L	L	L,G	L,G
Scrum/Kanban boards	L	—	L,G	L	—	L	L,G
Innovation boards	—	—	—	—	—	—	L,G
Workshops/Seminars	—	—	—	—	—	L	L
Community of practice	—	—	—	L,G	L,G	L,G	L,G
Technical presentation	—	—	L	—	—	L	L
Technical forum	—	—	—	—	—	L,G	L,G

**In Table II, L indicates Locally, G— Globally and "—" not in practice*
Dispersed teams- **α, β, δ***; Distributed teams-* **γ, ε, ζ***; Hybrid team-* **η**

A. Knowledge Creation: Locally and Globally

We have found that distributed agile project teams practice different types of techniques for both local and global shared knowledge creation. *Pair programming*, *customer collaboration*, *Scrum/Kanban* boards and *community of practice* are explicit practices used by the teams to perform both local and global shared knowledge (see in Table II).

1) Pair Programming: Pair programming is used for both local and global knowledge creation. From the series of semi-structured interviews we have found that both local and remote team members work together in one workstation to solve specific problems. They help each other to share their thoughts and create knowledge through discussion. In two cases, we have found that teams do not perform pair programming for shared knowledge creation among remote team members. In the Epsilon(ε) project, all development team members are in one site, and for that reason they do not need to perform pair programming for global shared knowledge creation. However, Zeta(ζ) project is a collaboration with a Chinese team on the same product, but the development team does not have any dependency. The development teams working on different modules and later core developers merge all modules together for specific release. But the local teams in Zeta (ζ) project perform pair programming.

2) Pre-planning game/customer collaboration: In the development cycle the customer has an important role. Customer collaboration helps teams to build up technical-business collaboration on a project and also helps to set the direction of the project. In agile software development customers are always involved with the development teams by providing project requirements and performing acceptance testing. Through customer collaboration agile teams participate in creating local knowledge. Evidence was also found from different cases that customers are also involved with the remote development teams to create knowledge through continuous discussion and features feedback. We have also found that customers are involved in issue tracking systems, which helps both the project manager and the developers towards early iteration. In two cases (δandε), we found that customer collaboration performed only in the local sites for shared knowledge creation.

3) Scrum/Kanban boards: The are two types of boards used by the office to create knowledge and common understanding. A *Scrum board* is used for teams that plan their work in sprint. A *Kanban board* is used to manage and construct team work in progress. In table II it is shown that, teams use *Scrum* and *Kanban* boards for shared knowledge creation among both local and globally distributed team members. In two cases (γandη), we found that teams are using boards both locally and globally. In Gamma (γ) project, the remote team has a sub-Scrum board, which is replica of the main Scrum board. Along with that, the local team (in γproject) upload pictures of the main Scrum board into a repository every day. But in Eta (η) project, teams use a visual Scrum board to perform shared knowledge creation among distributed team members.

4) Innovation boards: Most innovative ideas are kept in the human mind as tacit knowledge. Due to continuous work

loads, sometimes it is impossible to have a discussion with a team member, or other knowledgeable person. So, rather than talking with someone, people share their ideas through the innovation board in an explicit way. In one interview the researchers found that teams are using innovation boards to share their ideas with both collocated and remote team members.

5) Workshop/Seminars: Weekly or monthly workshops and seminars are arranged through collaboration between business teams, technical teams and customers, in order to share knowledge about projects and the latest technologies. This kind of workshop facilitates common understanding and communication between different team members. Workshops also help to facilitate tacit knowledge sharing through *socialization*. In the studied cases, we only observed *large-scale* teams practicing these techniques locally, to create shared knowledge. Later, the theme of the workshops/seminars was shared among remote team members through repositories.

6) Community of practice: To succeed in agile projects, learning is an important asset for agile teams. Agile teams practice two modes of learning: *peer learning* and *community learning*. In *peer learning*, team members start learning through interacting and collaborating with team members. *Community learning* is accessing and conceiving information that is available in knowledge archives or in discussion forums. We found community of practice within different projects, where it performed to share knowledge creation among local and remote team members.

We also found that to create shared knowledge, teams perform *technical presentations*. But these activities are only performed in the local site and later slides or documents are shared among remote team members. Technical forums are also in practice to perform shared knowledge creation between local and remote team members.

B. Knowledge Sharing: Locally and Globally

Knowledge exchange is always challenging in distributed agile teams due to a lack of face-to-face interaction among team members. Practitioners and researchers are trying to mitigate these challenges by initiating different kinds of techniques and tools. From the studied cases we observed that practitioners maintain different types of tools and techniques to share knowledge among globally distributed teams. Based on the findings, these knowledge sharing techniques are listed in Table III.

All studied projects are concerned with using repositories to share knowledge between local and remote team members. Most of the task and product related knowledge is kept in the repositories, which are easy to access by the remote team members. Different teams also depend on daily scrum, weekly sprints status, discussion forums, online conferences and common chat rooms to share knowledge between local and remote team members. Electronic boards are helpful for sharing knowledge across remote teams. Only one case was found where knowledge is shared between both collocated and distributed teams through electronic boards.

1) Repositories: To share knowledge among distributed sites, local teams used different types of repositories like *Wiki, JIRA, Redmine, Confluence and GitHub etc*. These types of repos- itories provide efficient mechanisms to access codified knowledge. From the gathered data it is evident that (see in Table III) practitioners are most dependent on repositories to share knowledge among both local and distributed team members.

Wiki, according to Ulrike Cress [11], provides new opportunities to learn and use collaborative knowledge building and sharing, through social interaction and individual learning. In different cases we found that *wikis* are helpful for starting new threads and discussing issues with other team members. It is also helpful for new team members as it (the *wiki*) provides detailed information about features, documents and so forth.

Project and Issue tracking: Nowadays, almost all medium and large distributed or dispersed agile teams are using *JIRA/Redmine* to track issues, bugs, tasks, deadlines, codes and hours. As collaboration and content sharing tools practitioners used *Confluence* to share docs, files, ideas, specifications, diagrams and mockups. During the interview one project leader said,

...Most of the time we share tacit knowledge between both local and global teams. After that, the information is converted through Redmine to make it explicit... **Project Leader - Alpha project**.

It is also evident that, in one case we found local teams upload *Scrum* board pictures, slides and workshop information in the repositories, which is codified and easy to access by the remote team members.

2) Pair programming: Pair programming plays an important role in creating and sharing developers knowledge in both locally and globally distributed project. In pair programming, two developers work together at one computer with a common goal [38]. In the studied projects, we found teams are using pair programming techniques to share knowledge among remote team members. Team members use Skype to share screens among remote team members. Along with that we also found that teams use *TeamViewer* and *VPN* services to share the same computer screen with remote team members, in order to perform pair programming.

3) Daily Scrum/weekly sprints status/online conferences: Scrum meetings are a source for sharing project progress information among team members. Usually a Scrum standup meeting is held in collocation. From the gathered data we found that distributed teams practice Scrum standup meetings with Internet Relay Chat (IRC), Skype or other group chatting software. Through daily Scrum/weekly sprints status/online conferences local teams share knowledge with distributed team members. In one case (ζ), we found that due to less dependency, the development team does not need to perform Scrum meetings/weekly sprint status. But team (ζ) maintain online conferences in order to share knowledge among remote team

TABLE III: Knowledge sharing techniques among different sites

Techniques	α	β	γ	δ	ε	ζ	η
Repositories	L,G	L,G	L,G	L,G	L,G	L,G	L,G
Pair programming	L,G	L,G	L,G	L,G	—	L	L,G
Version control	—	—	—	—	L,G	—	—
Screen sharing	G	G	G	G	—	—	—
Daily scrum	L,G	—	L,G	—	L,G	L	L,G
Weekly sprint status	L,G	—	L,G	L,G	G	L	L,G
Common chat room	—	—	L,G	L,G	L,G	L,G	L,G
Technical forum	—	—	—	—	—	L,G	L,G
Discussion forum	—	—	L,G	L,G	—	L,G	L,G
Electronic board	—	—	—	—	—	—	L,G
Online conference	G	—	G	G	G	L,G	L,G
Rotation/Visit	—	—	—	—	G	G	G

**In Table III, L indicates Locally, G— Globally and "—" not in practice*
Dispersed teams- **α, β, δ***; Distributed teams-* **γ, ε, ζ***; Hybrid team-* **η**

members (if needed). However, apart from Beta(β) project, other projects explicitly maintain *online conferences* globally for knowledge sharing among distributed team members.

4) Common chat room: Common chat rooms are useful for exchanging knowledge among distributed teams. From the empirical findings we observed that for faster and quicker communication among distributed team members, *medium-* and *large-scale* teams maintain common chat rooms.

In one case, a software architect said that *...the Sprint management system handles all task related knowledge but for the domain related knowledge sharing we maintain a common chat room, which helps us to resolve specific problems within a short time* **- Software Architect, Delta project**

But, in another case we found that, it is not an efficient way to communicate among distributed teams due to language barriers, common understanding, technological factors and so forth. Frequently misunderstandings occur and things go wrong. To mitigate these types of problem, practitioners also suggested different types of mitigation techniques.

5) Technical forum: The idea behind a technical forum is *learning through sharing knowledge*. Technical forums are like communities of practice which create a network between technical team members. They are self-organizing groups that consist of individuals who share information, experience and technical skill on a specialized discipline [28]. Technical forums assist distributed teams in quick problem solving and reduce development time since team members do not get stuck on recurring issues. Building trust between team members in the distributed environment is challenging; so knowledge sharing through technical forums can build trust between developers. Technical forums help to create and share both local and distributed knowledge. We have found that large-scale team members practice *technical forum* techniques to share knowledge among remote team members.

6) Electronic board: The office boards hold a lot of knowledge which is difficult to share among distributed teams. The interviews revealed that practitioners are using electronic boards to share and access knowledge both locally and globally. Electronic boards hold the tasks list to perform , latest information and along with that necessary technical and business information are regularly updated in a wiki. Electronic boards help to decrease the communication overhead.

In one case an interviewee said *...I am not satisfied with the current tools; Its tough to describe designs to new team members. Visual aids are helpful during discussions* **- Project Manager, Beta project.**

7) Rotation/Visits: The primary intention of team member rotation between different sites is knowledge sharing. Due to frequent face-to-face interaction with product owners, on-site team members get more business and domain related information than offshore team members [49]. A lack of face-to-face meetings and poor socialization also causes a lack of trust among distributed team members [30]. Rotation between on-site and distributed team members promotes the sharing of business and domain related knowledge across the teams. From the data gathered we found that, both *distributed* and *hybrid* teams visit remote sites and rotate team members to increase the trust and communication bandwidth between team members. But the studied *dispersed* teams never visit and rotate with remote team members.

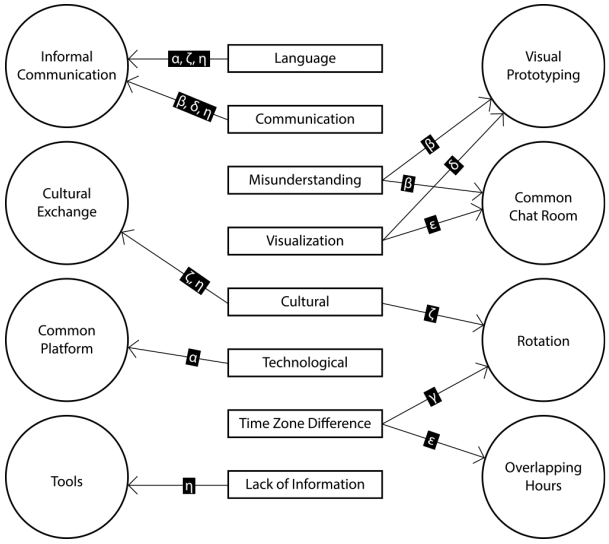
One of the distributed teams line managers said *Visits to remote sites are highly costly. So, we rotate team members and mostly, the duration of the rotation between team members is 3-6 months.* **- Line Manager, Zeta project**

We have also found that teams practice *version control, screen sharing and discussion forums* to maintain knowledge sharing among both local and remote team members.

C. Challenges faced by practitioners during knowledge sharing among distributed teams

In agile software development most of the knowledge is tacit, which resides in the human mind rather than documentation. This codified tacit knowledge is shared among between locally and globally distributed team members through tools. The knowledge sharing approach varies between team members due to experience levels. The types of problem this leads to are search availability and difficulty finding the right knowledge at the right time. We have also found that to share tacit knowledge between remote team members, teams maintain a *common chat room* and *online conference* (see in Table III). In one project

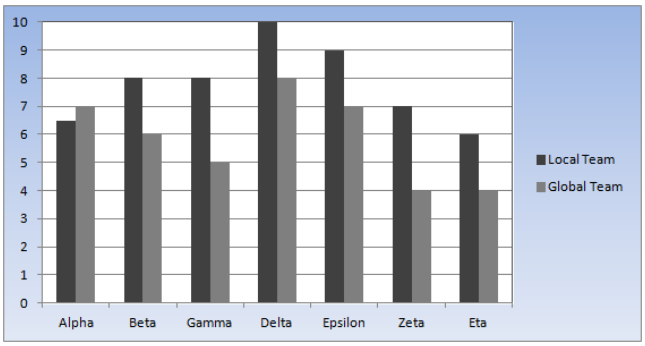
(β), we found that the team does not share tacit knowledge among dispersed team members. But, based on the situation, sometimes the team performs pair programming through screen sharing, to resolve problems. Challenges faced by the practitioners during knowledge sharing among distributed team members are shown in Figure 1. Mitigation techniques applied by practitioners are also shown in the same Figure 1.



Dispersed teams- **α, β, δ***; Distributed teams-* **γ, ε, ζ***; Hybrid team-* **η**

Fig. 1: Knowledge sharing challenges and mitigation techniques

In Figure 1, arrows indicate the mitigation techniques applied by practitioners for a specific challenge. Based on the severity of *communication*, *language* and *cultural* challenges frequently faced by practitioners during knowledge sharing in distributed agile projects (see Figure 1). Distanced teams are also struggling with *misunderstanding* and *visualization* challenges.



Not satisfied (0-3), Satisfied (4-6) and Highly satisfied (7-10)
Dispersed teams- **α, β, δ***; Distributed teams-* **γ, ε, ζ***; Hybrid team-* **η**

Fig. 2: Success of knowledge sharing

Though teams face different types of challenges during knowledge sharing among distributed team members, we identified successful knowledge sharing in both locally and globally distributed agile teams from the seven cases studied. Based on the seven cases the above graph (see Figure 2) has been drawn. An *ordinal scale* is used to map the interviewee satisfaction with their KM activities in both local and global teams. Figure 2 depicts that in project (*Alpha*) interviewees think knowledge sharing activities are more successful among distributed team members than in the local team, due to language barriers (*non-native English speaker*) between local team members. It is also evident from the gathered data that *dispersed* feature teams are more successful in their knowledge sharing among distributed team members than *distributed* and *hybrid* teams.

VI. Discussion

A. Codification of Knowledge

According to Polanyi [39], *Individuals know more than they can say*. Polanyi classified human knowledge into two categories. *Tacit knowledge*, which is very difficult to describe or express: this type of knowledge is transferred through demonstration. Tacit knowledge has an important cognitive dimension which consists of mental models, beliefs and perspectives [3,34,35]. So it cannot be easily characterized by clear expressive language. *Explicit knowledge*, is easily written down and codified. It is easily possible to characterize explicit knowledge in textual or symbolic forms. This kind of knowledge resides in textbooks, memos and technical documents. Codification of knowledge is the conversion of tacit knowledge into explicit knowledge in a written, verbal or visual format. The extraction process of tacit knowledge into explicit is called *externalization*. Tacit knowledge cannot be interpreted fully even by an expert [1]. This type of knowledge is more deeply placed in action and is hard to express in words [22]. Nelson *et al.* [31] conclude that it is impossible to describe all the necessary aspects of organizational tacit knowledge for successful performance. In organizations, most of the tacit knowledge is work related, which is learned informally as the team works [51]. Codification extract tacit knowledge into explicit ; it is a challenging task, so an expert needs to understand the essence of the tacit knowledge in order to increase the degree of explicitness of knowledge. Surprisingly, we found from our results that all studied cases are concerned about knowledge codification. To codify tacit knowledge, teams are using Wiki, JIRA, Confluence etc. In local sites, technical presentations and discussion forums are also taken into account as knowledge codification strategies. Later, teams share codified knowledge among remote team members through repositories and that is helpful for the remote team members to reuse codified stored knowledge.

B. Knowledge Management Strategies in Practices

We found that knowledge management schools are in practice, according to the results in Table II and III. We used Earls [17] framework to select types of strategies,r schools or practices in the different projects that applied to managing

TABLE IV: Knowledge sharing strategies in practice

Projects	**Locally — Globally**									
	Systems	*Cartographic*	*Engineering*	*Organizational*	*Spatial*	*Systems*	*Cartographic*	*Engineering*	*Organizational*	*Spatial*
Alpha	+	-	+	+	+	+	-	+	+	-
Beta	+	-	+	+	-	+	-	+	-	-
Gamma	+	-	+	+	+	+	-	+	+	-
Delta	+	+	+	+	+	+	+	+	+	-
Epsilon	+	-	+	+	-	+	-	+	+	-
Zeta	+	+	+	+	+	+	+	+	+	-
Eta	+	+	+	+	+	+	+	+	+	+

[+] In practice, [-] Not in practice

knowledge locally and globally.

Based on the evidence from the different cases, we found that knowledge management schools were in use to manage knowledge both locally and globally. It is also evident that *systems*, *cartographic*, *engineering*, *organizational* and *spatial* schools are practiced in distributed agile projects to manage knowledge both locally and globally (see Table IV). Both *commercial* and *strategic* schools are focused on a business perspective (*patent*, *copyright*, *trademark*, *know-how* and *intellectual assets* [17]) and there is also no evidence found within gathered data sets that indicates those schools (*commercial* and *strategic*) are in practice. For that reason those schools are not taken into account in this research.

1) Systems school: This schools philosophy is to codifying knowledge with the help of technology. Organizations use repositories for storing and sharing knowledge. These knowledge repositories usually store domain specific information. The codification of knowledge can be compared with the *externalization* of knowledge by Nonaka [15]. It is easy to realize the benefits of knowledge bases and the systems school is the most researched school [4]. These knowledge bases become richer and more useful over time. As shown in Table IV, the *systems* school is in practice in all cases to manage knowledge locally and globally. Though search functions are a difficult issue in the systems school, practitioners depend on it because across distances this school effectively perform knowledge sharing activities using repositories.

2) Cartographic school: This school focuses on the mapping of organizational knowledge and aims to build knowledge directories by disclosing who knows what [17]. This is sometimes achieved by yellow-pages, which ensure the accessibility to others of a knowledgeable person within the organization for knowledge exchange. Though knowledge maps and directories on company intranets might be helpful for distributed team members to have an idea of who knows what, in distributed projects it seems challenging to put into practice. This is because it needs joint effort and commitment from both local and remote team members. In collocation, it seems easier to find a knowledgeable or experienced person because they knew each other well. In globally distributed projects who knows what and what is where are important issues for effective knowledge exchange. We have found that the *cartographic* school is practiced by different projects (δ,ζ,η) to exchange knowledge both locally and globally. This strategy is also practiced in different companies by introducing the idea of knowledge brokers: this helps other developers to consult with knowledgeable and experienced software engineers [46]. Knowledge brokers are knowledgeable and experienced software engineers who will communicate with other developers, provide them with information or listen to them.

3) Engineering school: This school of knowledge management focuses on business process re-engineering [4] and knowledge flows in organizations. This school has a more empirical attention than other schools, which focuses on managing knowledge about software development processes and improvement of software development processes. More specifically, this school focuses on formal routines, mapping of knowledge flows, project reviews, and social interactions. Software process improvements like CMMI can be regarded as a stimulus for knowledge flow throughout the organization. This school supports explicit knowledge sharing and in distributed projects, temporal distance does not affect this school. In globally distributed projects coordination is one the major challenges and the engineering school focuses on the coordination process and aims to ensure knowledge flows within the organization though shared databases. The processes of using tools *(i.e the installation manual for GitHub or SVN with eclipse)*, quality code writing techniques, testing and reviews are all documented in repositories to share among distributed teams. Practice of this school is found in the studied projects.

4) Organizational school: The philosophy of the *organizational* school is to create a network by collaborating between communities to share or pool knowledge. This school of knowledge management focuses on organizational structure. These structures are often referred as knowledge communities [4]. This is a networking approach for people to communicate and share knowledge. Based on the seven cases, this school is in practice for knowledge sharing both locally and globally.

5) Spatial school: The intention of the spatial school is to encourage socialization (tacit to tacit knowledge) as a means of knowledge exchange [25]. The spatial school is more

concerned with the development and utilization of the social capital which develops from people interactions, formal or informal, repeatedly over time [17]. The spatial school focuses on designing office space to promote knowledge sharing [15]. Organizations use different office settings to promote communication between people. For example, in the case of software organizations agile methodologies may use boards, charts or other tools to create spatial knowledge. Sometimes, even common spaces like conference rooms, dining rooms or places for refreshment and activities are also places where knowledge can be shared. Five cases were found that practice the *spatial* school to manage knowledge locally and only one case (*hybrid team*,η) was found that to practice the *spatial* school to manage knowledge both locally and globally. The *hybrid team* uses visual boards to communicate among collocated and distributed team members.

VII. Conclusion and Future Work

The aim of this research was to discover the knowledge sharing techniques, strategies applied and challenges faced by the practitioners in distributed agile projects.

To perform knowledge management activities in a distributed agile project, different teams practice different types of approaches. But, in general, we found that different types of knowledge creation and sharing techniques are applied by practitioners to perform knowledge management activities in distributed agile projects. Along with that, we also found different types of strategies practiced by the team members to manage knowledge both locally and globally.

For the first research question, we found that:

- To perform shared knowledge creation in a distributed agile project, team members practice: pair programming, customer collaboration, Scrum/Kanban boards, innovation boards, workshops/seminars, learning, technical presentation and technical discussion techniques.
- In globally distributed agile projects, teams practice different types of strategies to perform shared knowledge creation such as: *systems, engineering, organizational and cartographic* schools. We observed that the spatial school is in practice for local knowledge creation but when the project is distributed, this school is used less due to expensive tools.

For the second research question, we found that:

- To share knowledge among distributed sites, team members practice different type of techniques: repositories, pair programming, version control, screen sharing, daily scrums, weekly sprint status, common chat rooms, technical forums, discussion forums, electronic boards, online conferences, rotations/visits etc.
- *Systems*, *engineering* and *organizational* school strategies are explicitly in practice to share knowledge among distributed team members. These strategies foster effective knowledge sharing activities for team members in distributed agile projects. In distributed development, who knows what and what is where need to be known by employees, for effective knowledge sharing: this is associated with the cartographic school. But, in distributed agile projects, this school has is used less due to social-cultural distances. The spatial school facilitates knowledge sharing by using office space but in distributed agile projects this strategy is not explicitly in practice to share knowledge among remote team members.

For the third research question, we found that:

- During knowledge sharing among distributed team members, practitioners faced different types of challenges, such as: language, communication, misunderstanding, visualization, cultural, technological, time zone difference and lack of information.
- To mitigate those challenges, practitioners also apply different types of mitigation techniques, such as: informal communication, cultural exchange, common platform, tools, visual prototyping, common chat rooms, rotation, and overlapping hours.

Through a series of semi-structured interviews from agile practitioners, we investigated knowledge sharing activities in distributed agile projects. Communication, coordination and collaboration are the keys to fostering knowledge sharing between team members in agile software development. However, we have seen knowledge sharing in distributed agile projects is challenging, due to factors such as communication difficulties, language barriers and cultural barriers. To mitigate those challenges and succeed in knowledge sharing within and across borders, practitioners adopt different types of techniques to manage knowledge both locally and globally. Along with these techniques, we have also noticed that, practitioners adopt different types of strategies to manage knowledge both locally and globally. Though systems, engineering and organizational schools are explicitly in practice, the spatial school has less concern with managing knowledge in distributed agile projects. With closer observation between software engineering and schools Bjørnson and Dingsøyr found that there is a heavy focus on the systems and engineering schools [4]. There are also limited number of studies focusing on the organizational school, but no studies in software engineering were found, that focus on the spatial aspect [4]. Agile software development is more related to *socialization*, which includes the spatial schools concepts of knowledge sharing strategies. There is a lot of knowledge residing in the office space and office space fosters knowledge sharing through spatial knowledge management strategy. In the future, it will be interesting to find the spatial school being practiced in distributed agile projects.

VIII. Acknowledgement

We are thankful to our dear supervisor Dr. Darja Šmite, Associate Professor and this paper also published as a Blekinge Institute of Technology's Master thesis, 2013.

References

[1] P.K.K. Ahmed, K.K.K. Lim, and A.Y. Loh. *Learning through knowledge management*. Routledge, 2012.

[2] T. Basit. Manual or electronic? the role of coding in qualitative data analysis. *Educational Research*, 45(2):143–154, 2003.

[3] A. Bennet and M.S. Tomblin. A learning network framework for modern organizations: Organizational learning, knowledge management and ict support. *VINE*, 36(3):289–303, 2006. *DOI:10.1108/03055720610703588*.

[4] F.O. Bjørnson and T. Dingsøyr. Knowledge management in software engineering: A systematic review of studied concepts, findings and research methods used. *Information and Software Technology*, 50(11):1055–1068, 2008, *DOI: 10.1016/j.infsof.2008.03.006*.

[5] F.O. Bjørnson and T. Dingsøyr. A survey of perceptions on knowledge management schools in agile and traditional software development environments. *Agile Processes in Software Engineering and Extreme Programming*, pages 94–103, 2009, *DOI:10.1007/978-3-642-01853-4_12*.

[6] I. Bleijenbergh, H. Korzilius, and P. Verschuren. Methodological criteria for the internal validity and utility of practice oriented research. *Quality & Quantity*, 45(1):145–156, 2011, *DOI:10.1007/s11135-010-9361-5*.

[7] A. Boden and G. Avram. Bridging knowledge distribution-the role of knowledge brokers in distributed software development teams. In *Cooperative and Human Aspects on Software Engineering, 2009. CHASE'09. ICSE Workshop on*, pages 8–11. IEEE, 2009, *DOI:10.1109/CHASE.2009.5071402*.

[8] A. Boden, G. Avram, L. Bannon, and V. Wulf. Knowledge management in distributed software development teams-does culture matter? In *Global Software Engineering, 2009. ICGSE 2009. Fourth IEEE International Conference on*, pages 18–27. IEEE, 2009, *DOI:10.1109/ICGSE.2009.10*.

[9] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006, *DOI:10.1191/1478088706qp063oa*.

[10] T. Chau and F. Maurer. Knowledge sharing in agile software teams. *Logic versus approximation*, pages 173–183, 2004, *DOI:10.1007/978-3-540-25967-1_12*.

[11] U. Cress and J. Kimmerle. A systemic and cognitive view on collaborative knowledge building with wikis. *International Journal of Computer-Supported Collaborative Learning*, 3(2):105–122, 2008, *DOI:10.1007/s11412-007-9035-z*.

[12] J.W. Creswell. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage Publications, Incorporated, 2008.

[13] B. Curtis, H. Krasner, and N. Iscoe. A field study of the software design process for large systems. *Communications of the ACM*, 31(11):1268–1287, 1988, *DOI:10.1145/50087.50089*.

[14] T.H. Davenport and L. Prusak. *Working knowledge: How organizations manage what they know*. Harvard Business Press, 2000, *DOI:10.1023/A:1011199223173*.

[15] T. Dingsøyr, F.O. Bjørnson, and F. Shull. What do we know about knowledge management? practical implications for software engineering. *Software, IEEE*, 26(3):100–103, 2009, *DOI:10.1109/MS.2009.82*.

[16] S. Dorairaj, J. Noble, and P. Malik. Knowledge management in distributed agile software development. In *Agile Conference (AGILE), 2012*, pages 64–73. IEEE, 2012, *DOI:10.1109/Agile.2012.17*.

[17] M. Earl. Knowledge management strategies: Toward a taxonomy. *Journal of management information systems*, 18(1):215–233, 2001.

[18] U. Flick. *An introduction to qualitative research*. Sage Publications Limited, 2009.

[19] G. Guest, K.M. MacQueen, and E.E. Namey. *Applied thematic analysis*. Sage Publications, Incorporated, 2011.

[20] M T Hansen, N Nohria, and T Tierney. Whats your strategy for managing knowledge? *Knowledge Management Critical Perspectives on Business and Management*, 77(2):322, 2005.

[21] T. Hildenbrand, M. Geisser, T. Kude, D. Bruch, and T. Acker. Agile methodologies for distributed collaborative development of enterprise applications. In *Complex, Intelligent and Software Intensive Systems, 2008. CISIS 2008. International Conference on*, pages 540–545. IEEE, 2008, *DOI:10.1109/CISIS.2008.105*.

[22] D. Hislop. Mission impossible? communicating and sharing knowledge via information technology. *Journal of Information Technology*, 17(3):165–177, 2002, *DOI:10.1080/02683960210161230*.

[23] H. Holz and F. Maurer. Knowledge management support for distributed agile software processes. *Advances in Learning Software Organizations*, pages 60–80, 2003, *DOI:10.1007/978-3-540-40052-3_7*.

[24] RK Kavitha and I. Ahmed. A knowledge management framework for agile software development teams. In *Process Automation, Control and Computing (PACC), 2011 International Conference on*, pages 1–5. IEEE, 2011, *DOI:10.1109/PACC.2011.5978877*.

[25] M.B. Lloria. A review of the main approaches to knowledge management. *Knowledge Management Research & Practice*, 6(1):77–89, 2008, *DOI:10.1108/JKM-10-2012-0316*.

[26] W. Maalej and H.J. Happel. A lightweight approach for knowledge sharing in distributed software teams. *Practical Aspects of Knowledge Management*, pages 14–25, 2008, *DOI:10.1007/978-3-540-89447-6_4*.

[27] G.R. Marczyk, D. DeMatteo, and D. Festinger. *Essentials of research design and methodology*, volume 2. Wiley, 2010.

[28] R. McDermott. Learning across teams. *Knowledge Management Review*, 8(3):32–36, 1999.

[29] M.B. Miles and A.M. Huberman. *Qualitative data analysis: An expanded sourcebook*. Sage Publications, Incorporated, 1994.

[30] N.B. Moe and D. Šmite. Understanding a lack of trust in global software teams: a multiple-case study. *Software Process: Improvement and Practice*, 13(3):217–231, 2008, *DOI:10.1007/978-3-540-73460-4_6*.

[31] R.R. Nelson and S.G. Winter. *An evolutionary theory of economic change*. Belknap press, 1982.

[32] S. Nerur, R.K. Mahapatra, and G. Mangalaraj. Challenges of migrating to agile methodologies. *Communications of the ACM*, 48(5):72–78, 2005, *DOI:10.1145/1060710.1060712*.

[33] B. Nicholson and S. Sahay. Embedded knowledge and offshore software development. *Information and organization*, 14(4):329–365, 2004, *DOI:10.1016/j.infoandorg.2004.05.001*.

[34] I. Nonaka. A dynamic theory of organizational knowledge creation. *Organization science*, 5(1):14–37, 1994, *DOI:dx.doi.org/10.1287/orsc.5.1.14*.

[35] Ikujiro Nonaka. The knowledge-creating company. *Harvard Business Review*, 26(4-5):598–600, 2007.

[36] Ikujiro Nonaka and Noboru Konno. The concept of ba : Building a foundation for knowledge creation. *California Management Review*, 40(3):40–54, 1998.

[37] C. O'Dell and C.J. Grayson. If only we knew what we know: identification and transfer of internal knowledge and best practices. *California management review*, 40:154–174, 1998.

[38] D.W. Palmieri. Knowledge management through pair programming. 2002.

[39] M. Polanyi. *The tacit dimension*. University of Chicago press, 2009.

[40] K.F. Punch. *Introduction to research methods in education*. Sage Publications Limited, 2009.

[41] I. Richardson, M. O'Riordan, V. Casey, B. Meehan, and I. Mistrik. Knowledge management in the global software engineering environment. In *Global Software Engineering, 2009. ICGSE 2009. Fourth IEEE International Conference on*, pages 367–369. IEEE, 2009, *DOI=10.1109/ICGSE.2009.57*.

[42] C. Robson. *Real world research: a resource for social scientists and practitioner-researchers*, volume 2. Blackwell Oxford, 2002.

[43] P. Runeson and M. Höst. Guidelines for conducting and reporting case study research in software engineering. *Empirical Software Engineering*, 14(2):131–164, 2009, *DOI:10.1007/s10664-008-9102-8*.

[44] I. Rus, M. Lindvall, and S. Sinha. Knowledge management in software engineering. *IEEE software*, 19(3):26–38, 2002, *DOI:10.1109/MS.2002.1003450*.

[45] S. Sahay, B. Nicholson, and S. Krishna. *Global IT outsourcing: software development across borders*. Cambridge University Press, 2003.

[46] K. Schneider. *Experience and knowledge management in software engineering*. Springer, 2009, *DOI:10.1007/978-3-540-95880-2*.

[47] U. Sekaran. *Research methods for business: A skill building approach*. John Wiley & Sons, 2006.

[48] F. Shull, J. Singer, and D.I.K. Sjøberg. *Guide to advanced empirical software engineering*. Springer, 2007.

[49] K. Sureshchandra and J. Shrinivasavadhani. Adopting agile in distributed development. In *Global Software Engineering, 2008. ICGSE 2008. IEEE International Conference on*, pages 217–221. IEEE, 2008, *DOI:10.1109/ICGSE.2008.25*.

[50] D. Šmite, N.B. Moe, and P. Ågerfalk. Agility across time and space: Implementing agile methods in global software projects. 2010.

[51] R.K. Wagner and R.J. Sternberg. Tacit knowledge in managerial success. *Journal of Business and Psychology*, 1(4):301–312, 1987, *DOI:10.1007/BF01018140*.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1441–1447
DOI: 10.15439/2014F44
ACSIS, Vol. 2

Information security in IT global sourcing models

Prof. Kazimierz Perechuda
Wroclaw University of Economics
ul. Komandorska 118/120
53-345 Wroclaw, Poland
Email:
kazimierz.perechuda@ue.wroc.pl

Dr Małgorzata Sobińska
Wroclaw University of Economics
ul. Komandorska 118/120
53-345 Wroclaw, Poland
Email:
malgorzata.sobinska@ue.wroc.pl

***Abstract*—In the dynamic environment, organizations are required to build their competitive advantage not only on own resources, but also on resources commissioned from external providers, accessed through various forms of sourcing, including the sourcing of IT services. This paper presents some of the aspects of information security, in the context of the modern implementation of new IT sourcing methods. IT sourcing solutions are presented, as employed by modern companies, together with potential benefits offered. The main focus is put on the determination of the most important risks involved in information sharing in IT sourcing relations, as well as minimization and reduction of such risks, with particular attention to various cloud computing services on offer.**

***Index Terms*—management, IT sourcing models, cloud computing, information security**

I. Introduction

BUSINESSES are on the lookout for newer and more innovative ways to enhance competitiveness and get ahead of the growth curve. A new generation of advanced technologies – social, mobility, analytics and cloud – have taken the center-stage, promising to transform enterprises and help them do business better. Enterprises that embrace these technologies would be able to seamlessly redesign their business models, strategy, operations and processes to meet the new customer demands.

The business models employed by modern enterprises are increasingly more involved in problems related to the security and protection of information, data, and knowledge, particularly of the classified and undisclosed type.

In a sense, these business models can be viewed as based on knowledge and security. The classified knowledge (technical, technological, design, logistic, etc.) is one of the core competences of large network corporations, such as Renault, Mazda, Opel, Toyota, Deutsche Bank, and others.

The network structure of large corporations, while designed to provide competitive advantage in two areas, namely:

- outsourcing of ancillary functions, support functions, and even primary functions (as in the case of Opel assembly factory in Gliwice),
- centralized investment in R+D and new technologies (patents, inventions, improvements, copyrights),

may also increase the risk of uncontrolled 'leakage' of key undisclosed knowledge (technical, design, technological, financial, trade, etc.) to market competitors. This is a direct result of the increased access to core corporate knowledge offered to cooperating entities.

The most important aspect of this process is the natural outflow of hot knowledge, resulting from transmigration of knowledge agents (managers and long-term employees with unique competences and experience), in both the networked and non-networked systems.

II. New Global Sourcing Models Of Business

In the modern, 'flat' model of economy, networked enterprises build their competitive advantage through careful selection of sourcing agents. One of the most important criteria for such selection is the perceived level of security with respect to uncontrolled and undesired outflow of data, information, and knowledge from organizations to other entities outside their network structure.

Sadly, this particular criterion is rarely perceived as mission-critical. Companies tend to prioritize the aspect of compatibility between core competences of the potential sourcing partner with key competences and resources of the mother company. The increased asymmetry of key competences between sourcing partners may result in the following trends (Fig. 1):

- departure (short-term contracts, rapid capturing of the partner's know-how),
- unification (strengthening the cooperation, balancing the symmetry of undisclosed knowledge, participation in future projects).

New needs of enterprises result in the emergence of new types of global sourcing models, where sourcing can be defined "as the act through which work is contracted or delegated to an external or internal entity that could be physically located anywhere" ([1], p. 2]. Sourcing can also be defined as a comprehensive organizational strategy for distribution of business processes and other functional areas of the enterprise among cooperating partners. For the purpose of this study, sourcing is defined as a notion of superior level to the notions of outsourcing and insourcing ([2], p.17). The main

978-83-60810-58-3/$25.00

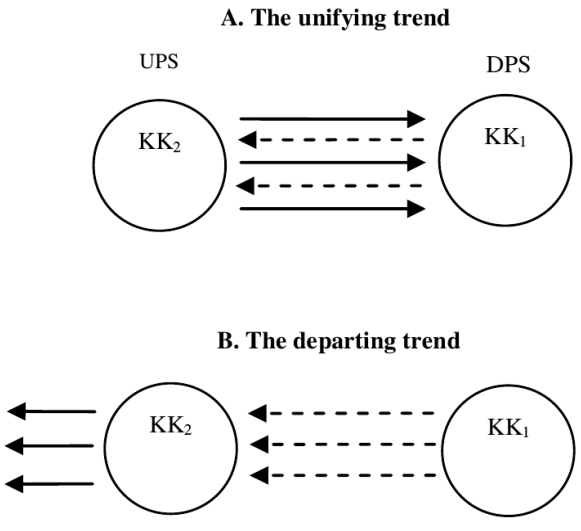


Fig 1. Trends in sourcing cooperation
Source: own research.

differences between sourcing models involve determination of the following factors:

- is the sourced function delegated to a dependent entity or an independent external supplier (or both), and
- is the sourced function performed on-site or off-site, is it performed onshore, nearshore (in a neighboring country) or offshore (in a remote location) ([1], p. 25).

A business model of IT sourcing may comprise the following types/models of sourcing cooperation/relation (own research, based on: [7], pp. 6-16; [1], pp. 26-42; [5], p. 1300; [6]]: facility management, selective outsourcing, tactical outsourcing, transformational outsourcing, transitional outsourcing, Business Process Outsourcing, joint ventures, benefit-based relationships, insourcing (staff augmentation), offshore outsourcing (foreign supplier), nearshore outsourcing (foreign supplier), onshore outsourcing (domestic supplier, "rural sourcing"), cosourcing, shared services, captive models and models based on Internet: cloud computing, software as a service, crowdsourcing and microsourcing. Figure 2 presents a graphic representation of a general model of IT sourcing.

A large number of modern organizations operate simultaneously in two business areas: the real and the virtual. The greater the availability of resources, the greater the potential impact. The more we expand the range of the network (new locations where functions/processes/services of an organization are implemented; greater number of sourcing providers; new areas of service delivery such as cloud computing), the greater the potential opportunities, but also the greater is the risk involved.

A decision to implement a particular sourcing model may be influences by the following factors:

Factors supporting the trends towards IT sourcing (own research, based [7], p.4):

- the global skills shortage,
- a more mobile workforce,
- the mounting cost of in-house developed software,
- the need to move fast, rapidly adopting new technologies and speeding up system development,
- the explosion of Internet technologies and services requiring a wide range of new skills and investments.

For the purpose of this study, the authors focus on the presentation and analysis of one of the most popular Web-based sourcing models – the cloud computing – without going into detailed analysis of other IT sourcing models in use.

Cloud computing allows users to access technology-enabled services on the Web, without having to know or understand the technology infrastructure that supports them. Nor do they have much control over it. It is an innovative new way to boost capacity and add capabilities in computing without spending money on new infrastructure, training new personnel or licensing software.

There are four basic types of clouds: private clouds (operated solely for the use of a single organization), community clouds (operated for a specific group that shares infrastructure), public clouds (which use cloud infrastructure available over a public network) and hybrid clouds (which combine the infrastructure of two or more clouds - public, community, and private).

The increased risk of cloud computing projects has to do with opening up the organization to a whole new space, which is not yet completely examined and "protected". The range of potential risk scenarios is impossible to predict at this moment, since they have yet to be observed in organizational practice. At the same time, the output and the use of the "new space-clouds" can increase the potential added value of using this type of sourcing, compared to more classical forms, such as the generic outsourcing and offshoring models.

New forms of contracting, and – consequently – new resource acquisition methods are required to help modern organizations survive in this age of innovation and strong competition. However, it should be noted that those new solutions, as any new ideas, come with new risks. Some of those risks will be discussed in the following chapters.

III. Information and knowledge security in the IT sourcing models

Information security is one of the key factors to be taken into account in the context of sourcing decisions, particularly those which involve cooperation with external partners. Potential contractors may operate from remote locations, often

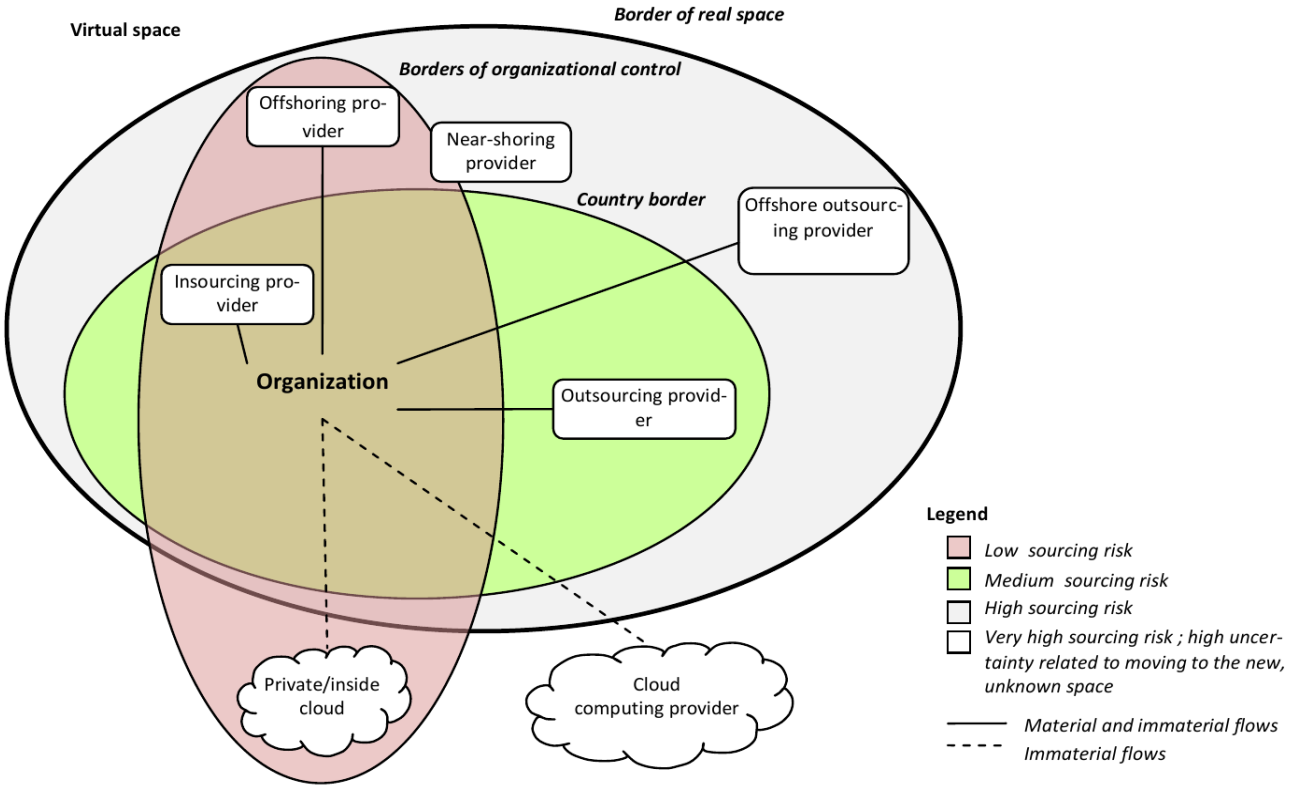


Fig. 2 A general model of IT sourcing
Source: Sobińska 2014

in diverse cultural, political, social, legal, and other settings. The problem of information security is also reported as one of the main concerns for potential cloud computing users and clients.

According to K. Liderman, information security comprises all forms (also verbal) associated with the exchange, storage, and processing of information. It represents the risks involved in information resource loss, as well as misinformation resulting from poor quality of information provided. It must be noted that Liderman makes a clear distinction between information security and 'security of information'. The latter, a subset of information security, is defined as "*justified trust (e.g. based on risk analysis and the adopted risk management procedures) that the organization will not face potential losses as a result of undesired (accidental or intentional) use of information stored, processed, and transmitted within the system, be it information disclosure, modification, removal, or rendering it unfit for processing purposes"* ([4], p. 22). The problem of information security, despite increased publicity in professional literature and media, is still trivialized by managers of many companies. And, while the managerial personnel declare their knowledge of risks involved in this area (as reported in many studies, see: ([11], p.19), in practice they tend to minimize their efforts with respect to information security. The main reasons for negligence are: the financial outlay, and the lack of competences. For instance, not many companies are able to perceive the risk associated with the fact that company information is no longer contained within the 'company perimeter'. Nowadays, information flows freely, and the risk of disclosure is far more pronounced. The lack of informed approach to the risk management process is a fast lane to disaster, since potential incidents may gravely affect the security or company development. *Modern organizations* face the serious challenge of implementing effective security strategies, with proper risk management as one of the main elements of the system [5].

In this context, outsourcing may be seen as a chance to improve information security (and the security of IT systems) by transferring the IT security responsibilities to an external provider with proper specialist knowledge and IT technologies. On the other hand, it may just as well add to the risk, by opening up sensitive company resources to external agents. Those resources may (unwillingly or intentionally) be used to detrimental effects.

Since cloud computing has become a hot topic in IT management, it may be useful to address some of the security issues involved in this form of cooperation. In terms of information security, the main difference between traditional IT structures and cloud solutions is the fact that cloud infrastructure is interspersed and shared among many users. In

addition, certain features of cloud computing, such as the need for continuous optimization, improved access, balancing the computing load across the nodes, etc., bring additional complications to the risk management process.

At present, three main categories of cloud computing solutions can be distinguished, namely: SaaS - Software as a Service, PaaS - Platform as a Service, and IaaS - Infrastructure as a Service. Factors of potential risk to information security include the following ([12], p. 230):

- IT system's resistance to intrusive attacks from outside,
- Resistance to internal attacks (with users able to access and capture information belonging to other users, by exploiting security holes and other vulnerabilities of the system),
- The security verification and encryption methods in use (whether the access codes and passwords are stored in secure, encrypted form, or stored and transmitted in open text).

Each of the above models of cloud computing (SaaS, PaaS, IaaS) employs a specific set of information security solutions ([12], pp.232-233):

- In the SaaS model, users rely on service provider in all matters concerning information security. The provider is responsible for restricting access to sensitive information, as well as supplying proper security measures to prevent intrusion or breakdown. The provider is also responsible for all matters concerning access verification and data encryption. However, the user is rarely able to examine the details of the security measures taken, to make sure that they are up to the desired standard of service.
- In the PaaS model, the service provider may choose to grant the software provider some control over the system security (for instance, the software provider may take on the responsibility of providing their own access verification and encryption systems), but any security issues beyond the application level, such as the security of host machines or network, come under the administration of the service provider. The provider may choose to pass on to the user selected details on the security measures in use.
- In the IaaS model, the software producer has a great deal of control over security mechanisms, since the cloud applications are run on virtual machines, independent and separated from virtual machines used by other users. However, applications in this model take longer to develop, and are decidedly more costly.

The majority of cloud computing service providers offer data backup solutions. This aspect is clearly important from the user's viewpoint, but it must be noted that data backup is not a 100% solution for all security concerns. The SaaS model, in particular, seems the most risky solution in the context of information security. In this model, both the software code and the data being processed are stored remotely (i.e. outside the subscriber's real machine). Consequently, the user has no access to computing operations, and is in no way able to modify it. SaaS offers the potential for server operator to modify the computing software or data processing procedures. Users must upload their data on the server for computing. The result is the spyware effect: the server operator receives user data freely and effortlessly, due to the character of the service rendered, and this gives him the unfair advantage (or even power) over the user. With the IaaS model, on the other hand, it is advisable to refrain from implementing needless functions on virtual machines, as well as making sure that all virtual machine images communicate over encrypted channels, so as to eliminate the risk of data capture on the network infrastructure level.

According to A. Mateos and J. Rosenberg, the security of the cloud computing environment is comparable to that of most internally managed systems, because:

- Most of the potential (and known) risk problems can be eliminated by employing the existing technologies, such as data encryption and virtual local area networks (VLAN), as well as standard tools, such as firewalls and packet filtering (encrypted data stored on cloud may in fact be safer than non-encrypted data used locally);
- Cloud computing solutions may be supplemented by additional controlling and auditing functionality, layered outside the environment of the host. Such a solution offers the user greater security of cloud computing, far better than any locally implemented solution (since the latter require considerable outlay and design expertise);
- Many countries enforce security measures on SaaS service providers, requiring them to restrict transmission of data and other copyright content to the contractor's country of origin ([13], p. 104)

As aptly put by J. Viega, one of the fundamental benefits of a cloud solution is the fact that those structures are fairly unrewarding for willful intruders, since the code – i.e. the most vulnerable element that can be tested for security holes and exploited – is stored on server side, instead of being sent to the client browser ([12], p.230). Data centralization in cloud structures, as opposed to decentralized distribution of data within the company network, allows for vast reduction of leak risk, since users are less inclined to store sensitive resources on their real machines. Furthermore, the access to a centralized resource storage and actual data use can be monitored more closely.

The concern for security of information exchanged in the course of company relations with external service providers, although well-founded, must be examined against any potential benefits offered by this particular type of business model. And the actual informational risk may be largely minimized by employing proper principles of management with respect to relations with external providers – this applies also to knowledge management.

IV. Ways to Reduce the Risk of Knowledge/Information Loss In It Sourcing Models

K. Liderman believes that information security can be enhanced by employing proper documentation of the security

system in use. This task serves the following purposes ([4], p. 120):

- to ensure proper level of protection with respect to information and to those elements of the system which are directly involved in data processing and storing;
- to track (and control) any changes introduced to the system;
- to satisfy the legal requirements that oblige companies to keep and produce on demand certain documents, such as 'security policy guidelines', 'safety instructions and procedures', etc. (the wording used in actual legal standards may vary).

The use of formal documents (information security policies or guidelines) may attest to the company's intent on keeping proper security standards in data protection. It may also help the organization build and maintain trust relations with customers and/or business partners. Lastly, it may also be used to stimulate the involvement of employees in all tasks and procedures related to data/information security.

With respect to basic technical security measures employed for the purpose of maintaining the informational stability of IT and telecommunication systems, Liderman provides the following classification of elements ([4], p. 158-159):

- data backup procedures;
- provision of independent backup power solutions;
- provision of backup solutions for data processing (or even for running the company business, if necessary), in a reasonably remote location from company HQs;
- doubling the key infrastructure: servers, routers, etc., to serve as backup;
- doubling the information packets;
- providing alternative transmission routes (doubling keys and operators);
- use of verified software offering suitable protection of transmitted and stored data (proper data transmission protocols, software-assisted verification of data integrity based on cryptographic methods, etc.);
- protecting the telecommunication and IT systems from unauthorized access – both physical (access to hardware and technical infrastructure) and logical (access to information resources);
- protection from intentional or accidental exposure to hazards (fire, flooding, strong electromagnetic impulses, etc.).

The most advanced security measures used in cloud computing data centers include (own research, based on [13], pp. 104-114):

- physical security – modern centers are often located in unassuming locations and buildings (often in residential areas), with good security and skillful use of barriers (also natural). Security services cover both the immediate area and the access to actual data facilities, using modern CCTV solutions, intruder alert systems, etc. Servers are kept in fortified bunkers, protected by 5-level biometric scanners (hand geometry recognition), round-the-clock patrols and traps (caging intruders in case of unauthorized entry). Physical security is solely in the hands of the cloud computing service provider, and the above measures are required for certification purposes (the SAS 70 Statement on Auditing Standards No. 70).
- access control in public clouds – these apply to verification of users accessing the cloud. The initial registration of a user is a multi-layered procedure, consisting of several overlapping secret questions and answers (e.g. the user's credit card details). Other levels of security verification may include invoice address, call-back verification over the phone (the *out of band* mechanism, based on employing a different channel of communication), login and authorization (the password should be strong), access keys (a good practice here is to provide regular key substitution service), X.509 certificates, paired keys (the latter being the most important element of user verification when working in cloud environment instances)
- network security and protection of data in large clouds. Passing the task on to the experts employed by the cloud service provider seems the best approach, since it may be reasonably assumed that the provider will be faster to respond to a potential intrusion attempt, and that the response will be adequate to the risk at hand. System security in public cloud models is verified at many levels: at the level of the host's operating system, at the level of the virtual machine's operating environment or the host system, at the stateful firewall, and at the level of signed API calls (the cloud application programming interface), with each subsequent level supplementing the capacities of its immediate precursor.
- The role and the responsibilities of the application owner. Cloud users themselves are responsible for security at the level of their host machine accessing the virtual instance. Since the users have full admin control over their accounts, services, and applications, they are responsible for basic security measures, such as the use of strong passwords, safe storage of passwords and private keys, as well as regular key substitution. Data stored in clouds should also be sufficiently protected – for example, by encrypting the resources prior to uploading them to the cloud, to make sure they cannot be read or modified during transmission and storage.

Modern organizations – both the IT customers and IT service providers – should strive to identify and recognize all processes, services and resources considered mission-critical or of key importance from the information security viewpoint. They should also perform a reliable analysis of information risks, and take suitable measures and procedures to minimize the risk over the course of the cooperation with external partners. Thus, irrespective of the security solutions on offer by the service provider, they should employ their own, independent backup procedures with respect to sensitive data – such backup may be of great value if the company de-

cides to withdraw from the contract (in such cases, the provider may refuse access to data stored in their system) or if the provider goes bankrupt.

Companies unwilling to put their trust in external providers, despite numerous obvious benefits offered by cloud computing solutions, can always reach for other models, such as those based on insourcing or the private cloud model.

The insourcing solution is based on internal management of IT services. If need arises, the company may purchase the lacking skills on the market for a limited time, for example by contracting additional personnel for the task. In this model, the organization retains its internal IT personnel and infrastructure, trusting in its ability to free the latent potential of its employees for the purpose of improving its IT services and making them more effective. From the viewpoint of the insourcing model, the internal IT department is formally perceived as a provider of services.

In the case of private cloud solutions, the decision to adopt this particular business model is made on the basis of four fundamental factors: security, accessibility, the size of user population, and the effect of scale (Table 1).

TABLE 1. PREMISES FOR ADOPTING A PRIVATE CLOUD SOLUTION

Factor	Description
Security	Applications require direct control and data protection, for confidentiality and safety reasons (for instance, governmental agencies use dedicated applications for processing of confidential and classified data – it is essential that they be kept from unauthorized access).
Accessibility	Applications require access to a predefined set of processing resources, and this type of access cannot be securely provided in a shared environment.
User population	The organization caters for many users, often in geographically remote locations, and they all require unrestricted access to computational processing resources (private clouds are used, for example, in large telecommunication corporation).
The effects of scale	Data centers and infrastructural resources are readily available, or can be expanded at minimum cost.

Source: own research based on ([13], p.116).

Private clouds offer better control and assurance that the resources will not be used by other customers, since they are not shared in public space. However, as any other solution, the private cloud model has its own limitations, such as (own research, based on [13], pp.119-120):

- limited scale of operation, compared to public clouds,
- the problem of adopting old applications to the cloud structure requirements without redesigning the very architecture of the system,
- limited potential for optimization and innovation of the methods and elements of the system,
- larger operational outlay compared to the public cloud solutions.

Even if the organization does not anticipate any integration with external providers when choosing their outsourcing solution, it may be advisable to keep an open stance in this respect, so that it may smoothly transition to another model if need arises, and not be too restricted with their choice of a potential provider.

V. CONCLUSIONS

New technologies, and the resulting new models and instruments for business, generate new and previously unforeseen risks and threats. Changes in company operating environment, brought about by globalization, increased competition, automation and – most of all – computerization, informatization and virtualization – require a new approach to information security in modern organizations.

As discussed in this paper, new IT sourcing models, especially cloud computing, offer some opportunities, but even if organizations themselves feel "cloud ready," they must anticipate the capacity requirements in the cloud. They must also be aware of new risks, and manage their IT security in accordance with the new operating conditions. The most important risk areas with respect to modern IT sourcing solutions (similarly to those observed in classical outsourcing models) include: the loss of control over the IT environment, inadequate protection of data, overdependence on external suppliers, the loss of potential to switch back to previous (self-contained) IT services, etc. A decision to adopt a particular IT sourcing solution should be based on such factors as: the size of the organization, the scale of operation, risk propensity, the adopted information security policy, the personnel strategy, and the budget.

It seems that migration to a cloud model is a good solution for companies intent on maximizing their profits (cloud computing services are decidedly more cost-effective) while at the same time retaining their high standards of security. What makes the cloud computing particularly attractive for business entities is the fact that they can pass most of the IT system security responsibilities on to the service provider. The providers of cloud computing services, being well aware of the fact that security concerns are the most important factor to restrain companies from choosing the cloud model, make huge investments in security solutions and infrastructure, as a way to emphasize their responsible approach to the security of their clients' resources. Companies which – for a number of reasons – are unable or unwilling to rely on external partners with their data, can reach for other sourcing models, such as the private cloud model or the insourcing model, to improve their IT effectiveness.

REFERENCES

[1] Oshri, I., Kotlarski, J., Willcocks, L.P., 2011, *The handbook of global outsourcing and offshoring*. Second edition, Palgrave Macmillan Ltd. – Houndmills Basingstoke Hampshire (UK) .

[2] Morgan, J.L., Bravard, R., 2010, *Inteligentny outsourcing. Sztuka skutecznej współpracy*, MT Biznes Sp. z o.o., Polska.

[3] Szpor, G., Wiewiórowski, W. R. (eds.), 2012, *Internet. Prawno – informatyczne problemy sieci, portali i e- usług*, Wydawnictwo C.H. Beck, Warszawa.

[4] Liderman, K., 2012, *Bezpieczeństwo informacyjne*, Wydawnictwo Naukowe PWN, Warszawa.
[5] Rot, A., Sobińska, M., *IT Security Threats in Cloud Computing Sourcing Model*, Proceedings of the Federal Conference on Computer Science and Information Systems (2013 Federated Conference on Computer Science and Information Systems (FedCSIS),), pp. 1299 – 1303.
[6] Sobińska, M., *IT management business model - sourcing IT services,* a chapter in *Networking Models in Virtual Enterprises,* K. Perechuda (ed.) – in review (2014)
[7] Sparrow, E., 2003, *Successful IT Outourcing*. Springer, London.
[8] *Strategies To Improve IT Efficiency In 2010.Using Predictive Analysis To Do More with Less*, April 13, 2010, A Forrester Consulting Thought Leadership Paper Commissioned By TeamQuest, http://www.teamquest.com/pdfs/whitepaper/forrester-it-efficiency-2010.pdf- accessed on 18.04.2013.
[9] Szpringer, W., 2008, *Wpływ wirtualizacji przedsiębiorstw na modele e-biznesu. Ujęcie instytucjonalne*, Oficyna Wydawnicza Szkoły Głównej Handlowej w Warszawie, Warszawa.
[10] Willcocks, L.P., Lacity, M.C., 2012, *The new IT outsourcing landscape. From innovation to cloud computing,* Palgrave Macmillan Ltd. – Houndmills Basingstoke Hampshire (UK).
[11] *Firmy lekceważą cyfrowe ataki*, Puls Biznesu, 27 Nov. 2013
[12] Viega J., 2010, *Mity bezpieczeństwa IT. Czy na pewno nie masz się czego bać?*, Helion.
[13] Mateos , A., Rosenberg, J., 2011, *Chmura obliczeniowa. Rozwiązania dla biznesu*, Helion, Gliwice.

Joint Agent-oriented Workshops in Synergy

Joint Agent-oriented Workshops in Synergy is a coalition of agent-oriented workshops that come together to build upon synergies of interests and aim at bringing together researchers from the agent community for lively discussions and exchange of ideas. For the first time JAWS was organized during the 2011 FedCSIS Conference.

Workshops that constitute JAWS in 2014 are:

- MAS&M'2014 - International Workshop on Multi-Agent Systems and Simulation
- SEN-MAS'2014 - 3rd International Workshop on Smart Energy Networks & Multi-Agent Systems

8th International Workshop on Multi-Agent Systems and Simulation

Multi-agent systems (MASs) provide powerful models for representing both real-world systems and applications with an appropriate degree of complexity and dynamics. Several research and industrial experiences have already shown that the use of MASs offers advantages in a wide range of application domains (e.g. financial, economic, social, logistic, chemical, engineering). When MASs represent software applications to be effectively delivered, they need to be validated and evaluated before their deployment and execution, thus methodologies that support validation and evaluation through simulation of the MAS under development are highly required. In other emerging areas (e.g. ACE, ACF), MASs are designed for representing systems at different levels of complexity through the use of autonomous, goal-driven and interacting entities organized into societies which exhibit emergent properties The agent-based model of a system can then be executed to simulate the behavior of the complete system so that knowledge of the behaviors of the entities (micro-level) produce an understanding of the overall outcome at the system-level (macro-level). In both cases (MASs as software applications and MASs as models for the analysis of complex systems), simulation plays a crucial role that needs to be further investigated.

Topics

MAS&S'14 aims at providing a forum for discussing recent advances in Engineering Complex Systems by exploiting Agent-Based Modeling and Simulation. In particular, the areas of interest are the following (although this list should not be considered as exclusive):

- Agent-based simulation techniques and methodologies
- Discrete-event simulation of Multi-Agent Systems
- Simulation as validation tool for the development process of MAS
- Agent-oriented methodologies incorporating simulation tools
- MAS simulation driven by formal models
- MAS simulation toolkits and frameworks
- Testing vs. simulation of MAS
- Industrial case studies based on MAS and simulation/testing
- Agent-based Modeling and Simulation (ABMS)
- Agent Computational Economics (ACE)
- Agent Computational Finance (ACF)
- Agent-based simulation of networked systems
- Scalability in agent-based simulation

Steering Committee

Cossentino, Massimo, ICAR-CNR, Italy
Fortino, Giancarlo, Universita della Calabria, Italy
Gleizes, Marie-Pierre, Universite Paul Sabatier, France
Hilaire, Vincent, Université de Belfort-Montbéliard, France
Pavon, Juan, Universidad Complutense de Madrid, Spain
Russo, Wilma, Universita della Calabria, Italy

Event Chairs

Cossentino, Massimo, ICAR-CNR, Italy
Fortino, Giancarlo, Universita della Calabria, Italy
Hilaire, Vincent, Université de Belfort-Montbéliard, France

Program Committee

Arcangeli, Jean-Paul, Université Paul Sabatier, France
Bernon, Carole, Universite Paul Sabatier, France
Botía, Juan, Universidad de Murcia, Spain
Davidsson, Paul, Malmö University, Sweden
Garro, Alfredo, University of Calabria, Italy
Gomez-Sanz, Jorge J., Universidad Complutense de Madrid, Spain
Gravina, Raffaele, University of Calabria, Italy
Guerrieri, Antonio, University of Calabria, Italy
Hassan, Samer, Universidad Complutense de Madrid, Spain
Jedrzejowicz, Piotr, Gdynia Maritime University, Poland
Klügl, Franziska, Örebro Universitet, Sweden
López-Paredes, Adolfo, INSISOC - University of Valladolid, Spain
Molesini, Ambra, Università di Bologna, Italy
Niazi, Muaz, Bahria University, Pakistan
Petta, Paolo, OFAI, Austria
Picard, Gauthier, EMSE, Saint Etienne, France
Seidita, Valeria, Università degli Studi di Palermo, Italy
Terna, Pietro, Università di Torino, Italy
Vizzari, Giuseppe, Università di Milano Bicocca, Italy

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1453–1460
DOI: 10.15439/2014F335
ACSIS, Vol. 2

Opening Pandora's Box: Some Insight into the Inner Workings of an Agent-Based Simulation Environment

Daniel Dawson
School of Computer Science,
University of Nottingham,
Nottingham, UK
Email: ddawson4417@gmail.com

Peer-Olaf Siebers
School of Computer Science,
University of Nottingham,
Nottingham, UK
Email: pos@cs.nott.ac.uk

Tuong Manh Vu
School of Computer Science,
University of Nottingham,
Nottingham, UK
Email:
psxtmvu@nottingham.ac.uk

***Abstract*— Agent-Based Simulation (ABS) environments are somewhat of a black box to many modelers in Social Simulation or Economics and their inner workings are often only understood by the computer scientists who developed them. We intend to shed some light into the inner workings of such systems. For this purpose we have developed our own simple ABS environment in C++ using hierarchical state machines. In this paper we provide insight into the design of our ABS environment and then test the performance of it by comparing it to that of an "off the shelf" commercial package. While some programming knowledge is required to understand the paper in all its depth we believe that non programming experts will also benefit from this paper as it provides an insight into the underlying mechanisms operating within an ABS using graphical representations and explanations that avoid heavy technical jargon.**

I. Introduction

An Agent-Based Simulation (ABS) environment is a system in which a population of agents (autonomous objects that behave in a predefined way) are created using a template in order to investigate the consequences of these agents acting together in an environment. The application area that we focus on in this paper is Social Simulation [1] and we use examples from computer games (which are related to Social Simulation but much easier to understand) to describe the agents we created during this investigation. The very simple experiments we conduct help us to understand situations which can otherwise be difficult to replicate. ABS experiments can sometimes yield unexpected results, for example an ABS constructed by Bonabeau [2] revealed that placing a column in front of an emergency exit can improve the flow of people out of the exit in an emergency situation, which is not the first idea which common sense would dictate.

This paper will describe and promote the understanding of the inner workings of an ABS environment that has been developed from the ground up. For the software engineering process (i.e. the development of the ABS environment) we take our ideas from the Multi-Agent Systems field. The models we implement during the validation phase are those typically created in the Agent-Based Modelling community (e.g. by Social Scientists). A good explanation about the relevant differences between both fields can be found in [3].

In order to explore how an ABS system works, and subsequently construct one, it is necessary to first explore the concepts that are involved in creating such a system. Simple ABS systems are generally implemented using finite state machines. Once the behaviour of agents gets more complex the introduction of hierarchical state machines becomes necessary to avoid the over-complication of finite state machines, leading to state machines that can be notoriously difficult to fully understand. Object-oriented design principles will be used in the construction of our tool in order to promote its extensibility, allowing anyone to add features or extend classes where it will benefit them. The Unified Modelling Language (UML) is a graphical notation that is often employed in Software Engineering for conducting object oriented analysis and design. AgentUML is an extension of UML that is specifically used for the development of multi-agent systems [4], as for example mobile agents [5], [6]. However, in the field of Social Simulation it is still rarely used for developing agent-based simulation models that represent social processes [7]. In this paper we use the UML on the one hand to show the structure of the proposed ABS environment (in form of a class diagram) and we use statecharts on the other hand for the design of our agent based models (i.e. to represent the behaviour of our agents) as proposed by [8].

In order to provide the reader with the necessary understanding of all of the topics within this paper we first provide a short introduction to object-oriented methods, agent-based modelling and the concepts of state machines. After this we illustrate the design of the ABS system which we have implemented. Finally we focus on the validation of our system in order to demonstrate that its components adhere to the standards of an ABS system. During the validation process we also take a look at the efficiency of our system compared to an "off the shelf" commercial package in terms of memory usage and runtime.

II. Background

A. Object Oriented Methods

Object-orientation is an important concept for designing software in order to promote extendibility of existing systems. The object model encompasses the core principles of abstraction, encapsulation, modularity, and hierarchy [9].

978-83-60810-58-3/$25.00

It leads to reusable components, wherever possible, rather than the more bespoke solutions which procedural programming often offers. It also breaks programs down into understandable chunks, and by designing software with object-oriented methods in mind it can be more easily extended and fixed when problems arise.

In order to promote object oriented principles in our design we have taken a number of design patterns into consideration while designing our system. A design pattern is a standardized way of implementing a certain feature in a program [10] and makes it easier to reuse successful designs and architectures [11]. This will be discussed in more detail in Section III-A.

UML is a graphical notation commonly used in software engineering for the purpose of object oriented analysis and design. Through UML it is possible to visualise, specify, construct, and document software applications. It acts as a specification language in which we can precisely and unambiguously capture our design decision [12]. Besides the benefits for software design some of the diagrams (e.g. use case diagrams and state machine diagrams) seem to lend themselves particularly well to Agent-Based Modelling (ABM) [13]. Therefore we use the UML notation not only for designing our ABS environment but also for modelling the agents that we use within our system.

B. Agent-Based Modelling

In order to get a good picture of what an ABS environment is, we first need to define what the term agent means, including the principles on which an agent acts and behaves. In the eyes of software engineers agents are simply "objects with attitude" [14] in the sense of them being objects with some additional behaviour added, for instance, mobility, inference, etc. But there are a number of different conflicting views on what an agent is, depending on the situation and discipline for which it is being used [15]. However, often there is a point where the views start to overlap with each other. Castle and Crooks [16] discuss different points of view and merge them together to form a universal definition of an agent, which varying disciplines can agree with. Closely related to their definition we understand an agent to be an autonomous object with some memory, which is able to make individual decisions based on influences from its environment (e.g. messages received from other agents).

The agents which we intend to create have the ability to make decisions based on internal transition trigger rules which might be influenced by the environment they observe. These transition trigger rules which have been programmed into the state machine of the agent and most often fall into one of three categories: condition-based, time-based, or message-based [17]. Details about different transition types can be found in Section III-C.

An agent is often described as having some sort of memory, which comprises of the last state they were in or in the case of composite states, the last super or sub-state they were in. This is the concept of state history, and there are two types: Deep history and shallow history. Deep history goes through multiple levels of composite state, and will return to the last state within a state within a state etc. Shallow history will only return the last state to within a state.

The last major thing to consider with agents is a form of control. As mentioned previously, finite state machines are often used in the creation of agents, as a means to describe the behaviour of an agent, or in other words, a template for how they should act, with conditions specifying when a transition should be made. This leads us onto the topic of Finite State Machines (FSMs).

C. Finite State Machines

A Finite-State Machine (FSM) is conceived as an abstract machine that can be in one of a finite number of states. It can change from one state to another when initiated by a triggering event or condition; this is called a transition [18]. There are different types of FSMs that can be used in a variety of different situations. We distinguish in this paper specifically between deterministic and stochastic FSMs. While deterministic FSMs are based on mathematical formulas and can be formally proven, stochastic FSMs use stochastic rules for deciding about transitions and therefore the exact outcome cannot be predicted; however it may be estimated using models and theories. A turnstile is a good example of a simple deterministic FSM. It can either be locked or unlocked, and there are predefined conditions that determine which state the turnstile is in, and the transitions it can take from each state. Fig. 1 demonstrates this.

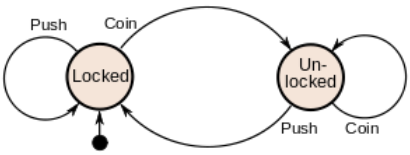


Fig. 1 A simple deterministic FSM (Source: [19])

More complex state machines are often used for economic models [20] and the transition function, which defines the transitions between states, is a lot more complex. For ABM we normally use stochastic FSMs. When behavioural models start to be implemented, FSMs can quickly become complex without some sort of organisation. In such cases Hierarchical State Machines (HSMs) are be introduced in order to keep such a model understandable. In HSMs states may contain other FSMs. This is often programmatically done with nested "if" statements.

D. Hierarchical State Machines

A HSM is similar to the composite state we see in UML state diagrams, and it provides the same functionality. One of the simplest ways to describe a HSM is by showing an example of one of its uses in Game AI design. Non Player Characters (NPCs) must have the ability to act in a complex way in order to give the player a challenge. If all of this behaviour were to be coded with a single if-then-else block, the code would quickly become hard both to manage, and for programmers to read and extend, specifically where the more recent generation of game AI is concerned. Fig. 2

shows the first level of a HSM of a typical NPC enemy in a typical first person shooter game.

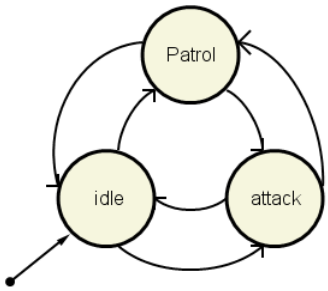


Fig. 2 Typical FSM for an NPC guard in a game

The HSM provides a structural concept for representing behaviour, which can otherwise be complex and resulting in somewhat of "Spaghetti" code when trying to split into logical if statements, in order to structure the functionality into logical categories of behaviour. Fig. 3 shows the hierarchical state "attack" from Fig. 2.

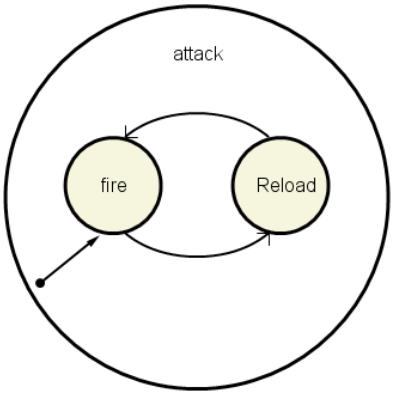


Fig. 3 Hierarchical attack state from NPC guard

Reference [21] talks about an object-oriented software tool that is used for creating the behaviour for NPCs in games by making use of HSMs. The system uses the logical components of a state machine, as well as making use of object-oriented principles such as inheritance and design patterns. This is similar to the work we are conducting, whereby state machines are being used as a template for the behaviour of entities. In our tool, more focus will be placed on the collaboration and communication of agents, unlike NPCs which often have no need to communicate or interact with each other.

E. UML State Machine Diagrams

The UML state machine diagram (also called statechart) is used to depict HSMs. Elements of this diagram are states, transitions, and composite states (which are equivalent to hierarchical states). Fig. 4 shows a UML state machine diagram of an office worker.

The office worker has three main states: "atHome", "elseWhere" and "atOffice". The "atOffice" state shows that while the worker may be at the office, there are still two sub-states that the worker can be in - "working" and "dozing". The statechart entry (the initial state the state machine is initialized into) is represented by the uppermost symbol in

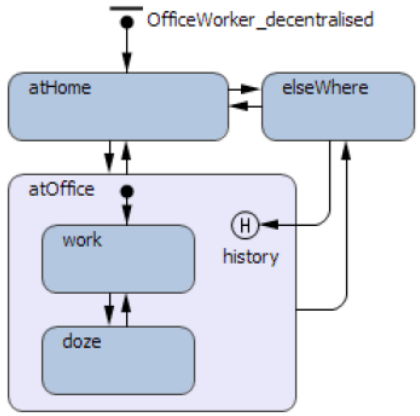


Fig. 4 UML diagram for an office worker (Source: [13])

the state-chart, the circle with an arrow with a line over it. The history state in this example can be described using the following scenario: If a worker was doing work, but then decided to take a break and go elsewhere, they would return to what they were doing before the break. The history state provides this capability. However, if the worker was dozing and then took a break, they would not start at the entry state of working when returning, but instead would return to the dozing state.

III. DESIGN

The design of the finite state machines that run the agents is based on the logical components of a state machine, as is the case in UML, where state machines have states, transitions, and composite states (state machines within states). The agent of this system contains the information relevant to the state machine, which includes the current state, last known state and history states for the super-state it was last in. This is a more memory-efficient way of storing the information, rather than having each agent assigned its own state machine, although the logic behind this is explained later.

A. Design Description

This section will describe the design of the system, including a simple class diagram of the main components of the agent. For the implementation we have decided to use C++ which in some cases has influenced our design decisions (e.g. multiple inheritance is not supported in Java). Fig. 5 shows the classes and initial relationships between each of the classes in our system.

The numbers in Fig. 5 represent the associations between objects, with * representing any number of. For example, the relationship between `Agent` and `StateMachine` is that there is one `StateMachine` to one `Agent`, and the relationship between `StateMachine` and `State` is that one `StateMachine` is composed of one or more `State` objects. The hollow arrowhead represents inheritance, showing that `CompositeState` inherits properties from both a `State` and a `StateMachine`. The `Attributes` of each object are the variables stored within the object, and the `Operations` are a list of methods which are used

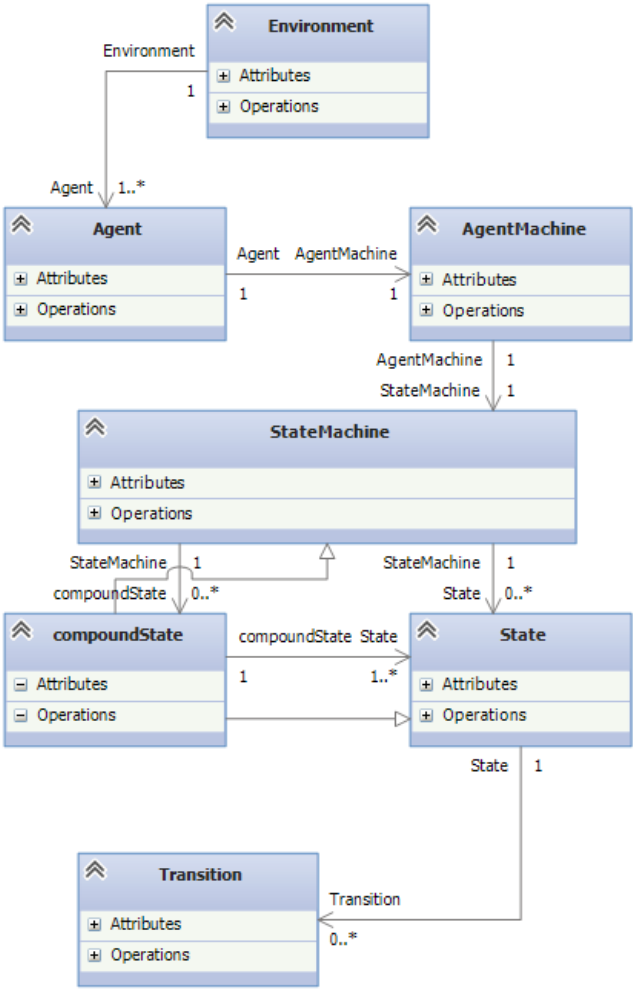


Fig. 5 UML class diagram for our ABS environment

within the object. As such, the object can be accessed as either a `State` or a `StateMachine`. The full class diagram including the methods that are used internally is available upon request.

The way the system is designed ensures object-oriented principles are taken into account so as any one of the objects in the system can be extended. The `CompositeState` class makes use of inheritance to reuse the code from `StateMachine` and `State`, since it exhibits the behaviours of both classes. A number of design patterns such as the Observer pattern, the State pattern, and the Model-View-Controller pattern have been taken into consideration during the design process. For example, the Observer pattern defines a one-to-many dependency between objects so that when one object changes state, all its dependents are notified and updated automatically [10]. In our case the `Agent` class is constructed using the ideas of the observer design pattern, allowing agents to constantly monitor transition conditions and allow them to enact the transitions themselves. This removes the need for the transition conditions to be explicitly checked.

In the following sub-sections we provide some more detailed design strategies for the key elements of our ABS environment. Here we look at the environment and agent design, the state machine design and explain the different time models we used.

B. Environment and Agent Design

The `Environment` in our ABS environment is a container for the agents, which manages their creation and deletion. This class also handles the sending of messages intended for all agents rather than singular agents. The agent class is designed to hold and handle the state machine for the instance of the agent, including telling the state machine to advance a time step in model time. There are two possible ways of controlling the state machine which links to an agent. The first (and simpler) way is to assign a state machine to each agent. The second (more complex) way, proposed by [21], is to use a simplified representation of a state machine without having to create a full machine for each agent; a way which can save a lot of memory. In the latter case the information for the current state of the agent is stored in the agent itself, whilst the state machine stores all of the logic for the states and transitions. The reason for storing information this way, rather than having a machine for each agent, is that even for simple state machines, a high number of objects will take up a lot of memory regardless of how small these objects are. Due to the object-oriented nature of this design, each state machine object requires the creation of every state and transition present in that state machine. The number of objects soon becomes very large, and becomes somewhat of a waste of memory, albeit at the expense of slower processing of state changes. However if a small number of agents were being created or the user were not concerned about memory usage, the gain in state change speed may be preferable.

Let us illustrate this principle with an example. Presume a state machine comprises of a total of 10 states and 10 transitions, and each object takes 1 byte of memory. In order to create 100,000 agents, 1 state machine + 10 states + 10 transitions + 1 agent object will be created for each agent. This totals 22 * 100,000 objects, so 2,200,000 objects in memory. If there is only 1 state machine acting as a template, there are 100,000 agent objects, plus 1 state machine, 10 states and 10 transitions, totalling 100,022 objects in memory. Of course, the larger the state machine is, the more effect it has on the size of the agent. Since we are trying to save memory here, the choice is only logical. We will provide evidence for the memory saving capabilities of this solution in Section IV when we validate our ABS environment.

C. State Machine Design

A state machine has logical components to it, which makes it easy to split up into the objects we talk about in object-oriented programming. The typical state machine is fairly simple and composed of a set of states, and a set of transitions. Composite states however, whilst being regarded as a state within the agent state machine, effectively contain their own state machine. This can be recursive, and there can be many composite states within another state. Fig. 6 shows

a UML diagram similar to the NPC in Fig. 2 but more complex.

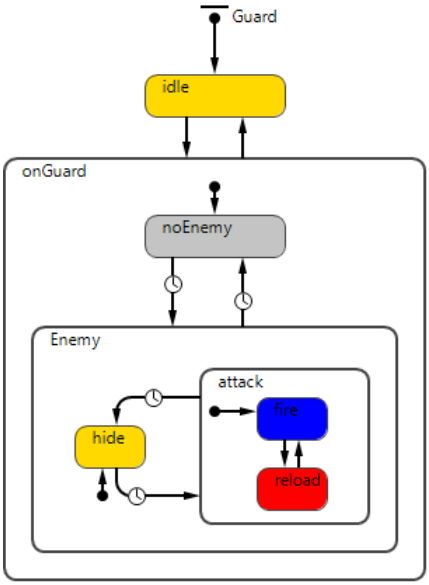


Fig. 6 UML version of a more complex NPC Guard

What we can see in Fig. 6 is that within each composite state there is an initial state pointer. This indicates that a composite state effectively contains a state machine, and so exhibits properties of both, a state machine and a normal state. Therefore it makes sense to reuse both, the state and state machine object, through use of multiple inheritance. Inheriting properties from multiple objects can be a complex problem in object-oriented programming [22]. However, in our case it seems an appropriate solution to the `CompositeState` problem.

Transitions within a state machine have certain conditions which cause the transition to trigger, and in the case of ABM the transitions can be triggered by a number of different things. We have considered the four following triggers which are often associated with transitions: condition; message arrival; rate; timeout. These are quite basic transitions which will cover transitions most people need to implement, however if any other types are needed, the transition class can be extended appropriately.

Condition-based transitions can be specified by the programmer when a certain condition is met. Since this condition could be any number of things (a block of code which can return true or false), there are two feasible ways of implementing such a transition. The first method would be by introducing time steps to the model. This would be using a synchronous time model, and therefore perhaps not the best way to implement things which reflect the real world, in particular in Social Simulation. We as people are not expected to make decisions every arbitrary unit of time, we simply make decisions when they need to be made. Asynchronous time, which does not specify time steps and instead relies on agents doing things when a condition is met, would reflect the real world more accurately, and is also more computationally efficient. One way of implementing such a time model would be to use the observer pattern. This means that once a condition is needed for a transition, an observer will be added to monitor the condition, and when the condition is satisfied, the appropriate object will be notified. In our case, this will be the lightweight object which is used to represent the agent instance of our state machine.

Time-based transitions are based on either timeouts, or a rate at which agents move from one state to another. The method in which this transition can be implemented is similar to Boolean condition based; however an easier way of implementing such a transition in C++ would be simply to set a timer, which notifies the agent instance of a state machine upon expiry. This eliminates the need for an additional observer to be added to the system.

Message-based transitions are the ones which are triggered upon receiving a message. No particular observer is required for this, as messages should be processed upon arrival at an agent. When a message arrives, it will be stored by the agent and a transition from the current state will be triggered if that particular message matches the condition of the agent.

D. Time Models

Agents need to periodically make decisions, which are based on the transitions in the state machine. There are two main ways of doing this: using an asynchronous and synchronous time model. Both ways have their advantages and disadvantages depending on what type of model is being created.

Asynchronous time models are the most commonly used when trying to model real world situations where an agent acts of their own accord at random time intervals [23]. This reflects best what happens in the real world in social systems. It can also be said that typically asynchronous models are more efficient in terms of computational expense, due to the fact that things are only triggered when they need to be triggered, whereas the same model run in asynchronous time will trigger at every time step regardless of whether it is necessary or not.

Synchronous time models on the other hand are where "time steps" are defined, and each time step triggers an agent to perform an action. This action can directly trigger a transition, or it can signal to the agent that it is time to make a decision on whether to make a transition or not. When using synchronous time models, time steps can notify an agent based on probability, effectively imitating asynchronous time models.

In our system we have the choice of which time model to use. Obviously there are advantages and disadvantages for both and it really depends on the application, which time model should be used.

IV. Validation

Our ABS environment has been implemented in Visual Studio 12 using C++ as the programming language. The "off the shelf" commercial package we have chosen for our validation experiments is AnyLogic 7, which is a multi-paradigm eclipse based simulation IDE that supports

graphical model design for all major simulation paradigms (including ABM) [24]. In order to assess run time and memory usage we run the AnyLogic 7 and the C++ implementation on their own with no other processes using significant RAM or CPU time. The experiments were run on the same system to ensure fairness.

For this paper we have conducted two validation experiments which test individual components of the agent, including simple behaviour and composite states. We used the synchronous time model for our tests due to the fact that only time-based conditions were included in the model. We have conducted further experiments using the asynchronous time model as well as all types of transition triggers but due to the limited space we cannot present them here. Results of those experiments are available on request.

A. Overview of Validation Experiments

We have constructed a number of experiments to test the individual components of our system. The purpose of these experiments is to verify that the models running in our ABS environment behave as expected and we also compare the performance (in terms of runtime and memory usage) of our system to that of AnyLogic 7.

With regards to creating the models both simulation systems have a very different approach: Our system requires some basic C++ programming while in AnyLogic you can use drag/drop to create UML charts and then simply set up the transition triggers. The AnyLogic model is then automatically translated into Java code and is ready to be executed from within the AnyLogic environment. For our system, in order to construct the state machines for the experiments, an empty class implementing `StateMachine` is created, first listing each of the states as variables, including composite states. The transitions are then defined, and finally all the states and transitions linked together by using an `addTransition` method.

B. Validation: Experiment 1

The first experiment focuses on testing the simple decision-making capabilities of an agent. The state machine used to control behaviour here consists of two states: `stateRed` and `stateBlue`. When created, each agent initially is in `stateBlue` and based on probability takes a transition that leaves it to a final state `stateRed`. Once in `stateRed` the agent will not change states any more. The UML statechart for the simple agent is provided in Fig. 7. The initial environment was run using 400 agents.

A counter was used in each instance to count the number of agents which are in each state initially, and after 5 time steps. Fig. 8 shows our system after 5 time steps (with 0 representing `stateBlue` and 1 representing `stateRed`), and Fig. 9 shows the counter variables for each state in AnyLogic.

It can be seen that there is a slight difference in the number of agents in each state, due to random number generation in each instance yielding different results. This shows that the agents in our system behave as expected. The

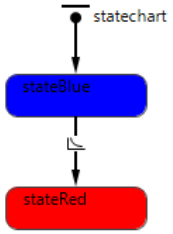


Fig. 7: UML state-chart for a simple state machine agent

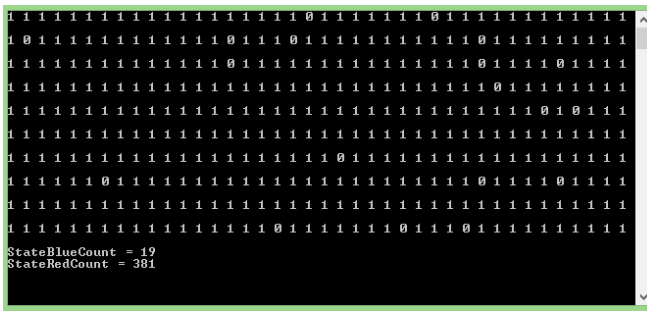


Fig. 8: Our system model after 5 time steps

Fig. 9: AnyLogic model after 5 time steps

experiment was then altered to run in virtual mode in AnyLogic, which simply executes transitions as fast as possible, and was also run without any GUI in our system. The results in each instance are shown in Tables I and II. Since the time taken cannot be easily recorded in AnyLogic, the number of time steps executed in virtual time over 10 seconds was gathered, and averaged per second. The time steps per second on our model was calculated using the time taken for all of the transitions to complete.

TABLE I.
EXPERIMENT 1 RESULTS USING OUR SYSTEM

Agents	Ram	Timesteps
500	0.5 MB	1000 / sec
100,000	14.4 MB	2.92 / sec
250,000	34.1 MB	1.68 / sec
1,000,000	134.5 MB	0.38 / sec

TABLE II.
EXPERIMENT 1 RESULTS USING ANYLOGIC

Agents	Ram	Timesteps
500	90 MB	13063 / sec
100,000	180 MB	19.7 / sec
250,000	320 MB	6.4 / sec
1,000,000	1,260 MB	1.34 / sec

As can be seen in Table I, the number of time steps executed per second is linearly affected by the number of

agents present in the model, and the memory used also grows fairly linearly with the amount of agents when concerned with a large agent population. Table II shows that AnyLogic also exhibits the same linear increase of memory usage with agent population size; however, this memory usage is far higher than our tool. The amount of time steps that is executed per second is also quite drastically affected in AnyLogic when a large number of agents are put into the simulation, although it is faster than with our tool. This could be due to the use of the lightweight state object which is used to save memory, and a performance increase may be obtained with the expense of more memory usage by assigning an individual state machine to each agent. It should be noted that AnyLogic can in fact execute multiple transitions per time step, so the number of transitions taken may be more than the time steps executed. This shows us that an improvement needs to be made to our tool in terms of the speed of the model, focusing on the use of the lightweight `AgentMachine` object used to save memory.

C. Validation: Experiment 2

This experiment implements the idea of composite state machines. In addition to the simple agent behaviour experiment, the `stateRed` state will now contain a clone of the previous state machine. The agent should move into the stateRed state and move between the states within this composite state. The UML statechart for the more complex agent is provided in Fig. 10.

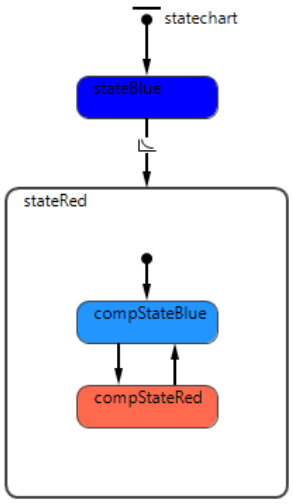


Fig. 10: UML state-chart with a composite state machine agent

As with the previous experiment, a counter was used in both AnyLogic and our tool to count the number of agents in each state. Fig. 11 and 12 illustrate the number of agents in each state after 5 time steps in both, our system and in AnyLogic. In Fig. 11, the numbers 0 and 1 represent `stateBlue` and `stateRed`, respectively, and the numbers 2 and 3 represent `compStateBlue` and `compStateRed`, respectively. As can be seen in Figs. 11 & 12, the number of agents in each state varies between our tool and AnyLogic due to random number being used to determine whether a transition should be made, however the agent behaves as expected with a composite state.

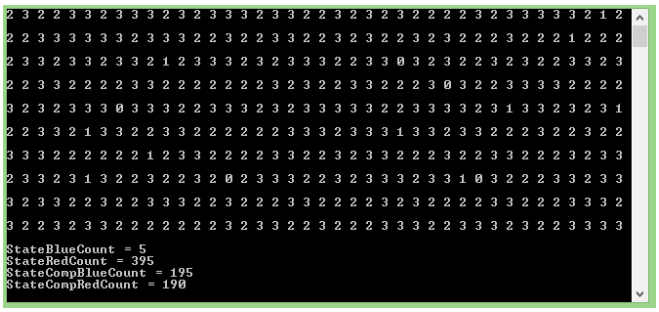


Fig. 11: Our system model after 5 time steps

Fig. 12: AnyLogic model after 5 time steps

As previously, the experiment was then altered to run in virtual mode in AnyLogic, which simply executes transitions as fast as possible, and without any GUI in our system. The results are presented in Table III for our system and in Table IV for AnyLogic.

TABLE III.
EXPERIMENT 2 RESULTS USING OUR SYSTEM

Agents	Ram	Timesteps
500	0.5 MB	1000 / sec
100,000	14.4 MB	2.99 / sec
250,000	34.1 MB	1.22 / sec
1,000,000	134.7 MB	0.29 / sec

TABLE IV.
EXPERIMENT 2 RESULTS USING ANYLOGIC

Agents	Ram	Timesteps
500	75 MB	9135 / sec
100,000	346 MB	16.5 / sec
250,000	540 MB	5.7 / sec
1,000,000	1,615 MB	1.5 / sec

Table III shows that the performance for our system has remained almost the same as in Experiment 1 in terms of memory usage and time steps executed per second. The same is true for AnyLogic in relation to the time steps executed per second, as can be seen in Table IV. However, memory usage has gone up quite significantly. This demonstrates the advantages of our state machine design. One can save potentially a lot of memory when constructing models that feature more complex statecharts.

V. CONCLUSIONS

This paper provides an insight into the often unexplored inside world of ABS environments. We have defined the different components (classes) of such a system, including `Environment`, `Agent`, `StateMachine`, `State`,

`Transition`, and `CompositeState`. For each of these, we have defined what is expected of them in an ABS system, the links between them, and in addition we have explored a variety of different ways in which these can be implemented whilst fulfilling the requirements of an ABS system. We have also designed and implemented some basic tests which showed that the components work as expected.

In Section IV we have compared the performance of our system to that of the "off the shelf" package AnyLogic. The results demonstrate that with a synchronous time model in mind, memory usage of our tool is much lower than that of AnyLogic. This demonstrates the value of our ABS environment when considering larger models where memory usage can be of high importance, and where large agent populations are being implemented. However, there is still much room for improvement in terms of performance, which has become apparent from our experimentation.

We have extensively used object oriented analysis and design principles to create a system that is easy for others to use and extend. This should allow others to easily implement features they might want to include. Through providing UML diagrams we are hopeful that we have helped non computer scientists to understand, perhaps for the first time, how ABS works internally. Through our tests we have demonstrated that our C++ implementation is a promising solution for use within the area of ABS when memory usage is of a higher priority than features.

Currently we are continuing our validation efforts by building more complex real world models in our ABS environment and comparing their performance to the performance of existing models that we have previously built in AnyLogic, e.g. [25]-[27]. With regards to extending the ABS environment we are considering to provide researchers with an easier way of creating simple state machines. XML would be a useful file format to consider for storing models of finite state machines, as all the information on states and transitions could be stored in a structured way, and many tools which are used to draw UML diagrams support the XML file format. Here we follow the idea of [21] where a custom file format is used to interpret HSMs.

REFERENCES

[1] M. W. Macy and R. Willer, "From factors to actors: Computational sociology and agent-based modeling", *Annual Review of Sociology*, 28(1), pp. 143-166, 2002. DOI: 10.1146/annurev.soc.28.110601.141117

[2] E. Bonabeau, "Agent-based modeling: Methods and techniques for simulating human systems", *Proceedings of the National Academy of Sciences of the United States of America*, 99, pp. 7280-7287, 2002. DOI: 10.1073/pnas.082080899.

[3] Y. Shoham and K. Leyton-Brown, *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*, Cambridge University Press, 2008.

[4] B. Bauer, J. P. Müller, and J. Odell, "Agent UML: A formalism for specifying multiagent software systems", *International Journal of Software Engineering and Knowledge Engineering*, 11(03), pp. 207-230, 2001. DOI: 10.1142/S0218194001000517

[5] G. Fortino, W. Russo, and E. Zimeo, "A statecharts-based software development process for mobile agents", *Information and Software Technology*, 46(13), pp. 907-921, 2004. DOI: 10.1016/j.infsof.2004.04.005

[6] G. Fortino, W. Russo, "ELDAMeth: An agent-oriented methodology for simulation-based prototyping of distributed agent systems", *Information and Software Technology*, 54(6), pp. 608–624, 2012. DOI: 10.1016/j.infsof.2011.08.006

[7] Engineering Agent-Based Social Simulations, CfP for a JASSS Special Issue, http://www.cs.nott.ac.uk/~pos/docs/pos-CfP-JASSS_EngineeringABSS.pdf [last accessed 12/04/2014]

[8] B. Hugues, "UML for ABM", *Journal of Artificial Societies and Social Simulation*, 15(1) 9, 2012.

[9] G. Booch, R. A. Maksimchuk, M. W. Engle, B.J. Young, J. Conallen, and K.A. Houston, *Object-Oriented Analysis and Design with Applications*, 3rd Edition, Pearson Education, 2007.

[10] E. Freeman, E. Robson, B. Bates, and K. Sierra, *Head First Design Patterns*, O'Reilly Media, Inc., 2004.

[11] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software*, Pearson Education, 1994.

[12] K. Barclay and J. Savage, *Object-oriented Design with UML and Java*, Butterworth-Heinemann, 2003.

[13] P.-O. Siebers, and B. S. Onggo, "Graphical representation of agent-based models in Operational Research and Management Science using UML", in *Proceedings of the 7th OR Society Simulation Workshop (SW14)*, 1-2 Apr 2014, Worcestershire, UK.

[14] J. Bradshaw (ed.), *Software Agents*, MIT Press, 1997.

[15] J. Ferber, Multi-Agent Systems: *An Introduction to Distributed Artificial Intelligence*, Vol. 1, Addison-Wesley, 1999.

[16] C. J. E. Castle and A. T. Crooks, "Principles and concepts of agent-based modelling for developing geospatial simulations", *Working Paper 110*, London: University College London, Centre for Advanced Spatial Analysis, 2006.

[17] M. Buckland, *Programming Game AI by Example*, Jones & Bartlett Learning, 2005.

[18] F. Wagner, R. Schmuki, T. Wagner, and P. Wolstenholme, *Modeling Software with Finite State Machines: A Practical Approach*. CRC Press, 2006.

[19] Wikipedia, "File:Turnstile state machine colored.svg.", http://en.wikipedia.org/wiki/File:Turnstile_state_machine_colored.svg [last accessed 12/04/2014]

[20] J. D. Farmer and D. Foley, "The economy needs agent-based modelling", *Nature*, 460(7256) pp. 685-686, 2009. DOI: 10.1038/460685a

[21] J. Ricardo, E. B. Passos, W. Esteban, C. Bruno and C. Pedro, "Dynamic game object component system for mutable behaviour characters", in *Proceedings of the VII Brazilian Symposium on Computer Games and Digital Entertainment*, 10-12 Nov 2008, Brazil.

[22] R. C. Martin, "Java and C++; A critical comparison", *ObjectMentor*, 1997.

[23] A. Borshchev, *The Big Book of Simulation Modeling: Multi-Method Modeling with AnyLogic 6*. AnyLogic North America, 2013.

[24] XJ Technologies, http://www.anylogic.com/ [last accessed 12/04/2014]

[25] P.-O. Siebers and U. Aickelin, "A first approach on modelling staff proactiveness in retail simulation models", *Journal of Artificial Societies and Social Simulation*, 14(2) 2, 2011.

[26] T. Zhang, P.-O. Siebers, and U. Aickelin, "Modelling Electricity Consumption in Office Buildings: An Agent Based Approach", *Energy and Buildings*, 43(10), pp. 2882-2892, 2011. DOI: 10.1145/2422531.2422535

[27] T. Zhang, P.-O. Siebers, and U. Aickelin, "Modelling the effects of user learning on forced innovation diffusion", in *Proceedings of the UK OR Society Simulation Workshop 2012 (SW12)*, 26-28 Mar 2012, Worcestershire, UK.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1461–1466
DOI: 10.15439/2014F81
ACSIS, Vol. 2

Improving the Social Capital of Trust-based Competitive Multi-Agent Systems by Introducing Meritocracy

Antonello Comi, Lidia Fotia and Domenico Rosaci
University of Reggio Calabria, DIIES Department
{antonello.comi,lidia.fotia,domenico.rosaci}@unirc.it

***Abstract*—A main issue in competitive Multi-Agent Systems is that of allowing self-interested agents to mutually cooperate in those cases where the local agent resources are not sufficient to satisfy user requests. To this end, it is necessary to introduce an internal organization for allowing agents to form suitable friendships and groups facilitating the collaboration. In the past, several approaches have been proposed aiming at forming agent coalitions which maximize the *profit* of the group or the individual agent. However, this viewpoint could introduce some negative side-effects, namely (*i*) it can lead to reward the most *aggressive* agents, also if they have bad social behaviours or (*ii*) it could introduce a sort of social flattening, without taking into account the differences among the agents in terms of merit. To face this issue, in this paper we propose an algorithm for forming friendships and groups which, instead of maximizing individual or global profit, tries to optimize a social capital represented by the mutual trust relationships.**

We theoretically prove that the application of our algorithm leads the competition to reward those agents exhibiting the most *virtuous* behaviours, introducing *meritocracy* in the system, not only rewarding effective agent performances but also encouraging correct behaviours.

I. Introduction

A COMPETITIVE Multi-Agent System (MAS) can be viewed as a virtual community composed of self-interested agents that often need to interact with each other to accomplish complex tasks. Although the introduction of cooperative behaviours seem necessary for obtaining better results, it is not possible for agents to assume that the others will cooperate sincerely, since agents are not interested in global outcome but only their own [4], [1]. For this reason, maybe some of these interactions will be done by some agent with the intention of deceiving its interlocutor, which is also a competitor. Therefore, in these systems, an agent has the problem of accurately selecting its partner, since both incomes and outcomes depend on this choice, and trust-based approaches are a possible solution to face this problem.

The issue of defining and measuring the trust between members of a community has been actively faced in the Social Sciences research field, and a common belief of most of the existing studies is that the concept of trust is strictly linked to the concept of *social capital* [15]. Social capital is a broad concept referring to the collective value associated with a community of individuals, and has been defined as "the density of interactions that is beneficial to the members of a community" [20].

Social capital is widely acknowledged as the most important asset of any social community and, more in general, of any organization. Therefore, the task of measuring trust is of the utmost relevance to understand the benefits a user can draw from her/his membership to a particular social network.

In our scenario, we suppose that the cooperation among agents can be improved by introducing a social organization in the agent community. More in particular, such an organization is realized by allowing that an agent can have some *friends*, collected in a friendship set, and moreover he can join with some *groups*, with the advantage that a collaboration requested to a friend or a member of a common group is obtained for free.

In this scenario the main question is: which is the best organization for the agent community? In our vision, we assume that the best organization is that maximizing the social capital, that we will define as the average advantage, in terms of trust, that the whole community obtains from that organization.

In the past, several approaches [9], [5], [19] have been proposed aiming at forming agent coalitions which maximize the utility of the group or the individual agent. In other words, these approaches have the purpose of optimizing some measure of *profit*. However, this viewpoint could introduce some negative side-effects. In particular, if the approach is based on maximizing the local profit, it can lead to reward the most *aggressive* agents, also if they have bad social behaviours (e.g. misleading or fraudolent agents). Moreover, if the approach tries to optimize a global profit, it could introduce a sort of social flattening, without taking into account the differences among the agents in terms of merit.

To face this issue, in our approach we propose an algorithm for forming friendships and groups which, instead of maximizing individual or global profit, tries to optimize a social capital represented by the mutual trust relationships.

In particular, we will provide two main contributions. First, we theoretically prove that our algorithm leads to continuously increase the social capital of the community. Then, enough interestingly, we also prove that the application of our algo-

This work has been partially supported by the TENACE PRIN Project (n. 20103P34XC) funded by the Italian Ministry of Education, University and Research.

rithm leads the competition to reward those agents exhibiting the most *virtuous* behaviours, i.e. those that the community perceives as the most competent and honest. In other words, our approach introduces meritocracy in the system, not only rewarding effective agent performances but also encouraging correct behaviours.

The plan of the paper is as follows. In Section II we describe a reference scenario for competitive agents. Then, in Section III we introduce our trust model while Section IV present the FGF algorithm we propose for forming friendships and groups. Next, in Section V we provide two important results related to FGF algorithm. In Section VI we discuss some related work and, finally in Section VII we draw our conclusions and discuss our ongoing research.

II. The reference scenario

Let U be a set of users and let A be a set of agents competing for satisfying the users' requests. In particular, each user $u \in U$ can submit to an agent $a \in A$ a service request r_γ falling in the category $\gamma \in C$, where C is a set of pre-defined categories. The user u will pay a fixed price p to the agent a, after the service has been provided.

Each agent a is provided with an expertise $e_a(\gamma)$ for each category $\gamma \in C$, where the expertise represents the agent's capability to correctly providing a service falling in γ. The expertise $e_a(\gamma)$ is a value ranging in the interval $[0..1]$, where $e_a(\gamma) = 1$ (resp. $e_a(\gamma) = 0$) means that a provides services falling in γ with the maximum (resp. minimum) quality of service. When the service has been provided, the user u returns a feedback f to a, where f is a value falling in $[0..1]$, such that $f = 0$ (resp. $f = 1$) represents the minimum (resp. maximum) satisfaction perceived by u for the service provided by a.

Each agent a has a set of *friend agents*, denoted by F_a, where $F_a \subseteq A$. Moreover, some groups of agents can be formed in this scenario. Let G be the set of these groups, where each group $g \in G$ is a set of agents, such that $g \subseteq A$. The names of agents and groups are registered in a Directory Facilitator (DF), as it is usual in multi-agent systems. The DF also provides a mapping $agents(g)$ receiving as input a group $g \in G$ and returning the set of the agents which are members of g.

In order to satisfy a service request concerning the category γ, the agent a can request the *contribution* of another user b, that can accept or refuse the contribution request. If b accepts, a price cp must be paid by a to b after the contribution has been provided by b. However, if b is a friend of a or a and b are in the same group, i.e. $b \in F_a \bigvee \exists g \in G : a, b \in g$, then b will accept the request and will provide the contribution for free.

In our scenario, we suppose that an agent can request a maximum number of $cMax$ contributions.

Moreover, in order to select those agents that are the best candidates for requesting contribution, the agent a can require the opinion of the other agents. In particular, a can send to an agent c a request for obtaining a recommendation $rec_b(\gamma)$ about the expertise $e_b(\gamma)$ of b in a given category γ. In other words, $rec_b(\gamma)$ is an estimation of $e_b(\gamma)$ provided by c. The recommendation request can be accepted or refused by c, and in the case of accepting, a price rp is paid by a to c. However, if c is a friend of a or a and c are in the same group, i.e. $c \in F_a \bigvee \exists g \in G : a, c \in g$, then c will accept the recommendation request and will provide the recommendation for free.

In our scenario, we suppose that an agent can request a maximum number of $rMax$ recommendations.

III. Trust measures: reliability, reputation and honesty

We assume that in the scenario described above it is possible that misbehaviours are adopted by some agent. In particular, an agent b that has been requested for a contribution by a could provide a service with a quality lesser than that corresponding to his actual expertise. Analogously, an agent c that has been requested for a recommendation by a about the agent b could indicate a value for the expertise of b different from the value that c really knows.

In order to manage these possible misbehaviours, the agent a associates three trust values to each agent b with which he interacted in the past. The first value, denoted by $rel(b, \gamma)$ is called *reliability* of b with respect to the category γ, and represents how much a trusts in the capability of obtaining good contribution from b in service falling in category γ.

When a receives from a user a feedback f for a service, he computes for each agent b which provided a contribution for that service, the component of the feedback associated with b. This component f_b will be obtained by considering the feedback f and possible additional information i_b that a has about the actual contribution provided by b. Formally, we define a mapping h such that $f_b = h(f, i_b)$. For example, assume that the service provided by a is represented by an integer value p. The user u when receiving p as the response to his request, will provide the feedback f as the percentage error with respect to the correct answer r. Then, in this case, f_b can be computed as the percentage error associated with the contribution r_b provided by b with respect to the correct answer r. In the case we do not have any additional information i_b, we assume that each feedback component $f_b = f$. In other words, in this case the agent a entirely transfers the responsibility of his feedbacks to his contributors. The reliability $rel(b, \gamma)$ is defined as the arithmetical mean of all the feedback components associated with b, related to feedbacks that a received from users for services falling in the category γ.

The second trust value, denoted by $hon(b)$, is called *honesty* of b when providing a recommendation. This value is computed by considering, for all the l feedback components f_c^1,..,f_c^l that a received from users and that are related to services which have been provided by a contributor c recommended by b, the percentage difference between f_c^l and the recommendation r_l provided by b and related to c (i.e. we compute $diff_l = \frac{|r_l - f_b^l|}{r_l}$). The value $hon(b)$ is defined as the arithmetical mean of all these $diff_l$ values.

Finally, the third value, denoted by $rep(b, \gamma)$, is called *reputation* of b with respect to the category γ, and represents how much the agents interrogated by a estimate the capability of b in the category γ. The reputation $rep(b, \gamma)$ is obtained as the weighted mean of all the recommendations that other agents provided to a about b in the category γ, where the weight of each recommendation provided by an agent c is considered equal to $hon(c)$.

For all the agents b that a did not contact in the past for requesting contributions in category γ, $rel(b, \gamma)$ is set equal to $crel$. Analogously, for all the agents b that a did not contact in the past for which a did not receive recommendations in category γ, $rep(b, \gamma)$ is set equal to $crep$. Finally, for all the agents b that a did not contact for requesting recommendations in the past, $hon(b)$ is set to $chon$. The values $crel$, $crep$ and $chon$ are cold start values associated with the agent a.

As a synthetic measure of trust for an agent b in the category γ, a considers the measure $trust(b, \gamma)$=$W_{rel} \cdot rel(b, \gamma) + (1 - W_{rel}) \cdot rep(b, \gamma)$ where W_{rel} is a weight ranging in [0..1] representing the importance that a assigns to the reliability with respect to the reputation.

IV. The friendship and group formation (FGF) algorithm

As described in Section II, for an agent a is preferable to ask for a contribution or a recommendation a friend or a member of one of his groups, since in this case he will not pay any cost. The difference between a friend fr and a group member m, from the viewpoint of a, is that fr has a direct connection with a, in the sense that a and fr mutually accepted in the past to become friends. Instead, m could not be a friend of a and thus he is available to give contributions and recommendations to a for free due to the fact they belong to the same group.

At each instant of time and for each category γ, the agent a can compute a set of *γ-preferable contributors* PC_a^γ, containing the agents b with which a interacted in the past for obtaining a contribution related to the category γ, and having (*i*) the highest $cMax$ trust values $trust(b, \gamma)$ and (*ii*) a trust value greater than a fixed threshold tc. The agents belonging to PC_a^γ are those that a would prefer to contact to have a contribution. Analogously, the agent a can compute a set of *preferable recommenders* PR_a, containing the agents b with which a interacted in the past for obtaining a recommendation, and having (*i*) the highest $rMax$ honesty values $hon(b)$ and an honesty value greater than a fixed threshold th. The agents belonging to PR_a are those that a would prefer to contact to have a recommendations.

In order to obtain the maximum performances, a would desire to have in his set of friends F_a, or in some group $g \in G_a$ with which he is joined, only the agents belonging to the sets of γ-preferable contributors PC_a^γ (for all the categories γ) and the set of preferable recommenders PR_a. In other words, he would desire to achieve the following goal:

$$\bigcup_{\gamma \in C} PC_a^\gamma \bigcup PR_a = F_a \bigcup_{g \in G_a} g \tag{1}$$

However, it is possible that (*i*) some agents that do not belong to $\bigcup_{\gamma \in C} PC_a^\gamma \bigcup PR_a$ there exist in $F_a \bigcup_{g \in G_a} g$ and (ii) some agents that belong to $\bigcup_{\gamma \in C} PC_a^\gamma \bigcup PR_a$ are not present in $F_a \bigcup_{g \in G_a} g$. Both these situations imply a disadvantage for a, that must be available to satisfy for free possible requests coming from the agents that belongs to $F_a \bigcup_{g \in G_a} g$ without balancing this cost with the possibility to have in its turn the best interlocutors for obtaining contributions and/or recommendations.

In the situation (i) the disadvantage for a can be represented by the percentage of the agents $b \in F_a \bigcup_{g \in G_a} g - \bigcup_{\gamma \in C} PC_a^\gamma \bigcup PR_a$ (w.r.t. the number of agents present in $F_a \bigcup_{g \in G_a} g$), since each of these agents will not never contacted for requesting help while they can contact a obtaining help for free.

In the situation (ii) instead, the disadvantage for a can be represented computing, for each agent b of $\bigcup_{\gamma \in C} PC_a^\gamma$ that is not present in $F_a \bigcup_{g \in G_a} g$, the difference between $trust(b, \gamma^*)$ and $trust(alt_b, \gamma^*)$, where γ^* is the category in which b is one of the preferred agents (in case b is the preferred agents in more categories, γ^* is the category associated with the highest trust value) and alt_b is the best alternative to b among the agents of $F_a \bigcup_{g \in G_a} g$, i.e. the agent of $F_a \bigcup_{g \in G_a} g$ having the best trust value in category γ^*. Analogously, it is necessary to cmpute for each agent b of $\bigcup_{\gamma \in C} PR_a^\gamma$ that is not present in $F_a \bigcup_{g \in G_a} g$, the difference between $hon(b)$ and $hon(alt_b)$, where alt_b is the best alternative to b among the agents of $F_a \bigcup_{g \in G_a} g$. The whole disadvantage can be considered equal to the average of all these contributions.

In other word, at each time the disadvantage of a will be expressed by this formula:

$$D_a = \frac{P_1 + P_2 + P_3}{3} \tag{2}$$

where:

$$P_1 = \frac{\| F_a \bigcup_{g \in G_a} g - \bigcup_{\gamma \in C} PC_a^\gamma \bigcup PR_a \|}{\| F_a \bigcup_{g \in G_a} g \|} \tag{3}$$

$$P_2 = \frac{\sum_{b \in \bigcup_{\gamma \in C} PC_a^\gamma - F_a \bigcup_{g \in G_a} g} trust(b, \gamma^*) - trust(alt_b, \gamma^*)}{\| \bigcup_{\gamma \in C} PC_a^\gamma - F_a \bigcup_{g \in G_a \|} g \|} \tag{4}$$

$$P_3 = \frac{\sum_{b \in PR_a - F_a \bigcup_{g \in G_a} g} hon(b) - hon(alt_b)}{\| PR_a - F_a \bigcup_{g \in G_a \|} g \|} \tag{5}$$

Obviously, if the equality 1 holds, the disadvantage D_a will be equal to 0 (minimum value), while if (*i*) all the agents belonging to $F_a \bigcup_{g \in G_a} g$ are not preferred contributors or recommender, and (*ii*) all the alternative agents introduce a trust difference equal to 1, therefore the disadvantage D_a will be equal to 1 (maximum value).

The algorithm we propose is executed by the agent a for minimizing D_a during several *epochs*, such that in each epoch some preferred agents are joined to the set $F_a \bigcup_{g \in G_a} g$ replacing those agents having the worst trust or honesty values.

The period of time between two consecutive epochs is equal to a pre-fixed value T.

At each epoch, the algorithm is composed by the following two tasks, the former (called active task) dedicated to manage the requests that a sends to the other agents, the second (called *passive task*) managing the requests coming from the other agents:

A. Active task

The agent a executes this task in order to obtaining the friendship or the presence in a group of G_a of those agents belonging to $\bigcup_{\gamma \in C} PC_a^\gamma \bigcup PR_a$ but which do not yet belong to $F_a \bigcup_{g \in G_a} g$. To this end, the strategy of a is that of (*i*) first requesting the friendship of each missing agent b; (*ii*) if the missing agent b does not accept the friendship, then requesting to join with some group that contains b; (*iii*) if any group containing b does not accept the joining request, then trying to forming a new group that in the future could attact b, requesting the participation of other agents having similar necessities.

More in particular, this task is composed by the following steps:

1) The set $\bigcup_{\gamma \in C} PC_a^\gamma \bigcup PR_a$ is computed.
2) For each agent $b \in \bigcup_{\gamma \in C} PC_a^\gamma \bigcup PR_a - F_a \bigcup_{g \in G_a} g$, a request of friendship is sent to b.
3) If the friendship request is accepted by b, then b is added to F_a. Moreover, if $b \in PC_a^\gamma$, then the agent $k \in F_a$, $k \notin \bigcup_{\gamma \in C} PC_a^\gamma \bigcup PR_a$ having the worst trust value $trust(k, \gamma)$ is removed from F_a. Otherwise, if $b \in PR_a$, the agent $k \in F_a$, $k \notin \bigcup_{\gamma \in C} PC_a^\gamma \bigcup PR_a$ having the worst honesty value $hon(k)$ is removed from F_a.
4) If the friendship request is not accepted by b, then the agent a requires to the DF the set $GROUP_b$ of all the groups containing b as a member. For each group $g \in GROUP_b$, the agent a computes the disadvantage D_a^* that would derive if the group g is added to G_a, using the Formula 2. If $D_a^* < D_a$, then a sends a joining request to g.
5) If the joining request is accepted by g, then g is added to G_a. Moreover, analogously to the previous step, if $b \in PC_a^\gamma$, then the agent $k \in F_a$, $k \notin \bigcup_{\gamma \in C} PC_a^\gamma \bigcup PR_a$ having the worst trust value $trust(k, \gamma)$ is removed from F_a. Otherwise, if $b \in PR_a$, the agent $k \in F_a$, $k \notin \bigcup_{\gamma \in C} PC_a^\gamma \bigcup PR_a$ having the worst honesty value $hon(k)$ is removed from F_a.
6) If the joining request is not accepted by g, then a sends a *call for a new group* to all the agents belonging to $F_a \bigcup_{g \in G_a} g$.
7) When some agent affirmatively responses to the call for a new group, the new group is formed and registered to the DF.

B. Passive task

The purpose of this task is that of managing the requests of friendships coming from other agents, as well as the requests of joining that other agents send to groups with which a is joined. In particular, this task is composed by the following steps:

1) When a friendship request coming from an agent b arrives to a, then a will accept it if the insertion of b in the set F_a (with the consequent removing of an agent following the rules described in step 3 of the active task) implies a decrement of the disadvantage D_a. Otherwise, the request will be refused.
2) When a joining request coming from an agent b arrives to the administrator of a group g with which the agent a is joined, a voting is requested from the administrator to all the agents belonging to the group g. Each vote can be positive or negative. If the the majority of the votes is positive, then the joining request of b is accepted, otherwise it is refused. The agent a will give a positive vote if the insertion of b in the set F_a (with the consequent removing of an agent following the rules described in step 3 of the active task) implies a decrement of the disadvantage D_a. Otherwise, the vote of a will be negative.
3) If a call for a new group arrives from an agent b, then a accepts the proposal to join with the new group if the addition of b to $F_a \bigcup_{g \in G_a} g$ does not increase the disadvantage D_a.

V. Theoretical results

We define *Social Capital* SC of the whole multi-agent system the mean value of all the contributions $(1 - D_a)$ given by each agent a. In other words, the social capital represents the *average advantage* associated to a given internal organization of the MAS in friendships and groups.

Definition 1: The Social Capital of a MAS is defined as:

$$SC = \frac{\sum_{a \in A}(1 - D_a)}{\|A\|} \tag{6}$$

In this section, we provide two important results related to the FGF algorithm. First, we prove that at each iteration of the FGF, the social capital SC increases. In other words, we show that FGF achieves the purpose of creating relationships among agents tending at optimizing the global social utility.

Theorem 1: The social capital SC increases at each new iteration of the FGF algorithm.

Proof: At each iteration, each agent a performs only actions that can either (i) increase each D_a, in the case some preferred contributor or recommender accept his joining request or (ii) maintain unvaried D_a, if any agent does not accept hos joining request. However, D_a can decrease due to actions performed by other agents, as the cancellation of a preferred contributor or recommender from F_a or from a group of G_a. However, let b a preferred contributor in a category γ^* or a recommender that cancels himself from F_a or from a group of G_a. Then, the disadvantage D_a will increase for the necessity to replace b with the best alternative alt_b, and the increment will be equal to $trust(b, \gamma^+) - trust(alt_b, \gamma^*)$ (or $hon(b) - hon(alt_b)$ in the case of a recommender), that is lesser than 1. However, b cancels himself from F_a of

from a group of G_a only if a is not one of his preferred contributors or recommenders, thus the cancellation implies a decrement of D_b equal to 1. Overall, the sum of D_a and D_b will decrement of 1-$(trust(b,\gamma^*) - trust(alt_b,\gamma^*))$ (or $1 - (hon(b) - hon(alt_b))$. From this observation, we directly derive that the sum of all the agent disadvantages decreases at each iteration, and consequently the sum of all the contributions $(1 - D - a)$ increases. This proves the theorem. ■

In order to characterizing the global trustworthiness that an agent a receives from the whole community in a given category γ, we define the notion of *merit* μ_a^γ, as follows:

Definition 2: The merit μ_a^γ of an agent a in the category γ is defined as the number of agents that consider a as a preferred contributor or a preferred recommender.

Moreover, we define also the notion of *expected gain* of an agent a, for characterizing how much the agent a can expect to increment his bank amount at a given step of the competition. Denoting as $P_a(i)$ the probability distribution of the bank amount increment for a, i.e. the probability that the bank amount increment is equal to i, we have that:

Definition 3: The expected gain δ_a of the agent a is defined as the expected value of the probability distribution $P_a(i)$

Finally, we consider valid the following *mirror assumption*:

Assumption 1: Let a, b be two agents, such that at a given iteration $\mu_a^\gamma < \mu_b^\gamma$. In this situation the number u_a of users contacting a for a service request in the category γ will be lesser than the number u_b of users contacting b.

This assumption appears reasonable considering that if $\delta_a^\gamma < \delta_b^\gamma$, then the global satisfaction of the agent community for the performances of a is lesser than the satisfaction for the performances of b. Since the global satisfaction of the agent community is constructed based on the feedbacks received by the users, it is reasonable to think that in this situation also the users will prefer to contact a instead of contacting b. In other words, this assumption means that the users' choices specularly reflect the agents' choices. This will be particularly true if the trustworthiness of a, represented by the number of agents that consider a as a preferred interlocutor, actually capture the expertise of a. Obviously, the more the trust models of the agents are built strictly based on the users' feedbacks, similarly to the case of the trust model presented in Section III, the more the mirror assumption can be considered as valid. Furthermore, the more the adopted trust model is able to capture the actual expertises of the agents, the more the assumption above will reflect the real situation.

Theorem 2: At each iteration, for each pair of agents a, b such that $\mu_a^\gamma < \mu_b^\gamma$, the expected gain δ_a will be lesser than the expected gain δ_b.

Proof: Let a, b be two agents, such that at a given iteration $\mu_a^\gamma < \mu_b^\gamma$. Supposing as valid the mirror assumption, the number u_a of users contacting a for a service request related to γ will be lesser than the number u_b of users contacting b. Moreover, in this situation, the probability PI_a that a is contacted by other agents for a contribution or a recommendation related to γ is lesser than the corresponding probability PB_a, and therefore the expected number of the agents i_a contacting a for a contribution or for a recommendation related to γ is lesser than the expected number i_b of those contacting b. In the same way, the expected number of agents o_a contacted by a in the category γ is greater than the expected number of agents o_b contacted by b, due to the high probability that the expertise of a is smaller than that of b. Thus, also supposing for simplicity that both the price of a contribution or a recommendation is equal to p^*, the expected gain δ_a at the end of the current iteration is $u_a \cdot p + i_a \cdot p^* - o_a \cdot p^*$, that is smaller than the corresponding gain δ_b for the agent b. ■

VI. Related Work

This section discusses some previous work related to the issues of partner selection and collaboration among self-interested agents. Partner selection plays an important role in filling the deficit of distributed agents. Research on partner selection generally proposes various types of evaluation metrics for selecting appropriate partners. Local decision with local modeling [12], [16] takes into account the local model about potential partners and finds the most appropriate partners, for example, according to trustworthiness, reputation or quality of provided services. These models are constructed either by direct observation or communication with other agents. In social control research, agents need to evaluate other agents or the services provided by other agents in order to realize a distributed but secure control over the interactions among agents [8].

The negotiation-based approach involves explicit peer-to-peer communication for negotiation. The Contract Net Protocol (CNP) [21], [6] provides a mechanism for finding the best partners who provide necessary services at the least cost. The Adaptive Decision Making Framework (ADMF) [2] also deploys a negotiation-based partner selection scheme. Recently, the use of trust in competitive agent systems has been widely emphasized [17], [11]. In this context, trust measures have been exploited for forming clusters of agents [10], [3] and for generating recommendations in social network contexts [7]. The problem of detecting group of actors in a competitive social community based on trust has particularly been faced in [13], [18], [14]. None of the aforementioned approaches faces the issue of improving the social capital of the agent community by introducing meritocracy. Instead, those approaches try to use trust measures for recommending to an agent the best agents to contact as promising interlocutors, without the purpose of introducing a social advantage for the whole community. Instead, our approach is capable of achieving such an advantage, also realizing it through a meritocratic approach, that encourages the social actors to assume correct behaviours for increasing their social reputation.

VII. Conclusion

The problem of introducing a convenient organization into a community of self-interested agent is central in the context of allowing agents to collaborate for increasing the individual capability of providing services to users. If in the past the

most of the proposed approaches face such a problem trying to maximize the profits of the single actors or the whole community, however it is important to remark that these kinds of proposals lead to important negative effects as, for instance: (*i*) encouraging deceptive or fraudolent behaviours in the case the goal of the approach is that of rewarding the most aggressive agents or (*ii*) introducing a social flattening that does not take into account the different merits of individual agents, in the case the approach is focused on the optimization of the profit associated with the entire community. In this work, we argue that a possibility to avoid the above problems is that of using as objective function to maximize a measure of social capital, depending on the trust relationships existing among agents. In other words, in our vision the social capital of the community is represented by the strength of these trust relationships in the friendship lists and in the groups formed into the community. As a consequence, in this vision, those agents that are perceived as the most trustworthy are rewarded by our approach, thus introducing a form of meritocracy in the community. Our paper gives the theoretical demonstration that, at each iteration of the FGF algorithm (*i*) the social capital SC increases and (*ii*) for each pair of agents a, b such that $\mu_a^\gamma < \mu_b^\gamma$, the expected gain δ_a will be lesser than the expected gain δ_b. This second result has been proved under the *mirror assumption*, valid in the case the behaviour of the users when selecting the agents reflects the behaviour of the agents when selecting their collaborators.

Our ongoing research is now devoted to test the FGF algorithm in complex situations, characterized by large sets of agents, and in real cases when agents operate on the behalf of human users. In particular, we are planning to develop an application in the domain of grid/cloud services, that appears as a promising possibility of profitably using our approach for introducing a convenient organization in the service system.

REFERENCES

[1] Agent uno: Winner in the 2nd Spanish ART Competition, author=Muoz, V and Murillo, J, journal=Inteligencia Artificial, volume=12, pages=19–27, year=2008.

[2] K Suzanne Barber and Cheryl E Martin. Dynamic reorganization of decision-making groups. In *Proceedings of the fifth international conference on Autonomous agents*, pages 513–520. ACM, 2001.

[3] Francesco Buccafurri, Antonello Comi, Gianluca Lax, and Domenico Rosaci. A trust-based approach to clustering agents on the basis of their expertise. In *Proceedings of Advances in Agent and Multi-Agent Systems 2014. AMSTA 2014*, 2014.

[4] J Carbo, G Muller, M Gomez, and J Sabater-Mir. Improving the art-testbed, thoughts and reflections. In Proceedings of the Workshop on Competitive agents in "Agent Reputation and Trust Testbed", Salamanca, Spain, pages 1–15, 2008.

[5] Viet Dung Dang and Nicholas R Jennings. Generating coalition structures with finite bound from the optimal guarantees. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 564–571. IEEE Computer Society, 2004.

[6] Randall Davis and Reid G Smith. Negotiation as a metaphor for distributed problem solving. *Artificial intelligence*, 20(1):63–109, 1983.

[7] Pasquale De Meo, Antonino De Meo, Domenico Rosaci, and Domenico Ursino. Recommendation of reliable users, social networks and high-quality resources in a social internetworking system. *AI Communications*, 24(1):29–50, 2011.

[8] Keith Decker, Katia Sycara, and Mike Williamson. Middle-agents for the internet. In *IJCAI (1)*, pages 578–583, 1997.

[9] Partha Sarathi Dutta and Sandip Sen. Forming stable partnerships. *Cognitive Systems Research*, 4(3):211–221, 2003.

[10] Salvatore Garruzzo and Domenico Rosaci. Agent clustering based on semantic negotiation. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 3(2), 2008.

[11] Salvatore Garruzzo, Domenico Rosaci, and Giuseppe M.L. Sarné. Integrating trust measures in multi-agent systems. *International Journal of Intelligent Systems (IJIS)*, 27(1):1–15, 2012.

[12] E Michael Maximilien and Munindar P Singh. Agent-based trust model involving multiple qualities. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 519–526. ACM, 2005.

[13] Fabrizio Messina, Giuseppe Pappalardo, Domenico Rosaci, Corrado Santoro, and Giuseppe ML Sarné. A distributed agent-based approach for supporting group formation in p2p e-learning. In *Proceedings of the thirteenth International Conference on Advances in Artificial Intelligence. AI*IA 2013*, 2013.

[14] Fabrizio Messina, Giuseppe Pappalardo, Domenico Rosaci, Corrado Santoro, and Giuseppe ML Sarné. Hyson: A distributed agent-based protocol for group formation in online social networks. In *Proceedings of the 7th International Workshop on Multi-Agent Systems and Simulation (MATES 2013)*, 2013.

[15] M. Paldam. Social capital: one or many? Definition and measurement. *Journal of Economic Surveys*, 14(5):629–653, 2000.

[16] Jisun Park and K Suzanne Barber. Information quality assurance by lazy exploration of information source combinations space in open multi-agent systems. *J. UCS*, 11(1):193–209, 2005.

[17] Domenico Rosaci. Trust measures for competitive agents. *Knowledge-based Systems (KBS)*, 28(46):38–46, 2012.

[18] Domenico Rosaci and Giuseppe ML Sarné. Matching users with groups in social networks. In *Proceedings of the International Conference on Distributed Computing (IDC 2013)*, 2013.

[19] Tuomas Sandholm, Kate Larson, Martin Andersson, Onn Shehory, and Fernando Tohmé. Anytime coalition structure generation with worst case guarantees. *arXiv preprint cs/9810005*, 1998.

[20] W. Sherchan, S. Nepal, and C. Paris. A survey of trust in social networks. *ACM Computing Surveys*, 45(4):47, 2013.

[21] Reid G Smith and Randall Davis. Frameworks for cooperation in distributed problem solving. *Systems, Man and Cybernetics, IEEE Transactions on*, 11(1):61–70, 1981.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1467–1476
DOI: 10.15439/2014F298
ACSIS, Vol. 2

Common and Domain-specific Metamodel Elements for Problem Description in Simulation Problems

Patrizia Ribino[‡], Valeria Seidita[§‡], Carmelo Lodato[‡], Salvatore Lopes[‡] and Massimo Cossentino[‡]
[‡]Istituto di Reti e Calcolo ad Alte Prestazioni
Consiglio Nazionale delle Ricerche
Palermo, Italy
Email: {ribino,c.lodato,lopes,cossentino}@pa.icar.cnr.it
[§]Dip. di Ingegneria Chimica Gestionale Informatica Meccanica, University of Palermo
Palermo, Italy
Email: {valeria.seidita}@unipa.it

***Abstract*—It is well known that the multi-agent system paradigm is well suited for modelling and developing simulations of complex systems belonging to several application domains. Simulation study aims at developing simulation models useful for representing, studying and analyzing entities and their behavior in a system according to specific purposes. With our work we are trying to understand what are the right elements to be considered and included in the description of a simulation problem. In order to root our resulting metamodel in the state of the art of multi-agent simulations we started from the study of twelve papers dealing with four different application domains: Crowd Dynamics, Traffic and Transportation, Electricity Power Engineering and Supply Chain and Logistic. From this study we obtained a metamodel that may be used by an analyst as a guideline and concept repository for facing a new system design. The metamodel is the result of a well defined approach that is described together with the obtained results consisting in one core metamodel containing elements that are common to all the four application domains and some domain extension contents. These latter contain the elements that are specific of each of the studied domains and are not present in the others.**

I. Introduction

MULTIAGENT system paradigm is well suited for modeling and developing simulations of complex systems belonging to several application domains. As Klügl says in [1], multiagent simulation is considered a killer application of agent-based technology. But an engineering approach for developing simulation models is still lacking and several issues are still open. Moreover the author says "*a categorization of models with respect to relevant design concepts would be a good starting point advancing the knowledge on what is relevant for multiagent simulations* "[1]. This future research is suggested to address the challenge about: *which design concepts are relevant for which type of model?* Due to the wide variety of domains in which multiagent simulations are employed, our question is: *which concepts are relevant for which type of agent based simulation problem?* Answering to this question, in our opinion, helps to find a solution to the proposed challenge.

The main aim of our research is to develop a complete methodological approach for conducting simulation studies. Our approach will be composed of: *(i)* the phase devoting to describe the features of the simulation problem, *(ii)* the phase for verifying if the simulation model adheres to the real system and then *(iii)* the phase for defining experiments and implementing them in the simulation framework. The development of a specific MAS resides inside the whole simulation methodology.

In this paper we focus on the first part of our project, the one related to the first activities of a simulation study. Simulation study aims at developing simulation models useful for representing, studying and analyzing entities and their behavior in a system following specific purposes. Activities in simulation studies are mainly performed by the analyst and the model designer that have to create a model of the world starting from the description of the problem. But what are the right tools for them to build the simulation model that best fits their needs? And above all, agent simulation is used in several application domain presenting very different features that affect the creation of simulation models, so how may the model designer identify the elements helping him to construct the simulation model?

This starting point is very demanding, several authors in literature agree on that and propose different solutions. In this work we propose a preliminary work for solving this problem by focusing on the problem statement, our aim is to create a metamodel of all the elements and their relationships that we have to find in the problem description of a simulation study. As we said, several different simulation studies greatly differ for the application domain they refer to, so in order to be as much general as possible in the creation of the metamodel, we analyzed and studied papers by four different application domains (Crowd Dynamics, Traffic and Transportation, Electricity Power Engineering and Supply Chain and Logistic) in order to retrieve common and specific elements.

In other words, the contribution of this paper is in the identification of what elements are commonly used for describing some simulation problems and, moreover, what elements are used in specific application domains. From this study we obtained a metamodel that may be used by an analyst as a guideline and concept repository while describing a new problem. We do not claim that all situations may be faced

978-83-60810-58-3/$25.00

(or have to be faced) with the proposed metamodel but we think that if the elements we list in that have been frequently used in papers dealing with similar problems, it is likely that such elements may be useful again in the same context or in a similar one.

The rest of the paper is organized as follows: in section II we detail the motivations for our work against the related works in literature, in section III we show the process for bundling the metamodel and then in sections IV and V some discussions and conclusions are drawn.

II. Motivation and Related Works

Simulation is the discipline for designing a model of actual or physical systems. It abstractly represents a real system involving the elaboration of models where system behavior is reproduced following a set of hypothesis used to define different scenarios. Multiagent simulation focuses on the study and description of distributed behavior in a dynamic context [2] and consists in identifying simulation models where entities (such as agents), their behaviors, their interactions among themselves or with the environment in which they are situated, are described. Hence simulation studies prescribe to work with models of the problem and not with the problem itself.

Simulation results are used in place of experimentation over the actual or real system. If the model were not the closest approximation of the actual system then it might lead to erroneous considerations, faults and costly decisions.

The question we want to answer is: which are the activities to made in order to build adequate simulation models?

In [3] and [4] the life cycle of a simulation study is illustrated, it prescribes ten phases organized in a quite iterative way. Briefly, the first step is *communicating the problem* and *formulating it*, then defining the system objectives, creating the conceptual model and finally designing experiments. Defining and setting the objectives of the study is a very important step that may lead to the identification and the definition of the right simulation model to use for investigating a specific problem and analyzing its results.

In [5] the author investigates how to build a valid and credible simulation model. He does not describe a real and complete methodological approach, rather he presents a set of techniques for building a valid model. Law's approach starts with the identification of a set of steps to do for formulating the problem, these steps sound like advises more then techniques, indeed he say "Problem of interest is stated by the decision-maker. It may not be stated precisely or in quantitative terms. An iterative process is often necessary" and still "a kickoff meeting is necessary for discussing the overall objectives of the study, the scope of the model, the performance measures, etc.". Nothing is said, in Balci's, Nance's and Law's works, about *how* to perform these activities in order to retrieve the most useful elements for creating simulation models starting from problem description.

It is clear that the description of the problem, or of the domain under study, greatly affects the production of simulation models. The risk of influencing the result of simulation study is high. The model developer and then the decision maker have to be sure to work with problem domain description where useful elements are described. From a literature review and above all from [6] we can highlight the following elements: entities of the domain exposing behavior, the objective of the simulation, the parameters to be tuned for evaluating results, the interactions among entities, possible rules or constraints for interacting, resources and services from which entities take data for implementing their behavior.

Balci and Nance in [7] presented an high level procedure for guiding the analyst during problem formulation; he identified the need of formulating the problem in objective terms and distinguished between problems requiring prescriptive or descriptive solutions. The procedure he proposed is very detailed and is followed by a formulated problem verification activity and the measurement of the formulated problem. The overall approach may be resumed in fifteen very detailed tasks that, we may say, influenced all the successive work in problem domain description and from which the previous said elements may also be abstracted.

In the agent oriented context Garro et al. [8] propose a methodology for guiding domain experts from analysis to modeling. Our approach mainly focuses on the problem formulation and, starting from the same considerations of Balci and Nance, does not "still" propose detailed guidelines for problem statement description but instead provides the set of elements and relationships among them, a metamodel, that we abstracted from several works presented in literature. This metamodel is the base for obtaining the sequence of activities to do for building the most correct problem statement for a specific problem.

Today, it is quite usual that existing software engineering design methodologies start with the assumption that the Problem Statement document is already available, it is usually presented in the form of a text document, sometimes delivered using techniques such as interviews or ethnographic approaches [9] as often as not it is presented as an unstructured document containing a free description of the problem.

Our aim is to pose the base for creating guidelines for writing problem statements as a useful input for an agent-based simulation methodology and we want to answer to the following questions:

- *How the problem statement has to be done?*
- *Which are the essential elements it has to present?*
- *How to describe the specific problem that should be solved by the model in the simulation study?*

First of all, we considered that if the Problem Statement has to serve as a basis for obtaining a simulation model it has to contain and make explicit elements such as the entities involved in the real system, their behavior and interaction with the environment but also elements representing the objectives of the simulation and a fair set of parameters settings and measures.

This is very different from a common problem statement, for instance from the one devoted to illustrate elements useful for obtaining an object model (of a object-oriented methodology).

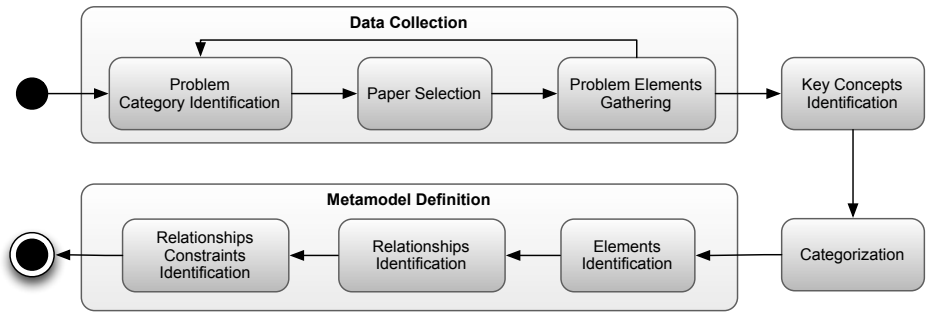


Fig. 1. Metamodel Building Process

In what does a simulation study differ from an agent based methodology? May differences be reported on the structure (or on how to build) of a problem statement? In order to answer to these questions we decided to explore and analyze the research products of several fields in which agent-based simulation is used in order to retrieve the common elements in use and to create a metamodel for supporting the creation of problem statements for agent-based simulation systems.

We analyzed literature from the fields of Financial Market, Urban Development, Traffic and Transportation, Crowd Dynamics and Logistic and Supply Chain Management in order to identify the shared elements in all these categories of problems and to conduct a generalization process that is detailed in the following section.

The reason for which we create a metamodel for representing abstract simulation study elements lies in the deep experience we have in creating and documenting design methodologies [10], [11], [12], [13]. Indeed, in the PRoDe approach [14], elements and relationships among them are used for representing dependencies among elements and establishing the right sequence of tasks/activities to do for reaching a design result. This is what we mean to do with the problem domain, once having determined the elements and their relationships that have to be present in a problem statement we might be able to determine the set of activities for instantiating them. This results, in the context of our complete building-in- progress methodology, in a set of guidelines to be used for writing the most useful problem statement for a specific need.

III. METAMODEL DEFINITION

A critical point for a simulation study is to accurately formulate the problem to be addressed in order to obtain a high-quality simulation model. Problem statement is the first activity by which the addressed problem is translated into a well formulated one that is sufficiently detailed for allowing specific modeling activities. Several studies have highlighted that agent-based simulations seem to be more suited in the field of socio-economic systems, which covers a wide range of problem domains. These domains share some common concepts but differ from some others. Hence, for answering the question *"which concepts are relevant for specific types of agent based simulation problems?*", we are leading a systematic analysis of several scientific papers in order to determine: (i) what are the domains of simulation problems commonly faced with an agent-based approach; (ii) what is the set of shared concepts useful for describing such problems and (iii) what is the set of concepts that differ among domains of agent based simulation problems.

As we previously said, we are working on the development of a methodology that covers the entire life cycle of a simulation study starting from the problem statement activity. In this paper we propose a preparatory study whose aim is building a metamodel that may be applicable in different agent based simulation studies for supporting the problem statement creation activity.

In the following we show the process we performed for building such metamodel and its preliminary results.

A. Metamodel Building Process

In order to build the metamodel for agent based simulation problems, we followed the process shown in Fig.1. The first three steps allow us to collect data and information that will be the starting point for our analysis. Then, a key concepts identification, followed by a categorization activity, enables us to make some reasoning useful for determining the metamodel. In the following we show the results of these steps.

a) Data Collection Phase: Firstly, we performed a literature review for identifying problem categories addressed by means of agent based simulation approaches. The preliminary results of this research highlighted that agent based simulations are applied to solve problems in the following domains (and not only[1]): *(i) Financial Markets* [15], [16] where simulations are used for studying the behavior of individual investors, the dynamics of markets, trading mechanism and so on; *(ii) Urban Development* [17], [18] that studies models for urban planning, city dynamics, individual residential behaviors; (iii) *Traffic and Transportation* [19], [20], [21] dealing with simulation models for transportation planning, design and operations; (iv) *Crowd Dynamics* [22], [23], [24] that studies the behaviors of individuals, groups in several critical scenarios often with the

[1]This research is not intended to be exhaustive. The results we present are only preliminary. We are doing further researches in order to obtain more accurate results.

aim of buildings design; (v) *Social Networks* that studies the evolution and dynamics of networks [25], [26]; (vi) *Logistics and Supply Chain Management* that studies processes inside different nodes of supply chain or the whole supply chain in order to find the best organizational structure for collaborative companies working together [27], [28], [29]; (vii) *Electric Power Engineering* [30], [31], [32] deals with the generation, transmission, distribution and utilization of electric power as well as the electrical devices connected to such systems.

The second step in the Data Collection phase was papers selection for each of the previously identified categories. The selection was based on two criteria: details of the simulation problem description and its relevance/impact of the paper. Finally, we analytically gathered elements from each paper. Such elements were opportunely ordered according to their semantic and functional similarity.

So far, we have examined only four problem domains: Traffic and Transportation [19], [20], [21], Crowd Dynamics [22], [23], [24], Electric Power Engineering [30], [31], [32], Logistics and Supply Chain Management [27], [28], [29]. In this preliminary work we selected 12 papers, three for each domain.

The results of this phase are sets of domain dependent concepts. All the common concepts of each domain have been grouped by analyzing the text describing each domain in the papers. Each element has been identified firstly looking at its explicit presence in sentences, the same was made for relationships, and secondly analyzing the whole domain description text in order to find unexpressed implicit knowledge.

b) Key Concepts Identification and Categorization: The key concepts identification (see Fig.1) is the conceptual activity we performed for processing sets of domain dependent concepts coming from the previous stage. Our aim was to infer a set of "higher level" abstraction categories that can be more widely applicable to generic agent based simulation problems (see Fig.2). For instance, in the domains we studied, we found concepts like room, road and city that may be represented, at an higher level, by the "spatial position" concept.

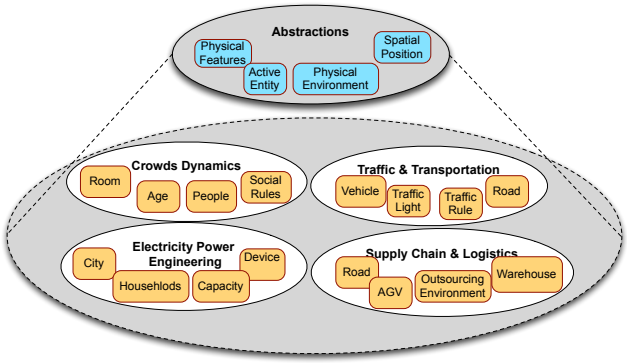


Fig. 2. Identification of higher level abstraction categories for the problem description metamodel.

The results of this activity are listed below:

1) A problem statement of an agent based simulation study describes active entities. For representing such entities, each domain uses concepts such as: (i) people, persons, pedestrian, individuals, human beings in the field of crowd dynamics; (ii) cars, drivers, buses, vehicles, people, AGVs, trucks etc... in the domain of traffic and transportation; (iii) electric cars, devices, households, electricity sellers, buyers, players etc... in the domain of power engineering; (iv) companies, suppliers, production, customers, resources,organization etc... in the domain of logistic and supply chain management.
2) Active entities may be located in a physical space. Their local position could either be a geographic position or a particular place (for example a room, a corridor, a road, in a station and so on);
3) Active entities are described by means of features and behaviors. The kinds of features we discovered in our domains are: physical, psychological and technical features. To give some examples in the domain of Crowd Dynamics a problem statement may contain sentences such as: "*Agents in a hurry will not respect others' personal space . . . More polite agents will respect lines and wait for others to move first. . .*" [23]; "*Agents are given different psychological (e.g., impatience, panic, personality attributes) and physiological traits*" [23]; "*Human individuals are different from each other by age, body dimension, mobility and personality*"[33]. Examples in Power Engineering are: "*The market operator manages the pool using a market-clearing tool to set market price and a set of accepted selling and buying bids for every negotiation period*" [30]. "*Room air conditioning Schedule time on: 15 min Schedule time off: 35 min, Power: 221 W . . . The simulated vehicles are able to move to and between different cities and recharge from time to time*"[31]. In the domain of Traffic and Transportation, typical features are described with sentences such as "*Each driver or pedestrian is assigned an agent's profile that includes height, width, velocity steering . . .* " [21]. In the Logistic and Supply Chain domain we may find sentences such as *"...forest is harvested by small-size entrepreneurs responsible for felling trees, for cross cutting them into appropriate length logs, and for hauling then to roadside"* [34].
4) The problem statement includes a description of a real world portion (physical or abstract). The represented environment may contain active entities, static and dynamic objects, groups of entities. Some examples: "*Each smart city starts with a number of households, vehicles and power stations. . . . Its area is placed at the origin of a coordinated system or as a neighbor to an existing reference city e.g. on north, south, east or west*" [31]. "*The electricity market environment typically consists of a pool that players submit their bids to, which can be symmetric or asymmetric, and a floor for bilateral*

contracts" [30]. "*Ten thousand agents that fill a corridor that is 300m long*" [22]. "*The floor plan contains a number of office spaces organized along hallways and corridors. There are two egress exits, exit A on the west and exit B on the south*" [33]. "*The static structure of a bus network is composed of four elements: itinerary, line, bus stop and bus station*" [20]. *"is a warehouse consisting of: number of Gates where the articulated lorries leave their containers waiting to be unloaded; number of Sorting Area with twelve Sorter Places (where the pallets are left in order to be addressed toward next destination)..."* [29].

5) The problem statement describes the dynamics occurring in the real world portion. These dynamics are expressed in terms of actions and interactions of active entities. They can also be constrained by some regulations. To give some examples: "*When people are crossing portals, care must be taken to avoid intersection between agents leaving and agents entering*"[23]."*High truck A is stopped in the lane closest to the sidewalk and obscures the pedestrian's view. Vehicle B approaches the crosswalk from the second lane and the view of the driver is obscured too*"[21]. "*Players negotiating on the pool must prepare a bid for the 24 periods of the spot market . . . The market operator must assure that the economical dispatch accounts for the specified conditions, which might imply removing entities that have presented competitive bids but whose complex conditions were not satisfied*" [30], *"a network of facilities is responsible for orchestrating all operations from the forest to the customers, including the operations of many sawmills" "Once dried, bundles are disassembled to be planed, cut to length, sorted, and graded according to standard rules or customer specifications"*[34]. "*In a normal situation(non-panic), people will respect lines and wait for others to walk first*" [23]. "*People movement can also be restricted due to environmental constraints imposed by the spatial geometries*" [33]. "*The length of time that elapses during an assembly depends on a specified manufacturing delay parameter. It also depends on whether the assembly is performed in parallel or sequentially*" [28].
6) The problem statement may contain information about entity groupings and roles of the entities. Examples: *"Members in a hierarchically structured group (such as families) tend to stay together and follow the leader*"[33]. "*Each household belongs to one city and it is randomly placed within the city area. Households have a number of appliances that consume and/or produce energy*" [31].
7) The problem statement defines the objectives of the simulation study. Examples taken from the examined papers are: *"Egress Analysis for building design"*[33]. *"Traffic analysis for support decision for example road safety"* [21]. *"Efficient management of electricity network"*[31]. *"The goal of both simulations was to service customer orders at or near 100%, based on the three day requirement"* [28].
8) The problem statement defines parameters to be tuned in order to reach simulation objectives. These parameters may refer to the system as a whole or to its entities. Some examples of system parameters: *different spatial distribution of the occupant* [33], *density of crowd* [22], *changing supplier or add a distribution center* [35]. Examples of entity parameters are: *distance to obstacles*[23], *vehicles velocity and distance from cross walk* [21], *inventory levels, lead time and transportation time at different locations to better understand the dynamics of the supply chain* [35].

TABLE I
KEY ELEMENTS AND THEIR MAPPING WITH THE ANALYZED DOMAINS.

			PROBLEM DOMAINS			
			Crowd Dynamics	Traffic & Trasportation	Power Engineering	Logistics & Supply Chain
PROBLEM ELEMENT	INDIVIDUAL ENTITY	Active Entity	X	X	X	X
		Physical Features	X	X	X	X
		Psycological Features	X	V	-	V
		Technical Features	-	-	X	-
		Behavior	X	X	X	X
		Behavioral rules	X	X	-	V
	REAL WORLD PORTION	Physical Environment	X	X	V	V
		Abstract Environment	-	-	V	V
		Static Object	X	X	X	X
		Dynamic Object	X	X	-	X
		Physical Features	X	X	X	X
		Spatial Position	X	X	V	V
		Structural Constraints	V	X	V	V
	SCENARIO	Rules/Constraints	X	X	X	X
		Interaction	X	X	X	X
		Actions	X	X	X	X
	SOCIAL STRUCTURE	Role	X	X	X	V
		Groups	X	V	V	V
		Social Rule	X	X	_	V
	SIMULATION PURPOSE	Objective	X	X	X	X
		Parameter	X	X	X	X

Keys: X = Total Mapping; V = Partial Mapping; - = No Mapping

After Key Concepts Identification, the Categorization of Concepts is carried out in order to find high level concepts as explained in Fig. 2. Two work products result from this activity, one describing the concepts (see Table I) and one describing the relationships discovered into the analyzed domain(see Table II).

In Table I we represent the set of high-level concepts resulting from this activity grouped according to a more general category they are related to. In particular, a total

TABLE II
AN EXCERPT OF RELATIONSHIPS DETECTED AMONG CONCEPTS.

PROBLEM ELEMENT			INDIVIDUAL ENTITY						REAL WORLD PORTION						
			Active Entity	Physical Features	Psycological Features	Technical Features	Behavior	Behavioral Rules	Physical Environment	Abstract Environment	Static Object	Dynamic Object	Physical Features	Spatial Position	Structural Constraints
	INDIVIDUAL ENTITY	Active Entity	interacts												
		Physical Features	owns												
		Psycological Features	owns												
		Technical Features	owns												
		Behavior	shows	influence	infuence	infuence									
		Behavioral Rules	follows				may be constrained								
	REAL WORLD PORTION	Physical Environment													
		Abstract Environment							may contain						
		Static Object		characterize					may contain	may contain					
		Dynamic Object							may contain						
		Physical Features	owns				is influenced by				has	has			
		Spatial Position	may be located on								may be located on	may be located on			
		Structural Constraints					is constrained by		imposes						
	SCENARIO	Rules/Constraints	follows/is constraine					are	imposes						are
		Interaction	are involved in												
		Actions	performs/ is involed				aggregate								
	SOCIAL STRUCTURE	Role	adopt				may be determined								
		Group	may belong to												
		Social Rule	respects				may be influenced by								
	SIMULATION PURPOSE	Parameter	is correlated	may be	may be	may be	may be corelated to	may be					may be	may be	may be
		Objectives													

mapping (x) means that the concept has been encountered in all the papers we examined for the specific domain. Partial mapping (v) means that only some papers of the same domain use the concept the cell refers to. No mapping is represented by (-).

For space constraints, we only show an excerpt of the work product related to the relationships (see Table II). This is a table where the kind of relation discovered between two concepts is reported. The values of darker cells are the same of their symmetric cells. Only to name a few, we have detected the following relationships: *(i)* an active entity interacts with active entities; *(ii)* the behavior may be constrained by behavioral rules; *(iii)* physical environment may contain static object; *(iv)* a static object may be located on a physical space and many others.

c) Metamodel Definition: The study of the simulation problems under analysis, Table I and Table II highlighted that, although the problem domains are quite different, they share some high level features and differ for some others. For instance, looking at the table's columns, we may see that Abstract Environment is not present in the Crowd Dynamics and Traffic and Transportation domains and it is only partially present in the other two. The Physical Environment has always been found in the Crowd Dynamics and Traffic and Transportation domains and only partially in the Power Engineering and in the Logistics and Supply Chain management domains. So, we may say that the problem statement metamodel for the simulation study of a specific domain may be composed of a core metamodel, where common elements for all the studied domain may be found, and extensions where some specific and particular elements are shown (Fig. 3).

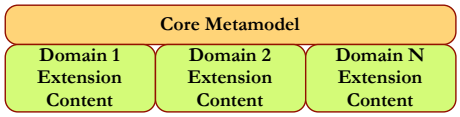


Fig. 3. Metamodel Scheme

Hence, the *Core Metamodel* contains all the elements we identified derived from the set of concepts that are shared among all domains under analysis. As a consequence, the role of the *Core Metamodel* will be to provide a minimum set of elements to be described in a problem statement for a non specific agent based simulation study. The *Domain Extension Contents* are collections of metamodel elements that allow for defining "more specific and domain dependent" contents for a

problem statement of agent based simulation studies belonging to a particular domain.

The resulting *Core Metamodel* is shown in Fig. 4. It aims to support the definition of an agent based simulation problem by identifying those elements that, once defined in the problem statement, will provide a shared starting point for the development of an agent based simulation model. The *Domain Extension Contents* for Traffic and Transportation and Power Engineering domains are shown in Fig. 5 and Fig. 6 respectively. As concerns, the *Domain Extension Content* for Crowds Dynamics domain, it entirely overlaps with that Traffic and Transportation Extension Content. For space constraints, we do not show the Logistics and Supply Chain Extension Content.

In particular, Table I, along with other considerations arisen from the activities we performed, helped us to define the elements of the metamodel. Whilst, Table II helped us to define the relationships among elements of the metamodel. The total or partial mapping also provided us information to define the cardinality of the metamodel relationships. For instance, from Table I we defined the metamodel elements: Active Entity, Behavioral Rule, Social Rule, Structural Constraint and Behavior. In Table II, we can see that these elements are linked by the following relationships: *(i)* Active Entity follows Behavioral Rules; *(ii)* Active Entity respects Social Rules; *(iii)* Behavior is constrained by Structural Constraints. Hence, we chose to model these relationships in our metamodel by introducing an abstract element named *Rule* whose Structural Constraints, Behavioral Rules and Social Rules are specializations. This choice allowed us to model the aforementioned relationships with only one link between Behavior and Rule labeled *is constrained by*. The Active Entity is related to Rule by means of its Behavior. In the same way, we deducted some other relationships and elements of the metamodel that do not explicitly appear in the tables.

In the following we describe the Core Metamodel elements showing also, when present, the related elements of the Domain Extension Contents.

Looking at Fig. 4, a generic *Simulation Problem Statement* is composed of a *System to be simulated* and *Simulation Purposes*. The *System to be simulated* in an agent based simulation study can be described by means of a set of interacting entities (i.e.: *Active Entities*) that own some distinctive attributes concerning their individuality (i.e.: *Features*) and that show particular *Behaviors*. A *Physical Feature* is a particular kind of *Feature*. It allows to describe the body properties of the entities (i.e: weight, height, etc...). The *Feature* may also be *Psychological* (i.e: impatient, polite, etc...) and *Technical Feature* (i.e.: energy consumption, capacity, etc...), we found the first in the description of the Crowd Dynamics domain and in the Traffic and Transportation domain (see the Extension Content in Fig. 5) whereas we found the second in the description of the Active Entities in the domain of Power Engineering (see the Extension Content in Fig. 6). Whilst, the *Behavior* is constrained by *Rules*. Rules can be *Behavioral Rules* (see Fig.5), *Structural Constraints* and *Social Rules* (see Fig. 5). A *Behavioral Rule* is an explicit statement or principle governing the functionality, the conduct or the procedures within a particular domain, commonly prescribing what is possible or allowable. *Social Rules* and *Structural Constraints* are defined in the following.

Commonly, the primary purpose of a simulation study is to analyze the behavior of a system under some conditions. Thus, the simulation purpose is defined by identifying the *Questions* to be answered and by determining what are the *Parameters* to be investigated from which the resulting simulation model will depend on. In this context, a Parameter denotes a measurable factor that can be varied during simulation experiments and may characterize the system and/or determine its behavior (i.e: *System Parameter*). *Parameters* are related to each *Active Entity* present in the system so, since *Active Entities* own *Features* and show *Behaviors*, parameters may also characterize or determine *Behavior* and may be related to *Features* through the *Entity Parameter*. For instance the velocity of a vehicle is a physical feature described together with the related active entity, so, in this case, a physical feature may also be realized by an *Entity Parameter*.

Moreover, the *System to be simulated* concerns a *Real-World Portion* that includes *Active Entities* and passive entities (i.e.: *Objects*). These latter do not exhibit behaviors but contribute to the description of the *System to be simulated*. The *Real-World Portion* is a *Physical Environment* in the domains of Crowd Dynamics and Traffic and Transportation and it may impose several *Structural Constraints* (see Fig. 5) derived by its physical configuration. Whilst, it can be also an *Abstract Environment* like it is in the domain of Power Engineering (see Fig. 6) for electricity markets.

In the same way, *Objects* may be *Static* or *Dynamic Objects* in the Crowd Dynamics and in the Traffic and Transportation domain but only static in the Power Engineering domain[2] (see the Extension Contents in Fig. 5 and Fig. 6 respectively).

Moreover, *Active Entities* and *Objects* are located on a *Spatial Position* when the problem concerns a *Physical Environment*.

For describing the *System to be simulated* it is also necessary to identify *Scenarios* to be investigated. A *Scenario* is a description of the *Interactions* that occur among entities and/or the *Actions* individually performed by *Active Entities*. A *Scenario* describes a way in which the *System to be simulated* behaves in specific situations.

Finally, the *System to be simulated* may be further described by means of a *Social Structure*. The *Social Structure* describes the way the *System to be simulated* is organized according to specific *Roles* played by *Active Entities*. A *Role* describes what an active entity is able to do. It expresses a set of behaviors showed by an Active Entity when it is involved in a social

[2]We would like to point out again that these considerations are the result of a preliminary work. We think to be likely that Dynamic Objects can be found in the Power Engineering Domain, but at the moment this does not emerge from the examined papers. Whether the element *Dynamic Objects* should be present also in the domain of the Power Engineering, then both elements, *Dynamic Object* and *Static Object*, will be moved within the Core Metamodel.

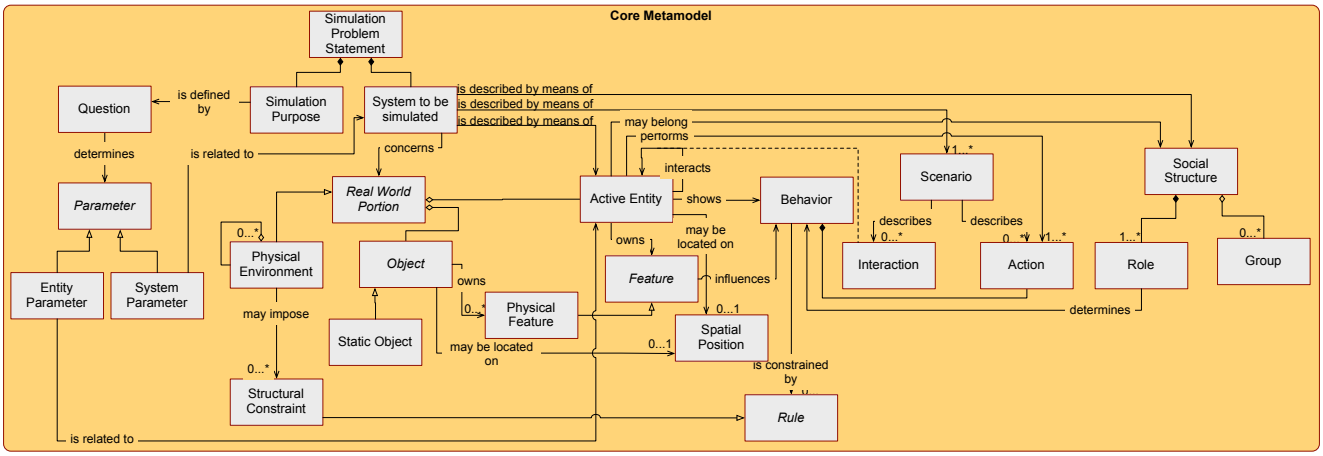


Fig. 4. The Core Metamodel derived from the domain under analysis.

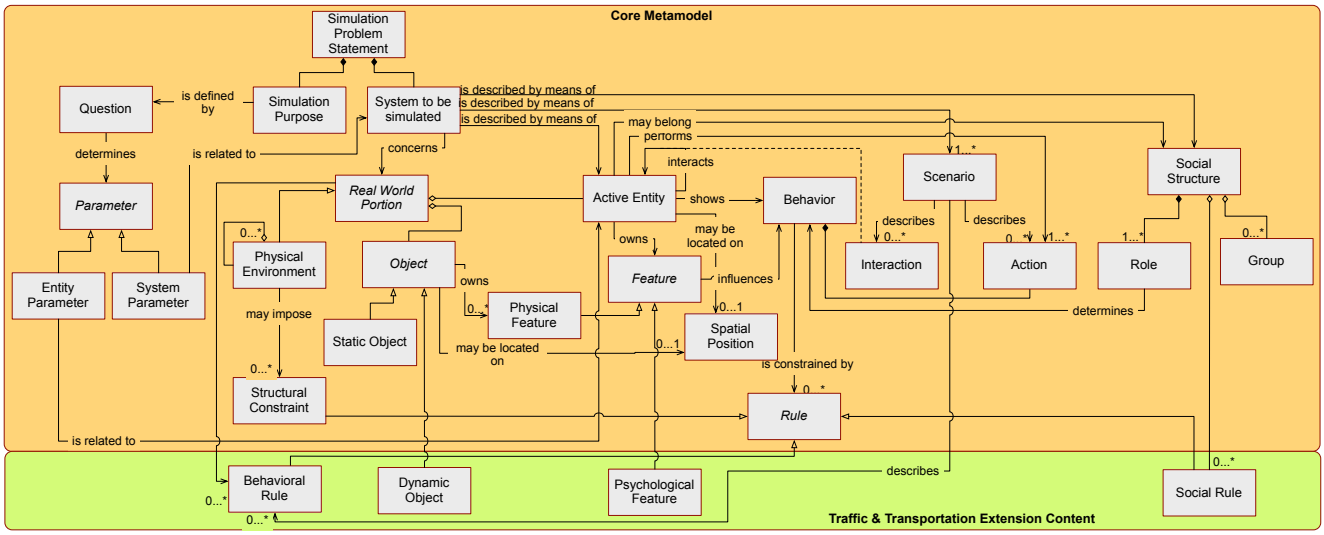


Fig. 5. Core Metamodel along with Traffic and Transportation Extension Content.

pattern. A *Group* defines an aggregate of active entities that can be together in the same place or that can be connected by some shared behavior or feature. A *Social Rule* has the same role of a *Behavioral Rule* but it is shared and followed by the members of a *Group*.

IV. Discussions

The focus of the Problem Statement Activity is to acquire all the necessary knowledge of the real-world system to be simulated and to point out the issues that the simulation studies have to address. In particular, without a definitive statement of the specific questions of interest to be addressed, it is impossible to decide on an appropriate level of knowledge detail useful to build an appropriate simulation model. A method to clearly define what type of the knowledge about the problem is relevant may be useful in order to avoid missing important parts of the problem description. The metamodel resulting from our study allows to handle two levels of knowledge at the same time. A knowledge related to the system to be simulated as a whole. It concerns aspects of the system that are not directly related to the individual entities and it helps to build the so-called environment model[3]. A knowledge relating to specific aspects of single entities that will be suited for building the agent models. Moreover, preliminary results we have obtained give us some suggestions about the knowledge to be included in a problem statement and how to describe it in a systematic way. The Core Metamodel highlights that the formulation of a simulation study may ground on five key aspects:

[3]An agent based simulation model is in many cases composed of two basic components. They are an agent model that focuses on each single agent and its behavior and an environment model that focuses on the structure and features of the real world portion in which agents act.

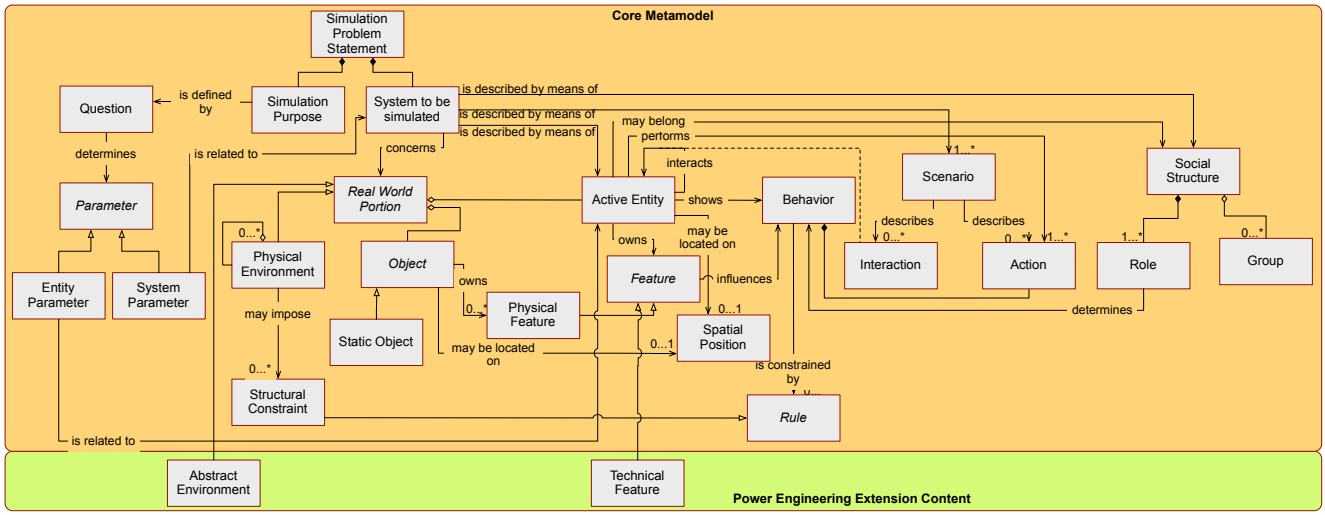


Fig. 6. Core Metamodel along with Power Engineering Extension Content.

1) the *issues* to be addressed by the simulation study made explicit by means of the metamodel element named *Simulation Purpose*. Instantiating this metamodel element means to formulate the questions to be answered that coincide with the definition of goals of the simulation study. As a consequence, this leads to associate goals with information that define them operationally and that has to be considered during the definition of the simulation model.
2) the *structural aspect* of the system to be simulated defined by its constituting parts. It is made explicit by instantiating the metamodel element named *Real World Portion*. This means to collect the knowledge necessary to describe the real-world part involved in the system to be simulated according to the issues to be addressed. Commonly, this concerns the description of the physical configuration of the real-world system under analysis (i.e: its components). As the Core metamodel shows, components may be static objects of the portion of the world enclosed by the system or active entities that are someone/something performing actions;
3) the *dynamical aspect* of the system to be simulated caused by processes taking place in the system. It is captured by defining actions and interactions that are performed by someone/something in particular operative scenarios.
4) the *organizational aspect* (if any) which refers to the social structure captured by defining the groups and sets of roles that may be present in the system to be simulated.
5) the *normative aspect* (if any) given by regulations that commonly constrains the dynamics or physicality of the system. It refers to any kind of regulations such as procedural rules, legal regulations, social norms, structural rules or physical laws. Usually, only some kinds of norm affect the system, this strongly depends on the particular domain. For example, in a simulation study concerning a manufacturing system normally the system is subjected to procedural rules and legal regulations. Vice versa, if the simulation is about the citizen of a town during an event then, probably, social norms could affect the system.

These key aspects may provide a first guideline about the kind of knowledge that can be managed during the problem statement activity. Moreover, basing on such a metamodel we are able to establish a starting point for the construction of a systematic approach for performing the problem statement activity by choosing the order that should be used for instantiating the metamodel elements. We already explored such an approach with the PRoDe method for developing software design processes starting from the underlying metamodel [14].

V. Conclusions and Remarks

Multiagent simulations are employed in several domains and for several simulation purposes. Finding a uniform way to address the problem description for a simulation study is an open issue. The work proposed in this paper is a preliminary step toward an ambitious objective: the definition of a complete methodological approach to obtain high quality simulation models for generic agent-based simulation studies. Such an approach will cover several activities starting from the problem statement to the results analysis trough the model validation. At the moment, we are addressing some questions related to the problem statement activity and in this paper we present our preliminary results.

As well as in many engineering contexts, we think that an accurate problem formulation is the first step to be accomplished in order to obtain a high-quality agent based simulation model. To do this, we are defining a metamodel-based approach in order to perform the problem statement

activity. The metamodel aims to guide the problem formulation highlighting the main elements and relationships that have to be made explicit. In this paper we present the rough metamodel we have already defined. This metamodel is composed of elements belonging to a shared level among several simulation problem domains useful for formulating generic simulation studies and several domain extension contents that allow us for detailing specific domain simulation studies.

REFERENCES

[1] F. Klügl, "Engineering agent-based simulation models?" in *Agent-Oriented Software Engineering XIII*, ser. Lecture Notes in Computer Science, J. P. Müller and M. Cossentino, Eds. Springer Berlin Heidelberg, 2013, vol. 7852, pp. 179–196. ISBN 978-3-642-39865-0

[2] O. Labarthe, E. Tranvouez, A. Ferrarini, B. Espinasse, and B. Montreuil, "A heterogeneous multi-agent modelling for distributed simulation of supply chains," in *Holonic and Multi-Agent Systems for Manufacturing*, ser. Lecture Notes in Computer Science, V. Marík, D. McFarlane, and P. Valckenaers, Eds. Springer Berlin Heidelberg, 2003, vol. 2744, pp. 134–145. ISBN 978-3-540-40751-5

[3] O. Balci, "Guidelines for successful simulation studies," in *Simulation Conference, 1990. Proceedings., Winter*, Dec 1990. doi: 10.1109/WSC.1990.129482 pp. 25–32.

[4] R. E. Nance, "The conical methodology and the evolution of simulation model development," *Annals of Operations Research*, vol. 53, no. 1, pp. 1–45, 1994. doi: 10.1007/BF02136825

[5] A. Law, "How to build valid and credible simulation models," in *Simulation Conference (WSC), Proceedings of the 2009 Winter*, Dec 2009. doi: 10.1109/WSC.2009.5429312 pp. 24–33.

[6] C. M. Macal and M. J. North, "Tutorial on agent-based modeling and simulation," in *Proceedings of the 37th conference on Winter simulation*. Winter Simulation Conference, 2005. doi: 10.1057/jos.2010.3 pp. 2–15.

[7] O. Balci and R. E. Nance, "Formulated problem verification as an explicit requirement of model credibility," *SIMULATION*, vol. 45, no. 2, pp. 76–86, 1985. doi: 10.1177/003754978504500204

[8] A. Garro and W. Russo, "< i> easyabms</i>: A domain-expert oriented methodology for agent-based modeling and simulation," *Simulation Modelling Practice and Theory*, vol. 18, no. 10, pp. 1453–1467, 2010.

[9] S. Viller and I. Sommerville, "Ethnographically informed analysis for software engineers," *International Journal of Human Computer Studies*, vol. 53, no. 1, pp. 169–196, 2000. doi: 10.1006/ijhc.2000.0370

[10] M. Cossentino and V. Seidita, "PASSI: Process for agent societies specification and implementation," in *Handbook on Agent-Oriented Design Processes*, M. Cossentino, V. Hilaire, A. Molesini, and V. Seidita, Eds. Springer Berlin Heidelberg, 2014, pp. 287–329. ISBN 978-3-642-39974-9

[11] A. Chella, M. Cossentino, L. Sabatucci, and V. Seidita, "Agile passi: An agile process for designing agents," *Computer Systems Science and Engineering*, vol. 21, no. 2, pp. 133–144, 2006.

[12] M. Cossentino, V. Hilaire, N. Gaud, S. Galland, and A. Koukam, "The ASPECS process," in *Handbook on Agent-Oriented Design Processes*, M. Cossentino, V. Hilaire, A. Molesini, and V. Seidita, Eds. Springer Berlin Heidelberg, 2014, pp. 65–114. ISBN 978-3-642-39974-9

[13] V. Seidita, M. Cossentino, and S. Gaglio, "Adapting PASSI to support a goal oriented approach: a situational method engineering experiment," in *Proc. of the Fifth European workshop on Multi-Agent Systems (EUMAS'07)*, 2007.

[14] V. Seidita, M. Cossentino, V. Hilaire, N. Gaud, S. Galland, A. Koukam, and S. Gaglio, "The metamodel: a starting point for design processes construction," *International Journal of Software Engineering and Knowledge Engineering*, vol. 20, no. 04, pp. 575–608, 2010. doi: 10.1142/S0218194010004785

[15] K. Boer-Sorban, *Agent-Based Simulation of Financial Markets: A Modular, Continuous-time Approach*. Erasmus Research Institute of Management (ERIM), 2008, no. EPS-2008-119-LIS.

[16] M. Lovric, *Behavioral Finance and Agent-Based Artificial Markets*. Erasmus Research Institute of Management (ERIM), 2011, no. EPS-2011-229-F&A.

[17] I. Benenson, "Multi-agent simulations of residential dynamics in the city," *Computers, Environment and Urban Systems*, vol. 22, no. 1, pp. 25–42, 1998. doi: 10.1016/S0198-9715(98)00017-9

[18] J. Perret, F. Curie, J. Gaffuri, and A. Ruas, "A multi-agent system for the simulation of urban dynamics," in *10th European Conference on Complex Systems (ECCS'10), Lisbon, Portugal*, 2010.

[19] P. Paruchuri, A. R. Pullalarevu, and K. Karlapalem, "Multi agent simulation of unorganized traffic," in *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*. ACM, 2002. doi: 10.1145/544741.544786 pp. 176–183.

[20] D. Meignan, O. Simonin, and A. Koukam, "Simulation and evaluation of urban bus-networks using a multiagent approach," *Simulation Modelling Practice and Theory*, vol. 15, no. 6, pp. 659–671, 2007. doi: 10.1016/j.simpat.2007.02.005

[21] G. Waizman, S. Shoval, and I. Benenson, "Micro-simulation model for assessing the risk of car-pedestrian road accidents," in *7th International Workshop on Agents in Traffic and Transportation, Valencia, Spain*, 2012.

[22] S. J. Guy, J. Chhugani, S. Curtis, P. Dubey, M. Lin, and D. Manocha, "Pledestrians: a least-effort approach to crowd simulation," in *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. Eurographics Association, 2010, pp. 119–128.

[23] N. Pelechano, J. M. Allbeck, and N. I. Badler, "Controlling individual agents in high-density crowd simulation," in *Proceedings of the 2007 ACM SIGGRAPH/Eurographics symposium on Computer animation*. Eurographics Association, 2007, pp. 99–108.

[24] A. Shendarkar, K. Vasudevan, S. Lee, and Y.-J. Son, "Crowd simulation for emergency response using bdi agent based on virtual reality," in *Proceedings of the 38th conference on Winter simulation*. Winter Simulation Conference, 2006. doi: 10.1109/WSC.2006.323128 pp. 545–553.

[25] F. Menges, B. Mishra, and G. Narzisi, "Modeling and simulation of e-mail social networks: a new stochastic agent-based approach," in *Proceedings of the 40th Conference on Winter Simulation*. Winter Simulation Conference, 2008, pp. 2792–2800.

[26] P. Holme and G. Ghoshal, "Dynamics of networking agents competing for high centrality and low degree," *Physical review letters*, vol. 96, no. 9, p. 098701, 2006. doi: 10.1103/PhysRevLett.96.098701

[27] L. A. de Santa-Eulalia, J.-M. Frayret, and S. D'Amours, "Essay on conceptual modeling, analysis and illustration of agent-based simulations for distributed supply chain planning," *INFOR*, vol. 46, no. 2, pp. 97–116, 2008.

[28] S. Buckley and C. An, "Supply chain simulation," in *Supply Chain Management on Demand*, C. An and H. Fromm, Eds. Springer Berlin Heidelberg, 2005, pp. 17–35. ISBN 978-3-540-24423-3

[29] M. Cossentino, C. Lodato, S. Lopes, and P. Ribino, "Multi agent simulation for decision making in warehouse management," *2011 Federated Conference on Computer Science and Information Systems, FedCSIS 2011*, pp. 611–618, 2011.

[30] Z. Vale, T. Pinto, I. Praça, and H. Morais, "Mascem: electricity markets simulation with strategic agents," *Intelligent Systems, IEEE*, vol. 26, no. 2, pp. 9–17, 2011. doi: 10.1109/MIS.2011.3

[31] S. Karnouskos and T. N. De Holanda, "Simulation of a smart grid city with software agents," in *Computer Modeling and Simulation, 2009. EMS'09. Third UKSim European Symposium on*. IEEE, 2009. doi: 10.1109/EMS.2009.53 pp. 424–429.

[32] N. W. Oo and V. Miranda, "Multi-energy retail market simulation with intelligent agents," in *Power Tech, 2005 IEEE Russia*. IEEE, 2005. doi: 10.1109/PTC.2005.4524755 pp. 1–7.

[33] X. Pan, C. S. Han, K. Dauber, and K. H. Law, "A multi-agent based framework for the simulation of human and social behaviors during emergency evacuations," *Ai & Society*, vol. 22, no. 2, pp. 113–132, 2007. doi: 10.1007/s00146-007-0126-1

[34] J.-M. Frayret, S. D'amours, A. Rousseau, S. Harvey, and J. Gaudreault, "Agent-based supply-chain planning in the forest products industry," *International Journal of Flexible Manufacturing Systems*, vol. 19, no. 4, pp. 358–391, 2007. doi: 10.1007/s10696-008-9034-z

[35] J. M. Swaminathan, S. F. Smith, and N. M. Sadeh, "Modeling supply chain dynamics: A multiagent approach," *Decision Sciences*, vol. 29, no. 3, pp. 607–632, 1998. doi: 10.1111/j.1540-5915.1998.tb01356.x

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1477–1485
DOI: 10.15439/2014F395
ACSIS, Vol. 2

S-MASA: A Stigmergy Based Algorithm for Multi-Target Search

Ouarda Zedadra and Hamid Seridi
LabSTIC Laboratory, 8 may 1945 University
P.O.Box 401, 24000 Guelma, Algeria
Department of computer science Badji Mokhtar
Annaba University P.O.Box 12, 23000 Annaba, Algeria
Email: zedadra_nawel1, seridihamid@yahoo.fr

Nicolas Jouandeau
Advanced Computing laboratory
of saint-Denis Paris 8 University
Saint Denis 93526, France
Email: n@ai.univ-paris8.fr

Giancarlo Fortino
DIMES, Universita'
della Calabria
Via P. Bucci, cubo 41c - 87036
- Rende (CS) - Italy
Email:g.fortino@unical.it

***Abstract*—We explore the on-line problem of coverage where multiple agents have to find a target whose position is unknown, and without a prior global information about the environment. In this paper a novel algorithm for multi-target search is described, it is inspired from water vortex dynamics and based on the principle of pheromone-based communication. According to this algorithm, called S-MASA (Stigmergic Multi Ant Search Area), the agents search nearby their base incrementally using turns around their center and around each other, until the target is found, with only a group of simple distributed cooperative Ant like agents, which communicate indirectly via depositing/detecting markers. This work improves the search performance in comparison with random walk and S-random walk (stigmergic random walk) strategies, we show the obtained results using computer simulations.**

I. Introduction

THE PROBLEM of finding multiple targets whose positions are unknown without a prior information about the environment is very important in many real world applications [1]. Those applications vary from mine detecting [2] [3], search in damaged buildings [4] [5], fire fighting [6], and exploration of spaces [7] [8], where neither a map, nor a Global Positioning System (GPS) are available [9]. The random walk is the best option when there is some degree of uncertainty in the environment and a reduced perceptual capabilities [10] because it is simple, needs no memory and self-stabilizes. However, it is inefficient in a two-dimensional infinite grid, where it results in an infinite searching time, even if the target is nearby [11], it results also in energy consumption and malfunction risks. To deal with these limits, some effective ways to coordinate multiple agents in their searching task need to take place. Recently many researchers have investigated bio-inspired coordination methods [12] [13], in which agents coordinate on the basis of indirect communication principle known as stigmergy.

Complexity of multi-target search solutions depends on simplifications considered over idealized assumptions, such as: perfect sensors [14], stationary environments [15], unlimited direct communication [16]. Even if these assumptions are far from real world applications, they provide first basic solutions. The algorithm presented in this paper avoids such type of assumptions. It makes the following contributions:

1) it is of very low computational complexity, in which agents have a very low-range of sensors;
2) it executes a search in nearby locations first by adopting spiral turns around the starting cell and around agents each other;
3) agents use stigmergic communication via digital pheromone;
4) it can be executed on known or unknown static obstacle-free environments or obstacle environments.

The rest of this paper is organized as follows. Section 2 discusses some related work. Section 3 describes the problem statement and formulation. S-MASA algorithm is described in detail in Section 4. Performance evaluation is given in Section 5. A comparison with the random walk and S-random walk strategies are given in Section 6 and Section 7 concludes the paper.

II. Related Work

The problem of searching a target may be considered as a partial area coverage problem that constitutes a key element of the general exploration problem [17] where coverage can be done by a single or multiple robots, with on-line or off-line algorithms. In the on-line coverage algorithms, the area and target positions are unknown, and are discovered step by step while the robot explores the environment, whereas, in the off-line algorithms, the robot has a prior information about the environment, target and obstacles positions, so it can plan the path to go through. Different approaches have been developed in the literature to solve area coverage using single or multiple robots. In this section, a brief overview of techniques that are used to solve the coverage problem using both single and multiple robots is presented. The single robot covering problem was explored by Gabriely and Rimon [18]. One of the most popular algorithms is the Spanning Tree Coverage (STC). In an STC algorithm, the robot operates in a 2D grid of large square cells. It aims to find a spanning tree for such grid, and allow the robot to circumnavigate it. This algorithm covers every cell that is accessible from the starting point, and it is optimal because the robot passes through each cell at least once [19]. Spiral STC is an online sensor based algorithm for covering planar areas by a square shaped tool attached

978-83-60810-58-3/$25.00

to a mobile robot. The algorithm incrementally subdivides the planar work into disjoint D size cells, while following a spanning tree of the resulting grid. The spiral STC covers every subcell accessible from the starting point, and covers these subcells in O(n) time using O(n) memory [20]. In this new version of STC, the spanning tree is stored in the onboard memory, which results in a dependency of the search area on memory size. With the aim of resolving the memory problem, Gabriely and Rimon propose in [21] the ant-like STC which forms the third version of the basic STC algorithm, that uses markers on visited cells. D-STC is introduced in [21] to solve the problem of uncovered partially occupied 2D-size cells, by visiting the previously uncovered cells, which results in worst-case scenarios, a twice coverage of the environment area. A generalization of STC to multi-robots is given in [22], the MSTC, in which a spanning tree is computed, and then it is circumnavigated by each robot. Another spanning tree construction using multiple robots based on approximate cellular decomposition is proposed in [23]. Another approach developed in [1], where the environment is subdivided into n concentric discs, each disc is covered by one robot, when the entire disc is completely covered, the robot move to the next disc not yet covered; an extension of this algorithm that uses heterogeneous robots is given in [17]. Instead of focusing on the on-board resources, some part of robotics literature use a single ant or a group of robots to cover an area robustly, even if they do not have any memory, do not know the terrain, cannot maintain maps of the terrain, nor plan complete paths. They use environmental markers such as pebbles [24], [25], [26] or pheromone like traces [27] or greedy navigation strategies [28].

Whether we deal with coverage, multi-target search as foraging task, we need at the first stage to search the corresponding area. A search is defined as the action to look into the area carefully and thoroughly in an effort to find or discover something [29]. In most search strategies based on random walk, the agent tends to return to the same point many times before finally wandering away, because it has no historical information about visited regions. But when time and energy consumption are determinants, it will be efficient to guide the agent to not visited regions and repulse it from visited ones. In [30] a cooperative and distributed coordination strategy (IAS-SS) is proposed, it is applied to exploration and surveillance of unknown environments. It is a modified version of the artificial ant system, where the pheromone left has the property of repealing of robots either than attraction. A guided probabilistic exploration strategy for unknown areas is presented in [31], it is based on stigmergic communication and combines the random walk movements and the stigmergic guidance. The paper [32], provide a simple foraging algorithm that works asynchronously with identical ants, based on marking visited grid points by pheromone. It lacks robustness to faults. Authors in [33], propose a swarm intelligence based algorithm for distribute search and collective clean up. In this algorithm, the map is divided into a set of distinct sub-area and each sub-area is divided into some grid. Each robot decides individually based on its local information to which subarea it should move. A direct communication via WIFI model is used between robots and their neighbors. The paper [11], introduce the ANTS (Ants Nearby Treasure Search) problem, in which k identical agents, initially placed at some central location, collectively search for a treasure in a two-dimensional plane, without any communication between them. A survey of online algorithms for searching and exploration in the plane is given in [34]. S-MASA is a simple search algorithm that uses pheromones to guide the search process, agents are reactive and do not need any memory. It can locate nearby targets as fast as possible and at a rate that scales well with the number of agents, so it operates as some animal species that search for food around a central location, known as central place foraging theory [35]. Table I gives a comparison between our algorithm and some of the related works according to the search process used.

Even if chemical substances [36], electrical devices such as Radio Frequency Identification Devices (RFIDs) [37] [38] [39] [40] are examples of real implementation of stigmergic communication in real world experiments, it is still important to understand and improve pheromone-based algorithms in simulations. By understanding the optimal conditions required for pheromone-based coordination, the real world implementations can also be better directed [31].

III. Problem Statement and Formulation

In a collective multi-target search task, there are a lot of targets randomly distributed in an area. The agents (robots) should find as fast as possible the targets and, after that, remove them, if we deal with a cleanup task, or transport them to a nest, if we deal with a foraging task. In this paper, a new search algorithm is proposed that enables a group of agents, each with limited perception capabilities to search quickly the targets. The algorithm presented here uses the principle of pheromone-based coordination where each agent deposits pheromone on its environment to inform the others about already visited areas. The finish time of the collective search is when all targets have been found. This section defines and clarifies some key terms which will be used in this paper.

- *Environment:* we assume that agents move in an N X M grid-based environment. It is divided into N X M cells. Each cell can be an obstacle, target or the base station, and can also contain an agent.
- *Agent:* simple reactive agents, with limited range sensor (can only perceive the four neighboring cells), have no memory and use the environment as their shared memory. Each agent has an initial position and heading (0, 90, 180 or 270).
- *Pheromone:* has a numerical meaning. It is represented by a color. The intensity of the pheromone at time t is set to arbitrarily chosen value c which is a small positive constant. It evaporates with time with a coefficient p fixed to 0.075 using equation 1 to avoid accumulation of pheromone.

TABLE I: A Comparison of Related Work

Reference	**[32]**	**[30]**	**[31]**	**[11]**	**[33]**	**S-MASA**
Multi-Agent	Yes	Yes	Yes	Yes	Yes	Yes
Heterogeneous	No	Yes	Yes	No	Yes	No
I.L	(0,0)	Room 1	Random	(0,0)	Random	Given
Online	Yes	Yes	Yes	Yes	Yes	Yes
Environment	I.2D.G	2D.G.R	B.2D.G	I.2D.G	B.2D.G	B.2D.G
Sensors	F.G.N	LS.R.S	F.N	L.R.S	F.G.N	F.G.N
Simulations	No	2DX	Robots	No	3DX	Agents
Redundancy	No	Yes	Yes	No	No	Yes
Robustness	No	Yes	Yes	Yes	Yes	No
Complete	Yes	Yes	Yes	Yes	Yes	Yes
Distributed	Yes	Yes	Yes	Yes	Yes	Yes
Collaboration	Yes	Yes	Yes	Yes	Yes	Yes
Communication	S.C	S.C	S.C	No	Direct	S.C
Problem	M.T.S	E.S	Exploration	ANTS	D.S.C	M.T.S and C

I.L: Initial Locations, I.2D.G: Infinite 2D grid, 2D.G.R: 2D Grid with 7 Rooms, B.2D.G: Bounded 2D Grid, F.G.N: Four Grid Neighbors, F.N: Five Neighbors, LS.R.S: LaSer Range Sensor, L.R.S: Low Range Sensor, M.T.S: Multi-Target search, E.S: Exploration and Surveillance, D.S.C: Distributed Search and Clean up, S.C: Stigmergic Communication, C: Coverage

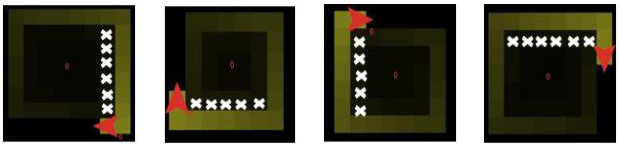
- *Motion policy:* each agent chooses the next cell to visit using a motion policy that is function of the presence of pheromone trail and obstacles. This policy helps the agent to decide where to go next.

IV. Design of the S-MASA Algorithm

The idea behind proposing this algorithm is to reproduce the behavior observed in water vortex dynamics. The vortex is a region in which a fluid flow is mainly a rotary movement about an axis, rectilinear or curved. So each agent tries to turn around the base station and around the other agents. Doing this with agents only is difficult and needs a great number of agents, but using pheromone to repulse agents from visited cells was very helpful to reproduce the structure of a vortex.

A. Basic S-MASA

In S-MASA, each agent started from an initial given position and oriented toward a given heading. To turn around the base station and around each other, each agent checks on his right cell if it is visited or not. If it detects a pheromone (Figure 1), it indicates to the agent that it is about to enter to a visited cell and therefore the agent keeps going forward its current heading, else the agent changes its heading and moves toward a new heading. S-MASA is further detailed in Algorithm 1.



(a) (b) (c) (d)

Fig. 1: S-MASA coordination principle: (a) Changing heading from 180 to 270 (b) Changing heading from 270 to 0 (c) Changing heading from 0 to 90 (d) Changing heading from 90 to 180, where white crosses represent already visited cells

Algorithm 1 S-MASA

Input: position and heading for each agent,
Output: iteration number,

```
while number of targets and boundaries are not reached do
    Move
    Lay pheromone
    Update Pheromone
end while
```

Move function is the motion policy. Each agent has initially a given heading (0, 90, 180 or 270) that allows it to move up, right, down or left in the four neighboring cells. The agent checks always its right cell which is the *up* cell if the heading is 270, the *down* cell if the heading is 90, if no pheromone is there it can change his heading to the new one using the move function and goes forward in that new heading. The move function is detailed in Algorithm 2.

Algorithm 2 Function Move

```
if (pheromone is detected in right cell) then
    go forward
else if (heading = 270) then
    set heading to 0
else
    set heading to heading + 90
end if
```

Update pheromone function is used for pheromone evaporation, using the equation

$$\Gamma_i(t+1) = \Gamma_i(t) - p * \Gamma_i(t) \quad (1)$$

Where: p is a coefficient which represents the evaporation of trail between time t and $(t+1)$ and is set to 0.075 to avoid unlimited accumulation of pheromone. S-MASA can be applied to environment with or without obstacles, the agent executes the function avoid obstacle to avoid obstacles, where

the agent follows in this case the obstacle boundary until a not visited cell is encountered, which means that agents are going around the obstacle in the direction of visited cells to guarantee the completeness of the algorithm.

B. S-MASA Extensions

The proposed algorithm allow to cover gradually the environment starting from the base station and reproducing by the way principle of central place foraging theory [35]. Although, this algorithm generates very efficient search results based on relatively simple motion rules, it can be extended to deal with dynamically changing environments, and with coverage problem in known or unknown environments.

V. Performance Evaluation

We used Netlogo framework [41] to evaluate the performance of our algorithm in two scenarios. In the first scenario we evaluate the algorithm by varying the number of agents from 5 to 30 agents in two environment configurations: obstacle-free environment and obstacle environment. In the second scenario, we evaluate the algorithm by varying the size of the environment from 20 X 20 cells to 100 X 100 cells. Obstacles in the two scenarios were defined in two ways: (i) given a desired percentage, cells were randomly designated as obstacles (ii) obstacles were specifically designed by hand. Then, one possible extension on S-MASA is discussed and related simulation results are illustrated. To evaluate average performance, each simulation is repeated 20 times, where time is defined as the number of iterations required by the agents to discover all the targets.

A. Scenario 1: Influence of Number of Agents on Performance

Agents start from initial given positions and each agent has a heading, we vary the number of agents from 5 to 30. The environment consists of a square of size 40 X 40 cells shown in Figure 2, free or with obstacles, with four targets distributed randomly. An example of execution of S-MASA on a group of 5 agents, 30 agents on obstacle-free environment, a group of 30 agents on obstacle environment (obstacles are uniformly distributed) and a group of 5 agents on obstacle cluster environment (where the distribution of obstacles in the environment gives a cluster or line shapes either than the uniform distribution) are illustrated in Figure 2.

Table II shows the performance of the algorithm in scenario 1 while the number of agents is varying from 5 to 30. It is represented graphically in Figure 3. The search time becomes dramatically faster with an increase in the number of agents. Note that there is no direct communication between agents, the only communication tool is the pheromone deposited in the environment. The standard deviation of the number of iterations reflects the impact of the random distribution of the targets between simulations. There is a linear decrease in the iterations number.

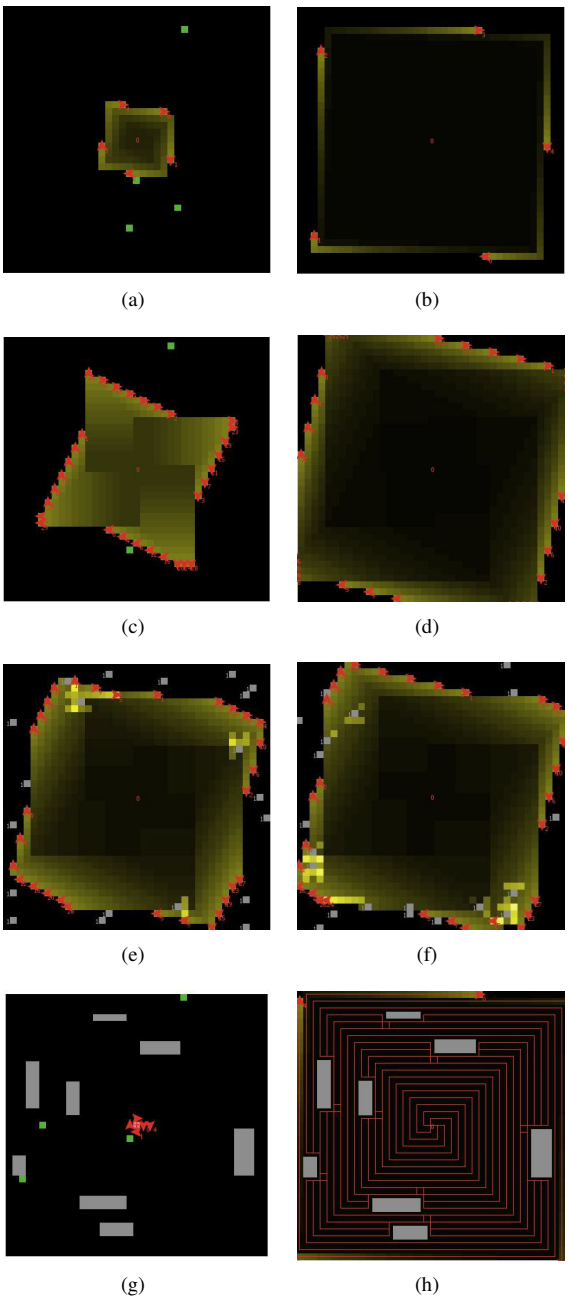


Fig. 2: The evolution of search achieved by S-MASA: (a), (b) Initial and final position of the 5-agents group in an obstacle-free environment. (c), (d) Initial and final position of the 30-agents group in an obstacle-free environment. (e), (f) Initial and final position of the 30-agents group in an obstacle environment. (g), (h) Initial and final position of 5 agents group in an obstacle cluster environment.

TABLE II: Effect of agent number on performance

	5	10	15	20	25	30
Iterations in free env	242,85	122,2	78,85	63,5	54,8	43,9
STD Deviation	46,62	24,84	17,87	14,15	11,35	10,15
Iterations in obstacle env	289,85	143,35	114,15	93,55	71,8	69,55
STD Deviation	56,76	36,93	18,29	18,58	22,97	22,24

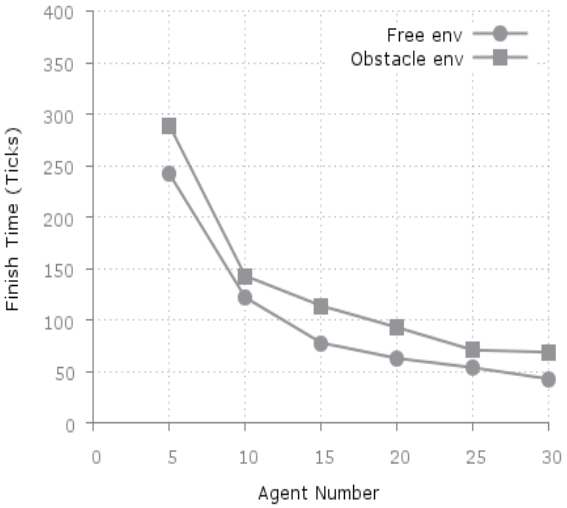


Fig. 3: Effect of agent number on performance in obstacle-free and obstacle environment

B. Scenario 2: Influence of Environment Size on Performance

We now show how the size of the environment affects the performance of the algorithm when the number of agents is set to 20. Also here we used an obstacle-free environment and an obstacle environment, just varying the size of the environment from 20 X 20 cells to 100 X 100 cells.

Table III shows the performance of the algorithm in scenario 2. It is represented graphically in Figure 4. The search time increases by increasing the size of the environment which is evident because the number of cells increases. The results show a difference in iterations number, S-MASA is robust to obstacles but this increase in number of iterations is due principally to the avoidance of obstacles that takes at least four iterations more, to go around a simple obstacle.

TABLE III: Effect of environment size on performance

	20X20	40X40	80X80	100X100
Iterations on free env	16,2	63,4	254,8	366,15
STD Deviation	3,28	13,48	50,47	101,03
Iterations on obstacle env	24,5	92,2	315,3	449
STD Deviation	8,06	23,37	71,49	131,06

C. Extension 1: S-MASA for Coverage Problem

Simulations presented in this section show that by changing the finish condition of the algorithm, the agents can achieve

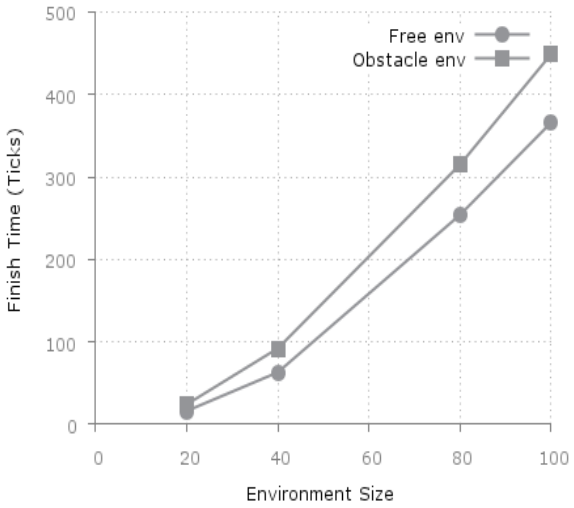


Fig. 4: Effect of environment size on performance in obstacle-free and obstacle environment

coverage mission as well as search one. The S-MASA algorithm can be applied for instance to known or unknown static environments, free or obstacle environments. Each simulation is repeated for 20 times in obstacle environments, because the obstacles are disseminated randomly in the environment and according to their position the agent take more or less iterations to go around the obstacle. Figure 5 represents the two simulations in obstacle-free and obstacle environment. As in scenario 1 and scenario 2, we test the performance of the algorithm on coverage problem by varying the number of agents and by varying the size of the environment in the two types of environments. Table IV and Figure 6 show the obtained results when varying the number of agents. There is a linear decrease in number of iterations when increasing the number of agents, and there is a difference between iterations in obstacle-free environment and obstacle environment, which are similar to Scenario 1 results. A possible reason is the random distribution of targets, so if there is one target close to boundaries, the search will be very close to coverage task and in the two tasks the number of iterations will be very close.

TABLE IV: Effect of number of agent on performance

	5	10	15	20	25	30
Iterations in free env	320	171	120	89	80	68
Iterations in obstacle env	354,25	206,7	164,3	138,6	126,25	111,55

Table V and Figure 7 show the obtained results when varying the environment size, because here there is no random distribution of targets or there are no targets, the coverage time in obstacle environment is greater than the coverage time in obstacle-free environment, but there is always an increase in the number of iterations in the two cases of simulations.

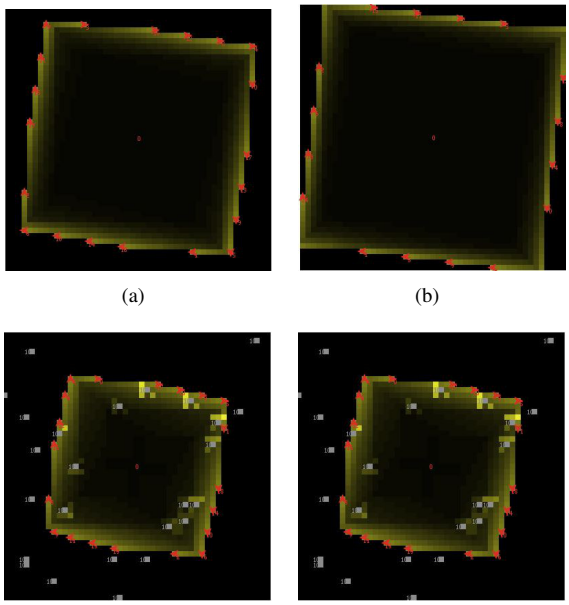


Fig. 5: The evolution of coverage achieved by S-MASA: (a), (b) 20-agents group in an obstacle-free environment in iterations 78 and 101. (c), (d) 20-agent group in an obstacle environment in iterations 44 and 108.

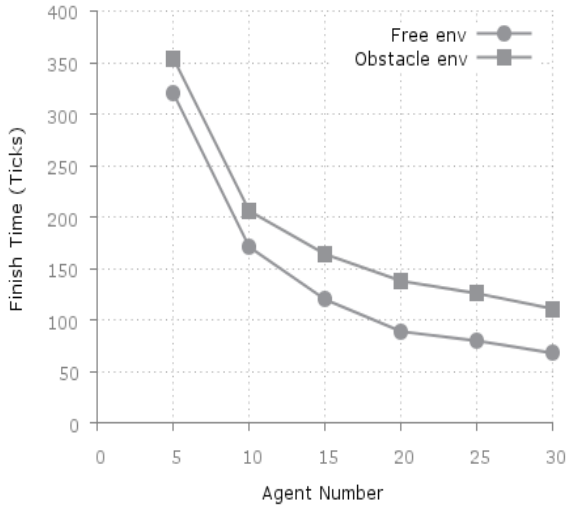


Fig. 6: Finish time of coverage in free-obstacle and obstacle environment when varying the number of agents

TABLE V: Effect of environment size on performance

	20X20	40X40	80X80	100X100
Iterations on free env	23	89	341	527
Iterations on obstacle env	55,4	135,7	435,9	625,1

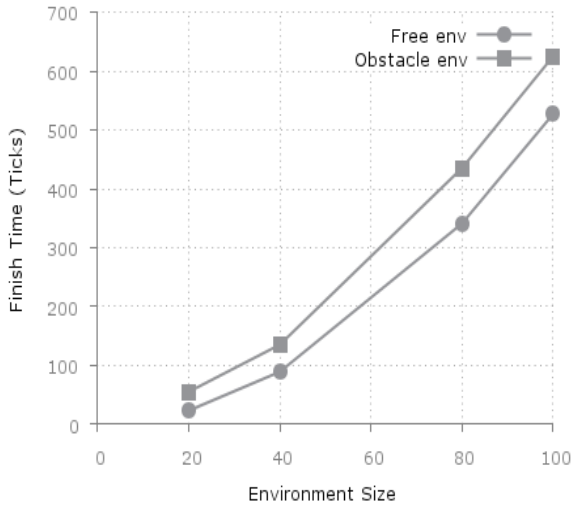


Fig. 7: Finish time of coverage in free-obstacle and obstacle environment when varying the environment size

VI. Comparison results

We compare S-MASA algorithm with two search strategies: the random walk, in which the agent chooses randomly one of the four neighbors even if it is already visited, that causes a increase in the global finish time; and the S-random walk, in which the agent chooses one of its four neighbors that are not visited yet, at each step the agent deposit a pheromone to mark already visited cells. Tables II and VI show the obtained results with S-MASA algorithm, random walk and S-random walk respectively when varying the number of agents from 5 to 30 in free-obstacle and obstacle environment where obstacles are uniformly distributed (2 *(e), (f)*). Figure 8 represents a comparison between these strategies according to Tables II and VI. Our algorithm performs much better than the random walk and the S-random walk, when the number of agents is less than 15. The results of the three strategies are close when the number of agents is more than 15, but our algorithm gives the best results.

TABLE VI: Effect of agent number on performance in random walk and S-random walk

	5	10	15	20	25	30
random walk free	2536,3	1365,65	932,6	567,05	508,7	487,3
STD Deviation	2021,98	1014,65	811,29	271,43	242,56	340,66
random walk obs	673,5	313,6	218,35	164,7	111,45	105,35
STD Deviation	853,37	223,99	207,19	83,20	43,07	34,95
S-random walk free	320	171	120	89	80	68
STD Deviation	354,25	206,7	164,3	138,6	126,25	111,55
S-random walk obs	798,7	543,1	306,7	205,7	182	157,2
STD Deviation	653,18	343,38	87,74	129,71	99,00	66,57

Tables III and VII show the obtained results with S-MASA algorithm, random walk and S-random walk respectively when varying the size of the environment from 20 X 20 cells to

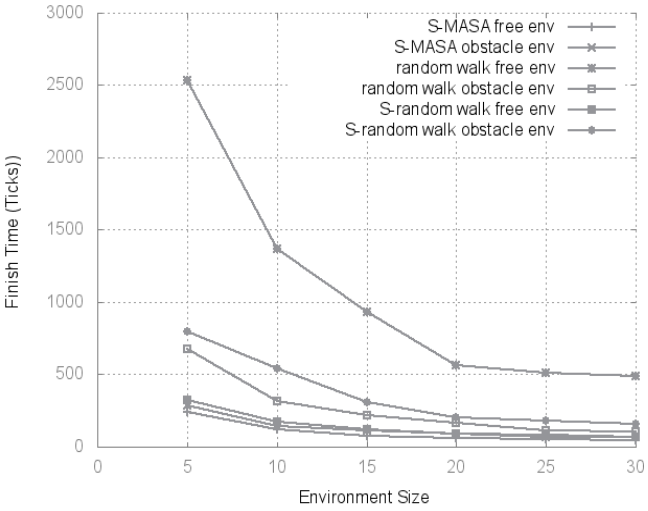


Fig. 8: Comparison of S-MASA with random walk and S-random walk when varying the number of agents (uniform distribution of obstacles)

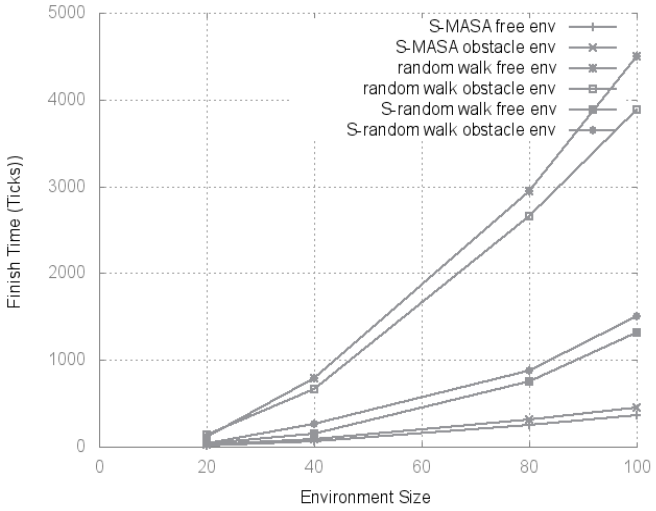


Fig. 9: Comparison of S-MASA with random walk and S-random walk when varying the environment size (uniform distribution of obstacles)

100 X 100 cells in free-obstacle and obstacle environment where obstacles are uniformly distributed (2 *(e), (f)*). Figure 9 represents a comparison between these strategies according to Tables III and VII. When the size of the environment is equal to 20 X 20 cells , the three strategies provide very close results. By increasing the environment size from 40 X 40 cells to 100 X 100 cells, our algorithm outperforms the two others.

TABLE VII: Effect of environment size on performance in random walk and S-random walk

	20X20	**40X40**	**80X80**	**100X100**
random walk free	108,2	789,65	2946,8	4501,85
STD Deviation	81,08	625,99	1398,71	2169,28
random walk obs	137,85	665,45	2651,5	3889,9
STD Deviation	68,93	376,37	1616,13	2307,50
S-random walk free	37,05	149,8	754	1312,25
STD Deviation	10,21	38,39	378,31	546,87
S-random walk obs	39,65	261,5	882,4	1502,75
STD Deviation	26,12	141,18	529,19	830,73

A comparison between the three strategies when varying the number of agents from 5 to 30 in an obstacle cluster environments (Figure 2 *(g), (h)*) is given in Figure 10, where our algorithm proves its performance among pure random walk and S-random walk. Table VIII shows the obtained results and the standard deviation in each simulation for the three strategies.

TABLE VIII: Effect of number of agent on performance in S-MASA, random walk and S-random walk

	5	**10**	**15**	**20**	**25**	**30**
S-MASA	242,85	122,2	78,85	63,5	54,8	43,9
STD Deviation	46,62	24,84	17,87	14,15	11,35	10,15
random walk	3472,95	1393,05	930,85	639,7	474,05	370,45
STD Deviation	2440,36	435,20	542,51	301,39	237,19	144,05
S-random walk	886,95	428,95	319,9	188,1	149,65	106,95
STD Deviation	717,16	230,36	398,51	96,82	79,94	39,23

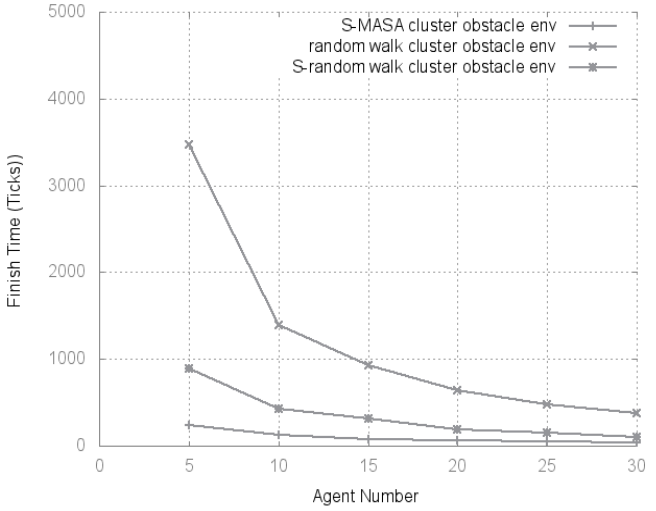


Fig. 10: Comparison of S-MASA with random walk and S-random walk when varying the number of agents (cluster obstacle environment)

VII. Conclusion

A multi-target search algorithm called S-MASA is presented in this paper. This algorithm reduces overall finish time without any direct communication between agents. Simulation results demonstrate the higher performance of our algorithm in comparison with the S-random walk which is guided by pheromones to repulse agents from visited areas and with random walk strategy. We believe that future work improvements should reduce searching time, consider more complex, dynamic and unknown environments in the context of foraging problem.

References

[1] S. Sarid, A. Shapiro, and Y. Gabriely, "Mrsam: A quadratically competitive multi-robot online navigation algorithm," in *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*. DOI: 10.1109/ROBOT.2006.1642109, IEEE, 2006, pp. 2699–2704.

[2] E. Acar, H. Choset, Y. Zhang, and M. Schervish, "Path planning for robotic demining: Robust sensor-based coverage of unstructured environments and probabilistic methods," *International Journal of Robotics Research 22(7-8)*, pp. 441–466, 2003.

[3] D. Gage, "Many-robot mcm search systems," in *Autonomous Vehicles in Mine Countermeasures Symposium, vol. 9*. DOI: 10.1.1.38.771, 1995, pp. 56–64.

[4] G. Kantor, S. Singh, R. Peterson, D. Rus, A. Das, V. Kumar, and G. Pereira, "Distributed search and rescue with robot and sensor teams," in *Field and Service Robotics*. DOI: 10.1007/10991459-51, Springer Berlin Heidelberg, 2006, pp. 529–538.

[5] J. Jennings, G. Whelan, and W. Evans, "Cooperative search and rescue with a team of mobile robots," in *8th International Conference on Advanced Robotics, ICAR*. DOI: 10.1109/ICAR.1997.620182, IEEE.

[6] A. Marjovi, J. unes, L. Marques, and A. de Almeida, "Multi-robot exploration and fire searching," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*. DOI: 10.1109/IROS.2009.5354598, IEEE, 2009, pp. 1929–1934.

[7] Landis and A. Geoffrey, "Robots and humans: Synergy in planetary exploration," in *SPACE TECHNOLOGY AND APPLICATIONS INT. FORUM-STAIF 2003: Conf. on Thermophysics in Microgravity; Commercial/Civil Next Generation Space Transportation; Human Space Exploration; Symps. on Space Nuclear Power and Propulsion (20th); Space Colonization (1st)*, vol. 654, no. 1. DOI: org/10.1063/1.1541377, AIP Publishing, 2003, pp. 853–860.

[8] K. Schilling and C. Jungius, "Mobile robots for planetary exploration," *Control Engineering Practice*, vol. 4, no. 4, pp. 513–524, 1996.

[9] M. A. Batalin and G. S. Sukhatme, "Spreading out: A local approach to multi-robot coverage," in *Distributed autonomous robotic systems 5*. DOI: 10.1007/978-4-431-65941-9-37, Springer, 2002, pp. 373–382.

[10] C. A. Pina-Garcia, G. Dongbing, and H. U. Huosheng, "A composite random walk for facing environmental uncertainty and reduced perceptual capabilities," in *Intelligent Robotics and Applications*. DOI: 10.1007/978-3-642-25486-4-62, Springer, 2011, pp. 620–629.

[11] O. Feinerman, A. Korman, Z. Lotker, and J. S. Sereni, "Collaborative search on the plane without communication," in *Proceedings of the 2012 ACM symposium on Principles of distributed computing*. DOI: 10.1145/2332432.2332444, ACM, 2012, pp. 77–86.

[12] R. L. Stewart and R. A. Russell, "A distributed feedback mechanism to regulate wall construction by a robotic swarm," *Adaptive Behavior*, vol. 14, no. 1, pp. 21–51, 2006.

[13] M. J. B. Krieger, J. B. Billeter, and L. Keller, "Ant-like task allocation and recruitment in cooperative robots," *Nature*, vol. 406, no. 6799, pp. 992–995, 2000.

[14] I. Roman-Ballesteros and C. PfeifferF, "A framework for cooperative multi-robot surveillance tasks," in *Electronics, Robotics and Automotive Mechanics Conference, 2006*, vol. 2. DOI: 10.1109/CERMA.2006.3, IEEE, 2006, pp. 163–170.

[15] M. Schwager, D. Rus, and J. J. Slotine, "Decentralized, adaptive coverage control for networked robots," *The International Journal of Robotics Research*, vol. 28, no. 3, pp. 357–375, 2009.

[16] J. C. dn S. Martinez, T. Karatas, and F. Bullo, "Coverage control for mobile sensing networks," in *IEEE International Conference on Robotics and Automation, ICRA'02*, vol. 2. DOI: 10.1109/ROBOT.2002.1014727, IEEE, 2002, pp. 1327–1332.

[17] S. Sarid, "Heterogeneous multi-robot search algorithms," Ph.D. dissertation, Ben-Gurion University of the Negev, 2011.

[18] Y. Gabriely and E. Rimon, "Spanning-tree based coverage of continuous areas by a mobile robot," *Annals of Mathematics and Artificial Intelligence*, vol. 31, no. 1-4, pp. 77–98, 2001.

[19] R. Meir, Y. Peleg, and N. Pochter, "lecture 2," in *Mathematical Foundation of AI*, 2008.

[20] Y. Gabriely and E. Rimon, "Spiral-stc: An on-line coverage algorithm of grid environments by a mobile robot," in *Proceedings IEEE International Conference on Robotics and Automation, ICRA'02.*, vol. 1. DOI: 10.1109/ROBOT.2002.1013479, IEEE, 2002, pp. 954–960.

[21] Y. Gabriely and E. Rimonn, "Competitive on-line coverage of grid environments by a mobile robot," *Computational Geometry*, vol. 24, no. 3, pp. 197–224, 2003.

[22] N. Hazon and G. A. Kaminka, "On redendency, efficiency and robustness in coverage for multi-robot," *Autonoumous System 56*, 2008.

[23] N. Agmon, N. Hazon, and G. A. Kaminka, "Constructing spanning trees for efficient multi-robot coverage," in *IEEE International Conference on Robotics and Automation, ICRA 2006*. DOI: 10.1109/ROBOT.2006.1641951, IEEE, 2006, pp. 1698–1703.

[24] X. Deng and A. Mirzaian, "Competitive robot mapping with homogeneous markers," *Robotics and Automation, IEEE Transactions on*, vol. 12, no. 4, pp. 532–542, 1996.

[25] G. Dudek, M. Jenkin, E. Milios, and D. Wilkes, "Robotic exploration as graph construction," *Robotics and Automation, IEEE transactions on*, vol. 7, no. 6, pp. 859–865, 1991.

[26] M. Bender, A. Fernandez, A. S. D. Ron, and S. Vadhan, "The power of a pebble: Exploring and mapping directed graphs," in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. DOI: 10.1145/276698.276759, ACM, 1998, pp. 269–278.

[27] I. A. Wagner, M. Lindenbaum, and A. M. Bruckstein, "Distributed covering by ant-robots using evaporating traces," *Robotics and Automation, IEEE Transactions on*, vol. 15, no. 5, pp. 918–933, 1999.

[28] I. A. Wagner and A. M. Bruckstein, "From ants to a (ge) nts: A special issue on ant-robotics," *Annals of Mathematics and Artificial Intelligence*, vol. 31, no. 1, pp. 1–5, 2001.

[29] D. C. V. Méndez and F. Bartumeus, "Random search strategies," *Stochastic Foundations in Movement Ecology, Springer-Verlag Berlin Heidelberg*, pp. 177–205, 2014.

[30] M. F. R. Calvo, J. R. de Oliveira and R. A. F. Romero, "Bio-inspired coordination of multiple robots systems and stigmergy mechanims to cooperative exploration and surveillance tasks," in *Cybernetics and Intelligent Systems (CIS), 2011 IEEE 5th International Conference on*. DOI: 10.1109/ICCIS.2011.6070332, IEEE, 2011, pp. 223–228.

[31] I. T. T. Kuyucu and K. Shimohara, "Evolutionary optimization of pheromone-based stigmergic communication," in *Applications of Evolutionary Computation*. DOI: 10.1007/978-3-642-29178-4-7, Springer, 2012, pp. 63–72.

[32] C. Lenzen and T. Radeva, "The power of pheromones in ant foraging," in *1st Workshop on Biological Distributed Algorithms (BDA)*, 2013.

[33] A. L. D. Liu, X. Zhou and H. Guan, "A swarm intelligence based algorithm for distribute search and collective cleanup," in *Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on*, vol. 2. DOI: 10.1109/ICICISYS.2010.5658776, IEEE, 2010, pp. 161–165.

[34] S. K. Ghosh and R. Klein, "Online algorithms for searching and exploration in the plane," *Computer Science Review*, vol. 4, no. 4, pp. 189–201, 2010.

[35] G. H. Orians and N. E. Pearson, "On the theory of central place foraging," *Analysis of ecological systems. Ohio State University Press, Columbus*, pp. 155–177, 1979.

[36] T. H. R. Fujisawa, H. Imamura and F. Matsuno, "Communication using pheromone field for multiple robots," in *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*. DOI: 10.1109/IROS.2008.4650971, IEEE, 2008, pp. 1391–1396.

[37] R. Johansson and A. Saffiotti, "Navigating by stigmergy: A realization on an rfid floor for minimalistic robots," in *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*. DOI. 10.1109/ROBOT.2009.5152737, IEEE, 2009, pp. 245–252.

[38] T. Sakakibara and D. Kurabayashi, "Artificial pheromone system using rfid for navigation of autonomous robots," *Journal of Bionic Engineering*, vol. 4, no. 4, pp. 245–253, 2007.

[39] L. M. A. F. V. A. Ziparo, A. Kleiner and D. Nardi, "Cooperative exploration for usar robots with indirect communication," in *Proc. of 6th IFAC Symposium on Intelligent Autonomous Vehicles, IAV, DOI: 10.3182/20070903-3-FR-2921.00094*, 2007.

[40] G. W. B. Ranjbar-Sahraei and A. Nakisaee, "A multi-robot coverage approach based on stigmergic communication," in *Multiagent System Technologies*. DOI: 10.1007/978-3-642-33690-4-13, Springer, 2012, pp. 126–138.

[41] U. Wilensky, "Netlogo. http://ccl.northwestern.edu/netlogo/,," in *Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL*, 1999.

3rd International Workshop on Smart Energy Networks & Multi-Agent Systems

The emerging smart infrastructure in energy networks represents a major paradigm shift in resource allocation management with the aim to extend the centralised supply management model, towards a decentralised supply-and-demand management that is expected to enable more efficient, reliable and environment-friendly utilisation of primary energy resources.

Together with this vision, there are new and complex tasks to manage, in order to ensure safe, cost-reducing and reliable energy network operations. This includes the integration of various renewable energy systems, like the photovoltaic or the wind energy, which are able to reduce the greenhouse gas emissions but that are working under greater uncertainty; as well as the interaction of transport and storage systems for energy that are envisioned through techniques like 'Power to Gas' and fuel cells, which are using the electrical and the gas transportation network.

Further tasks can be found in the fact that the market participants (e.g. simply households) are becoming more autonomous and intelligent through technologies like smart metering, which requires a coordinated demand side management for millions of producers, consumers or, if this applies, prosumers by means negotiations and agreements.

Information and communication technologies are key enablers of the envisioned efficiencies, both on the demand and the supply sides of the smart energy networks, where the agent-paradigm provides an excellent first modelling approach for the distributed characteristic in energy supply systems. On the demand side they aim at supporting end-users in optimising their individual energy consumption, e.g. through the deployment of smart meters providing real-time usage and cost of the energy and the use of demand-response appliances that can be controlled according to the user preferences, energy cost and carbon footprint. On the supply side they aim at optimising the network load and reliability of the energy provision, e.g. through active monitoring and prediction of the energy usage patterns, and proactive control and management of the reliable energy delivery over the networks. It is also envisaged that they will be able to influence the demand through the dynamic adjustments of the energy price in order to influence the end user behaviour and energy usage patterns throughout and across the energy networks for electricity, gas and heat.

Although a significant effort and investment have been already allocated into the development of smart grids, there are still significant research challenges to be addressed before the promised efficiencies can be realised. This includes distributed, collaborative, autonomous and intelligent software solutions for simulation, monitoring, control and optimization of smart energy networks and interactions between them.ion plays a crucial role that needs to be further investigated.

Topics

The SEN-MAS 2014 Workshop aims at providing a forum for presenting and discussing recent advances and experiences in building and using multi-agent systems for modelling, simulation and management of smart energy networks. In particular, it includes (but is not limited to) the following topics of interest:

- Experiences of Smart Grid implementations by using MAS
- Applications of Smart Grid technologies
- Management of distributed generation and storage
- Islands Power Systems, Microgrid Applications
- Real time configurations of energy networks
- Distributed planning process for energy networks by using MAS
- Self-configuring or self-healing energy systems
- Load modelling and control with MAS
- Simulations of Smart Energy Networks
- Software Tools for Smart Energy Networks
- Energy Storage
- Electrical Vehicles
- Interactions and exchange between networks for electricity, gas and heat
- Stability in Energy Networks
- Distributed Optimization in Energy Networks

Event Chairs

Derksen, Christian, University Duisburg-Essen, Germany

Kowalczyk, Ryszard, Swinburne University of Technology, Australia

Steering Committee

Derksen, Christian, University Duisburg-Essen, Germany

Lehnhoff, Sebastian, OFFIS - Institute for Information Technology, Germany

Kowalczyk, Ryszard, Swinburne University of Technology, Melbourne, Victoria, Australia

Nahorski, Zbigniew, Systems Research Institute - Polish Academy of Science, Poland

Program Committee

Andrew, Lachlan

Braubach, Lars, University of Hamburg, Germany

Dillon, Tharam, Curtin University, Australia

Elammari, Mohamed, University of Benghazi, Libya

Franczyk, Bogdan, Universitat Leipzig, Germany

Guttmann, Christian, Monash University, Australia

Klusch, Matthias, German Research Center for Artificial Intelligence, DFKI, Germany

Linnenberg, Tobias, Helmut Schmidt University, Germany

Maamar, Zakaria, Zayed University, United Arab Emirates

Moench, Lars, FernUniversität Hagen, Germany

Monti, Antonello

Ossowski, Sascha, University Rey Juan Carlos, Spain

Palensky, Peter

Schmeck, Hartmut, Karlsruhe Institute of Technology (KIT), Germany

Sonnenschein, Michael, Carl von Ossietzky University, Oldenburg, Germany

Sudeikat, Jan, Hamburg Energie GmbH, Germany

Tianfield, Huaglory, Glasgow Caledonian University, United Kingdom

Unland, Rainer, Universität Duisburg-Essen, Germany

Wang, Shuliang

Weber, Christoph

Weidlich, Anke

Werner, Andrej, University of Leipzig, Germany

Wong, Dennis, Swinburne University of Technology Sarawak Campus, Malaysia

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1489–1496
DOI: 10.15439/2014F251
ACSIS, Vol. 2

Multi-Agent-based Distributed Optimization for Demand-Side-Management Applications

Tim Dethlefs, Thomas Preisler, Wolfgang Renz
Hamburg University of Applied Sciences
Faculty of Engineering and Computer Science
Berliner Tor 7, 20099 Hamburg, Germany
Email: {tim.dethlefs | thomas.preisler | wolfgang.renz}@haw-hamburg.de

***Abstract*—Dynamic and volatile grid conditions caused by the growing amount of renewable energy producers require the operation of large-scale distributed Demand-Side Management (DSM) applications. This is one of the tasks of the aggregator role in smart grid operation according to the Smart Grid Architecture Model (SGAM). For the optimization of distributed demand-side loads under such conditions, Multi-Agent Systems (MAS) have been shown to provide an appropriate paradigm to model, simulate and deploy automated operating components.**

In this paper, we address an engineering problem that is still a matter of concern, namely the construction of efficient distributed optimization algorithms in conjunction with a generic software architecture. For this purpose, a distributed Multi-Agent architecture is presented with a generic consumer model and an energy exchange market as well as further roles and components. Ant Colony System Optimization is shown to effectively optimize consumers in a nature-inspired, self-organizing way.

The applicability of the proposed approach will be demonstrated in a use-case study where a group of heterogenous consumers optimize their runtimes in order to map their demand to the energy generation of a wind power plant in a self-organized fashion.

I. Introduction

THE DEVELOPMENT of small, affordable and profitable energy generators as well as the desired growth of renewable energy resources leads to an increasing decentralization and heterogeneity in the smart grid. The capacity and availability of the decentralized energy resources often depends on environmental influences so fast-reacting conventional power plants and energy storages will be needed to compensate such fluctuations.

The utilization of the demand-side potential, the *Demand-Side-Management* (DSM), could be an additional planning option which becomes achievable and affordable through the increasing degree of automation and information- and communication technology (ICT).

The management of both demand and production side can be more efficient and safer with planning and forecasting, but will require predictable and intelligent devices and appliances for domestic households as well as industrial applications. Those devices should be able to coordinate and optimize their schedule in order to achieve pre-qualification for markets or to lower their runtime costs. While in the energy domain no real quality of service, like frequency- or availability services, can be sold, demand-side orientated business models must focus on the flexibility and planning potential of the devices with dynamic pricing or stock exchange models. The scheduling and optimization of the highly heterogeneous and distributed devices needs adaptive solutions. It seems to be likely to use these intelligent embedded devices themselves for this task, because of their computational- and communication-capacities. So a lightweight and simple optimization algorithm and a distribution concept with a minimum of shared information are required.

Because of the distribution- and autonomy-properties Multi-Agent-Systems (MAS) are suitable as a paradigm for the logical representation of grid entities. Even the current top-down modus operandi of the grid could be considered as a distributed MAS-architecture regarding the distribution of loads, substations etc. Therefore, the architectural approach in this paper utilizes the MAS-paradigm towards the distributed optimization of consumers for DSM applications.

The remaining paper is structured as follows: Section II provides an overview on the related literature and covers different aspects of todays grid development towards an emerging DSM integration. In Section III a description of the distributed MAS-Architecture for DSM and optimization is provided. Section IV describes the reference implementation of an adapted Ant Colony System - Algorithm for distributed optimization. In Section V the applicability of the proposed approach is demonstrated in a use-case study where a group of heterogenous consumers optimize their runtimes in order to map their demand to the energy generation of a wind turbine in a self-organized fashion, before Section VI concludes the paper and gives an overview about future work.

II. Related Work

Energy supply particularly on the field of electricity currently experiences a profound change. Due to the availability and the versatile ways of use, electricity has become one of the most important energy sources. New applications and potential uses as well as laws, standards, guidelines and social developments generate new requirements for the energy generation, the grid-infrastructure and also the consumer-side. This Section introduces current laws and standards of the energy domain, it gives a brief overview about the situation on the energy markets and presents current grid-infrastructures

978-83-60810-58-3/$25.00

before an overview about related work on DSM and MAS-Simulation concludes the Section.

A. Laws and Standards

In the USA the development of energy distribution has mainly arisen from the oil crisis. Until today the *National Energy Conservation Policy Act* (NECPA) from 1978 regulates the energy consumption in the United States. Contrasting to California, where the electricity-demand has been stabilized by the 1972 decree on device efficiency, building regulation and energy efficiency for energy suppliers, which was manifested as a law in 2006 [1], the power consumption in the rest of the USA has risen steadily [2]. This led to the 2007 *Energy Independence and Security ACT* (EISA) as an extension of the NECPA. In 2000 Germany applied a law[1] to regulate the favored input of electricity from renewable energy resources and to guarantee a fixed buyback price for the next 20 years. Until 2020 the amount of electricity generated from renewable resources should have reached at least 35%, rising up to at least 80% in 2050. In 2013 this amount had already reached 23,4% [3]. Because of the volatility of renewable energy resources, the favorable input of such unsteady resources conflicts with the classical modus operandi and requires highly dynamic and regulable power plants to compensate energy deficiencies respectively the establishment of corresponding storage capacities as required by the Energy Industry Act[2]. This law demands that the security and reliability of the energy grid is secured through net- or market-based measures in case of disturbances or compromises. Besides the duties on the producer side, the law also permits the detailed acquisition of performance data and load balancing measurements on the consumer side through load- and time-depended tariffs.

In order to deal with this situation almost all parts of the grid from the producer to the consumer must be automated in the future. This requires a new role- and domain-model as defined by the *NIST-Framework-Roadmap* [4]. It is established as a quasi-standard for further publications in this area. European organizations and research facilities have adopted the model and extend it for local requirements, e.g. with the extension of the distributed energy resources (DER) domain as part of the *European Smart Grid Reference Architecture* and the *Smart Grid Architecture Model* (SGAM), a layered-architecture for smart grids representing the communication and information layers [5].

B. Energy Markets

Today's electricity market in Europe is highly regulated in order to maintain operational safety. Due to the big trade volume it is mainly accessible for large producers and consumers. The most important energy markets in Europe are the *European Energy Exchange* (EEX)[3] respectively the *European Power Exchange* (EPEX SPOT)[4] as power-based spot market located in Leipzig, Germany. In 2011 the German net-agency (Bundesnetzagentur BNetzA) reformed the bidding conditions for secondary and tertiary control by reducing the minimum submission size, enabling aggregation and allowed deliverance guarantees by third parties. Through the low investment risk and short amortization times both small and medium-sized energy producers as well as DER have gained increasing attractiveness in recent years [6]. But not without creating new challenges: Due to the increasing distribution of energy producers their coordination becomes more and more complex. Also because of their low production rate DER are not able to participate directly at the energy markets. A possible solution is the grouping of DER to larger entities, the so called *Virtual Power Plants* (VPP) [7], [8].

C. Communication Infrastructure

Many projects from the energy domain facilitate internet-technology like the TCP/IP protocol as communication infrastructure for their applications because of the well-defined standards and their wide dissemination [9]. Another interesting infrastructure approach for domestic household automation is given by the *OGEMA-Project* [10]. It deals with the connection of household devices via a gateway-interface to a coordination centre and allows the automated control of consumers depending on variable power prices. Alongside to the OGEMA-project, *OpenADR* is an established system for automated demand-response, focusing on top-down business models controlled by a system operator [11]. According to [9] there is still a lack of a common communication standard for demand-side management. [12] points out the need for a registry for DER and consumers which provides information about all available energy services. Therefore, the area-wide facilitation of DSM capacities is still a topic of current research areas.

D. Demand-Side Management and Multi-Agent Systems

Demand-Side Management (DSM) is defined by [13] as planning and implementation activities of utilities and operators in order to influence the consumer's demand so that the intended consumer-behavior is achieved. A study of the US *Federal Energy Regulatory Commission* [14] comes to the conclusion, that in the USA the peak-load will rise from 775 GW in 2009 to 900 GW in 2019. That would require 2000 new power plants. Here DSM is seen as a measurement to limit this growth to 800 GW and to stabilize it in order to deal with the increasing costs and environmental impacts [15]. Apart from energy savings also the purposeful use of volatile effects from the renewable energy resources is a significant DSM use-case in Germany [9].

Multi-Agent Systems (MAS) are an established technology in order to simulate different DSM applications in smart grids. Due to their distribution and autonomy properties they can be developed relatively similar to real infrastructures and grid entities. Almost all approaches share the fact, that agents are used to represent consumers or net-entities. Most approaches utilize the demand-side for feedback control and secondary

[1]Gesetz für den Vorrang Erneuerbarer Energien (EEG).
[2]Energiewirtschaftsgesetz (EnWG)
[3]www.eex.com
[4]www.epexspot.com

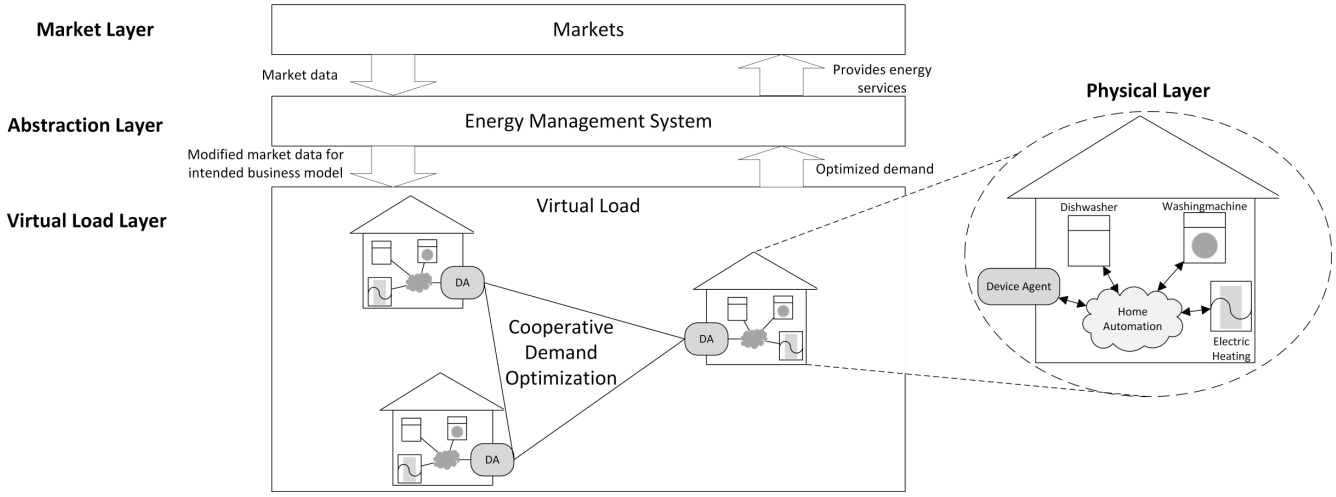


Fig. 1. Architectural model for the distributed optimization system.

and tertiary control energies. Hereby, the examined time frame ranges from near real-time applications as presented in [16] to applications like *PowerMatcher* [6] which focus more on day ahead planning. The simulation systems mainly differ on how the DSM is achieved. Here it can be differentiated between direct and indirect control. Direct control focuses on the used protocols and signaling, whereas indirect control approaches can be further differentiated between approaches based on price signals, auctions and optimization on the consumer-side. An example for the direct control of load is presented in [17] where the joint control of loads is researched. The advantage of this approach is the reliable system response for defined signals, as the connected consumers have to act accordingly to specified signaling. Additionally, the according service provider or operator has complete control over the loads. But this mode of operation implicates hard constraints on the consumer-side concerning the conditions under which load can be switched centrally. Thereby, the coordination centre requires detailed information about the controllable loads and time frames, in order to anticipate the effects of the according control signals. This approach seems to be more interesting for industrial consumers due to the complexity and heterogeneity of domestic household devices and appliances.

In the area of indirect load control very different approaches have been researched. Close to direct control is the use of price signals as described by [18]. It is based on the familiar top-down communication of price signals through a central control centre and therefore, a consequent enhancement of existing tariff models. But simple price minimization may cause peak-loads for automated consumers. Therefore, a real time observation is required so that the operator is able to counteract. Common are also auction-based approaches, where the consumer either bargains directly with the distributed energy resources or via aggregation agents [6], [16], [19]. All this approaches have in common, that the generator and load respectively the prices are tuned and negotiated in a bilateral way. A common challenge hereby is the a priori agreement on price barriers and acceptance levels for each DER or load through the producers and consumers. The last class of approaches focuses on the optimization on the consumer-side in order to generate an optimal degree of efficiency related to the according business- or operational-model. Both [20] and [21] describe exemplary use-cases with regard to specific business-models.

The usage of MAS for the aggregation of DERs to microgrids, modern, small-scale versions of the centralized electricity systems is described in [22], [23]. In [22] the aggregation of DERs and loads together to an autonomous entity a microgrid is described. Here a MAS approach as a branch of distributed artificial intelligence methods is introduced for DERs control in the microgrid. Similar work has been undone in [23] where a MAS for energy resource scheduling of an islanded power systems with DER is presented. It monitors, controls and operates an energy system consisting of a set of microgrids and lumped loads.

An interesting overview about the lastest applications of MAS in the smart grid context is given in [24].

III. Distributed Multi-Agent Architecture

Demand-Side applications usually contain large numbers of dynamic and inhomogeneous loads and thus require a distributed architecture with loose-coupled entities and generic models in order to handle the system-immanent complexities. Multi-Agents-Systems are a good metaphor for modeling and simulating such distributed grid components.

The abstract MAS architecture designed for such demand optimization problems shown in Figure 1 consists of four layers. The first layer is the market layer containing exchange markets for generation and energy services (see [4]). On a physical layer, each domestic household acts like an agent, monitoring the own environment through sensors, trying to achieve interests and goals (like a low energy price), while handling the system immanent restrictions (e.g. maintaining

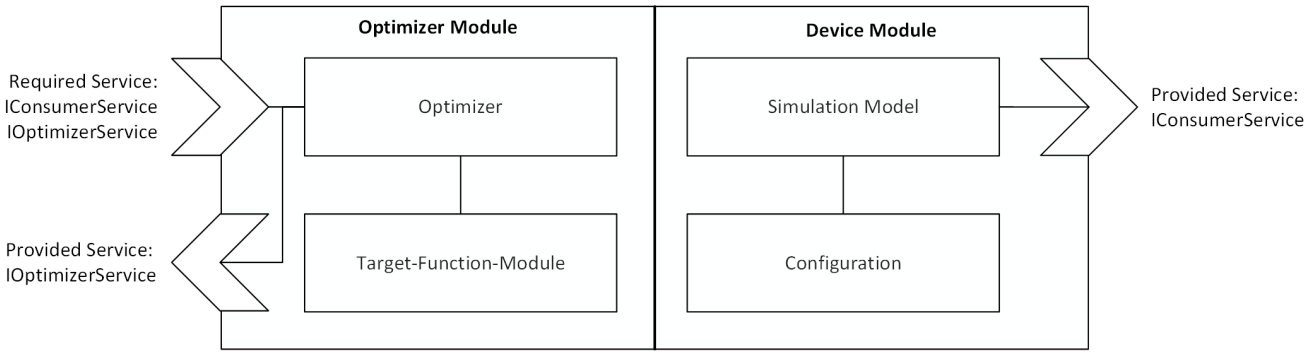


Fig. 2. Modular structure of the Device-Agent.

TABLE I
CLASSIFICATION OF LOADS

Class	Description	Control	Examples
User-driven loads	Loads that satisfy the users demand directly	User-controlled	Light, TV
Program-driven loads	The user starts the device but it may not run immediately	semi-automatic	Washing machines, dishwasher
Fully-automated loads	These devices have actors and sensors to maintain a certain state	Automatic, parameter driven	Electric heating or cooling (fridge)

a certain temperature in the household). This behavior, represented through Device-Agents, could be used by the Energy Management System as a virtual (and distributed) load to provide grid services to the market.

The **Device-Agent** (see Figure 2) is the most important entity in this energy system. Its role is to make the underlying physical layer transparent for the optimization system. With the Device-Agent as a gateway interface to the smart grid, it could abstract $1 : n$ physical devices and appliances, providing the accumulated DSM-potential of the underlying physical layer (like load-shifting). The Device-Agent needs actors and sensors to control the physical devices as well as an intrinsic simulation model in order to anticipate the runtime constraints of the physical devices connected to it.

The load of the Device-Agent is described by:

$$L\{\tau_1, \tau_2, \ldots, \tau_{n_m}\} = \sum_{i=0}^{n_m} \mathcal{L}(t - \tau_i), \tag{1}$$

with n_m as the variable number of starting times for the device, $\mathcal{L}(t-\tau_i)$ as the load of the agent, and τ_i as the discrete starting times of the loads.

An optional second module of the Device-Agent is the optimization module for the distributed, cooperative optimization of the loads. It implements a distributed lightweight optimization algorithm (described in Section IV) that optimizes the starting time of all participating Device-Agents utilizing the described iterative planning approach.

As stated in Table I loads could be classified into three groups [17]. The first class is only relevant for long term DSM efforts like strategic conservation [13]. The second and third classes can be planned iteratively using the provided *IConsumerService* of the related Device-Agent. An optimizer requests the first possible starting time vector from a Device-Agent and receives it as an integer array, then chooses one starting time and commits it back to the Device-Agent. Depending on the chosen starting times the Device-Agent calculates the next possible runtime-vector with the simulation model. This will be repeated until the Device-Agent returns an *end-of-planning*-flag. Program-driven devices require just one planning step, as the first vector contains all possible starting times from the earliest to the latest. Under the assumption that every planning step of a fully-automated-device depends on the previous condition of it, the iterative planning provides a generic way to plan such devices. Thus, the optimizer is able to plan the device without domain-specific knowledge and the whole simulation and planning logic remains at the device for security reasons and separation of concern.

The **Energy-Management-Agent (EMA)** is the domain specific representation of the control component. It is responsible for planning and operation tasks and therefore, requires knowledge about the intended *modus operandi* and the current status of the associated Device-Agents. The role of the EMA in the Smart Grid could be a versatile one. For example, the EMA could aggregate Device-Agents all over the balancing zone for secondary or tertiary control. Focusing more on local business models, the EMA may optimize the load of domestic households towards the available power provided by DERs in a local grid.

The infrastructure domain is subdivided into two parts: The infrastructure represented by a simple TCP/IP-network supporting the service-based communication of the agents and a **Registry-Agent**. The latter one provides and manages information about the Device-Agents to the EMA as an additional ancillary service, so it is able to find suitable Device-Agents

for the intended business model.

The **Energy Market** is a reactive agent or entity that provides a platform for an energy exchange. Here DERs can offer their generated power with the according prices, so consumers can evaluate and accept the offers.

At application runtime each Device-Agent has to register at the Registry-Agent giving information about its runtime capabilities, consumption capability, optional optimizer capability and some additional information (geographical or logical position in the grid). A User (Operator/Service Provider) defines the business plan parameters of the EMA i.e. the user group, markets or system targets. The EMA requests suitable Device-Agents at the Registry, reserves the control capability of the Device-Agents, defines a subset of optimizers out of the set of Device-Agents and parametrizes them (e.g. announces which Device-Agents are in the user group, which are the active/optimizing agents and which the passive ones). The optimizing Device-Agents iteratively plan the runtimes of all participating Device-Agents in the group and transmit the result to the EMA. The EMA then parametrizes every devices with its planed runtime.

IV. Implementation of a distributed optimization algorithm

In order to optimize the starting times of the devices, every starting time vector submitted by a Device-Agent could be interpreted as a set of nodes in a directed graph (see Figure 3). Thus, each optimizing Device-Agent receives the same ordered list of Device-Agents to be optimized from the EMA, so they are all traversing the identical graph. The edges of this graph are weighted with the operational price.

This interpretation of the problem domain allows to apply several meta-heuristics for combinatorial optimization. Furthermore, to enable the consumer-side and to use the potential of the embedded-systems, the applied algorithms must be distributable, lightweight and scalable for different environmental conditions. Thus, Meta-heuristics like Evolutionary Algorithms (EA) with their population-based concepts seem to be suitable for such load planning problems.

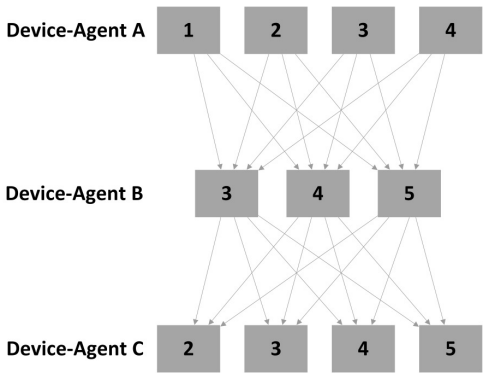


Fig. 3. Runtime vectors of three Device-Agents (A, B, C) interpreted as graph

A. Definition of the optimization problem

The optimization problem for load planning based on a power exchange market can be defined as the cost-minimization of the whole system:

$$I = min \sum_{t=0}^{T} \sum_{j=1}^{A} I_j(t) P_j(t), \tag{2}$$

where T is the number of time slots in the optimization period. Typically it ranges between 96 for quarter hour time slots and 1440 for minute-based time slots. A denotes the number of offers for the prices $I_j(t)$, and $P_j(t)$ the amount of power bought from that offer. A constraint is: $P_j(t) \leq P_{j,MAX}(t)$, meaning that the power bought may never exceed the offered amount.

The bought power is defined as:

$$\sum_{j=1}^{A} P_j(t) := L_N\{\tau_n\}(t) \tag{3}$$

with

$$L_N\{\tau_n\}(t) := \sum_{m=0}^{M} L_m(\tau_1, \tau_2, \ldots, \tau_{n_m})(t) \tag{4}$$

where M is the total number of devices and $L_N\{\tau_n\}(t)$ as the total consumption over all devices participating.

Propagating this problem to the virtual load by the EMS, the Device-Agents of the virtual load try to find cooperatively an optimal scheduling for price-minimization.

B. The distributed ACS

As a reference implementation an adapted version of the Ant Colony System (ACS, see [25]) was used. ACS is an algorithmic metaphor for food-searching ants, placing pheromones on their trail to indicate the shortest path between colony and food-source.

The problem described as a directed graph can be interpreted as a path-finding through the graph, a task ACS originally was designed for. In Figure 5 the global behavior of the optimizing module is illustrated. After the initialization phase, a Device-Agent (Master) with an optimizer module dispatches a certain number of Ant-Agents (Slaves) as its local colony. The local number of Ant-Agents and their serial or parallel execution (see [26]) depends on the properties and restrictions of each participating platform. Besides the nodes, the pheromones were also stored locally at the corresponding Device-Agent. Therefore, each Ant-Agent has to call the Device-Agent's service at least two times, one for the available nodes and one for the corresponding pheromone vector.

An Ant-Agent requests the initial vector of possible starting times from the first Device-Agent in the corresponding ordered list and randomly chooses the first staring time out of the vector and adds it to a temporary vector, before going iteratively through the following steps:

1) It requests the next possible starting time vector and the pheromone vector from the Device-Agent, transmitting the previously chosen starting time.
2) If the vectors are available, the Ant Agent chooses the next node based on the costs of operation (edge-weight) and the according pheromones on the trail.
3) If the Agent receives an *end-of-planning*-Flag it proceeds to the next Device-Agent in the ordered list (if available) and executes step 1.

Every optimizing Device-Agent has its own colony. After every full iteration, e.g. after each of the local Ant-Agent has traversed the graph once, the Master-Agent chooses the best-so-far path and places pheromones at the other Device-Agents along the path through service calls. Because of the different computational speeds of the Device-Agents the pheromone adjustment happens asynchronously. Further synchronizing efforts are not required but fast optimizers may place pheromones more often than slower ones.

In a reference scenario with 100 domestic households, the ACS performs best compared to a basic Ant-System implementation [27] and a greedy algorithm. (see Figure 4).

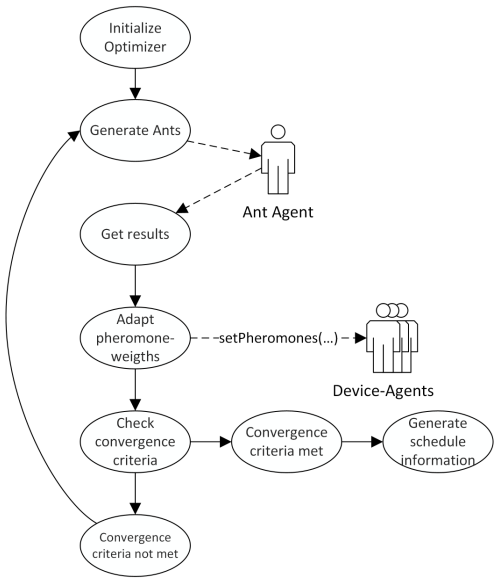


Fig. 5. Behavior of the device-agent optimizing the loads of other device agents.

V. Use-Case: Planning of wind energy resources

A typical use-case for consumer optimization is the day-ahead optimization of predictable and controllable loads for a given load profile. It is often necessary to plan the power consumption according to the optimal utilization of volatile DERs to achieve a higher efficiency. Based on weather forecasts it is in many cases possible to predict the electricity generation capacity of wind turbines with a sufficient accuracy one day ahead. Energy generated by such plans is keen available, as the operation does not require any additional resources and thus the marginal costs are low. Therefore, if the demand is planned accordingly to the generated energy the users may save money through the efficient use of the resource as only a low amount of external (and often expensive) control energy has to be bought in addition.

For this use case example the generated energy of a 330 kW Enercon E-33 wind turbine is used to meet the demand of flexible consumers in a reference scenario. Although the DSM-potential of domestic households is often to be considered relatively low, the results can be easily transferred and compared, as industrial applications have often very specific requirements and constraints. The considered property consists of 100 active domestic households. Accordingly to the dispersion in Germany 92% of them have a washing machine, 62% a dish washer and 4% an electrical heater. This results in 158 program-driven and 4 full-automatic Device-Agents representing and simulating the electrical consumers and appliances. The program-driven agents were statistically parameterized with regard to the *Smart-A Project-Study* [28]. The European-wide mean runtime probability of both washing machines and dish washers based on studies of the University of Bonn (Germany) were evaluated and implemented in the agent models. In order to calculate the runtime flexibility, the usage of power-up delays for the devices as described by the *Smart-A Project-Study* was analyzed. In case of the washing machines 56% of the users chose a power-up delay time period of 0 to 3 hours and 28% a time period of 4 to 6 hours. Therefore, by considering flexibilities up to six hours approx. 84% of the user preferences are considered in this use-case.

The study also shows the general acceptance of power-up delays for these type of devices. The broad degree of utilization alternates between approx. 38% in Sweden up to 81% in Italy in which the actual usage of power-up delays for every stage of washing program goes to approx. 43%. The situation for dish washers is similar. In 2007 39% of them were equipped with a power-up delay capability and 27% of the users were actually using it [28]. It can be assumed that the degree of usage and acceptance will increase further if the distribution of such power-up capabilities grows and more profitable DSM use-case utilizing this function will arise.

The fully-automated-device representing the electrical heater was parametrized with a temperature corridor of 1.0°C around the standard room temperature of 21°C. The heater can be turned on and off again in 15 minute intervals. The outside temperature based on temperature data from the Hamburg's air quality control net for September 15th 2013.

The turbine's power generation was simulated with the Greenius-Tool[5] based on weather data for the same day and scaled to the controllable power demand of 100 households in the reference property. So 10% of the turbine's nominal capacity can be used to meet the requirements of the flexible consumers. The wind power generation simulated here is extremely atypical for the property, where most of the energy is required in the evening hours due to the high amount of program-driven consumers. Therefore, even an optimized

[5]http://freegreenius.dlr.de

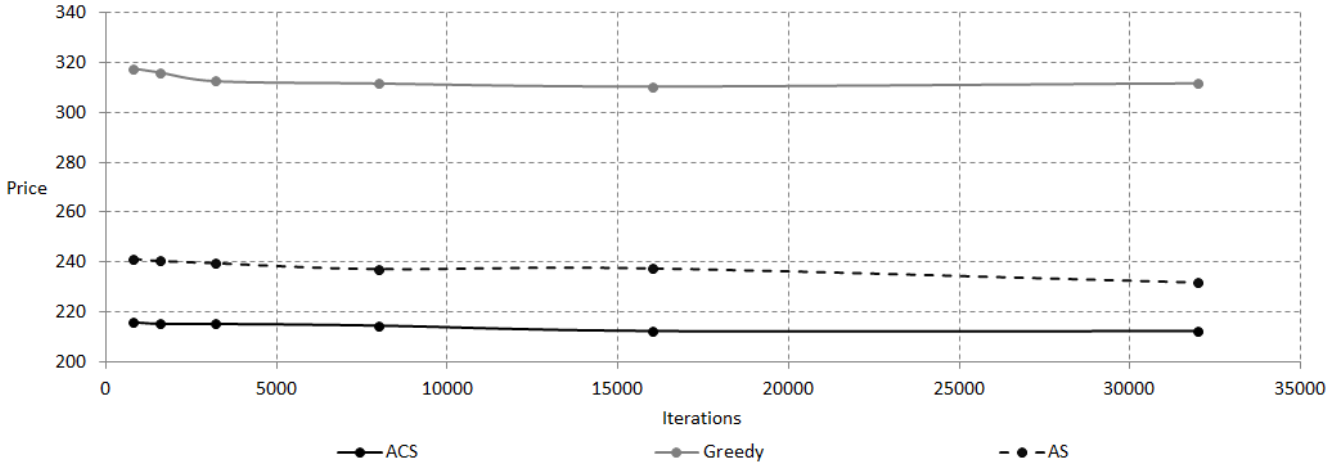


Fig. 4. Comparison between ACS, AS and a greedy algorithms.

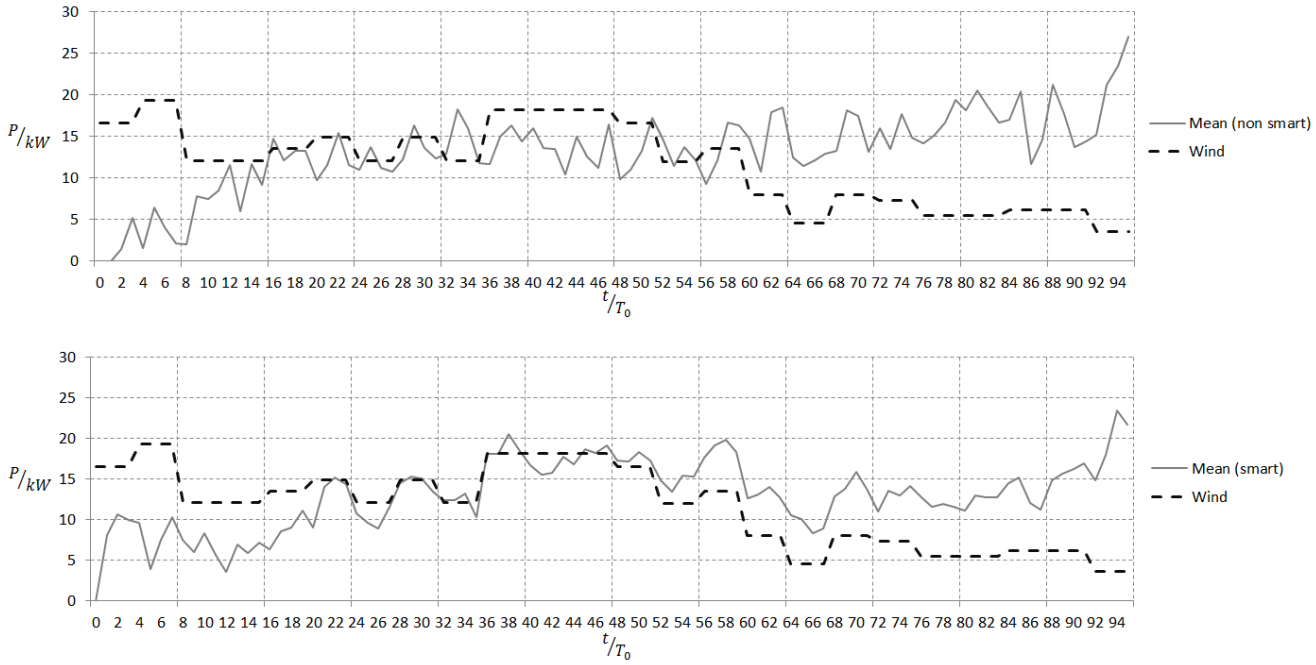


Fig. 6. Use-Case: Comparison of smart load control (below) and uncontrolled energy consumption (above).

schedule will not be able to utilize the available energy completely. But by utilizing the degrees of freedom of the four fully-automatic Device-Agents a significant energy saving can be achieved.

Figure 6 shows the aggregated results for 10 simulation runs. The upper chart shows the results for the uncontrolled energy consumption while the lower one displays the results of the ACS-based smart load optimization. For the latter parts of the day the energy consumption excels the generation as expected, but the optimized load control demands a significant lesser amount of external energy and utilizes especially earlier time slices better. The uncontrolled case utilizes 78,5% of the generated energy, while the controlled case uses 85,6% (+7,1%). This implies that for the uncontrolled case 100,0 kWh of external energy have to been bought in addition, while the optimized scenario requires only 78,5 kWh (-21,5 kWh). The addition of more fully-automated consumers, e.g. refrigerators or freezers should lead to further, significant improvements.

VI. Conclusion and Future Work

In this paper a Demand-Side-Management oriented Multi-Agent-System for distributed optimization was presented. The approach focusses on the automated use of flexibility potentials on the consumer-side. Through the use of a lightweight

and distributed optimization concept the demand-side was integrated into the electricity market. The defined generic Device-Agent supports the planning of both program-driven and fully-automated devices without requiring domain specific knowledge about the optimizing agents. The exchange market model allows the application of many different use-cases, including e.g. *flexible tariffs*.

The implementation of *Ant Colony System* (ACS) as a meta-heuristics for the optimization was described as well as the developed distribution concept of the algorithm. The presented use-case study includes 100 active domestic households with 158 Device-Agents and a wind turbine. The Device-Agents of the households optimized their demand towards the predicted power generation of the wind-turbine. It could be shown that the optimized planning lowered the demand for external energy about 21,5 kWh, even as the generation was extremely atypical for the normal load curve of the property.

Future work will cover several topics of the developed system. First, grid restrictions (avalanche-effect protection, see [20]) will be considered, which includes also the further development of the exchange market and its capabilities. A second aspect is to integrate a wider range of distributed optimization algorithms. Evolutionary Algorithms like Genetic Algorithms or other state of the art techniques of modern Operations Research like combined Branch-and-Cut and meta-heuristics will be examined as first tests have shown that ACS finds very good solutions but leaves room for performance improvement. Also specification for regional market layers as well as generic energy-service descriptions of the typical use-cases could be defined. Furthermore, field tests and the integration of the proposed system as extension into systems like OGEMA or OpenADR can be targeted.

REFERENCES

[1] C. Mitchell, R. Deumling, and G. Court, "Stabilizing california's demand," *Fortnightly Magazine*, vol. 03, p. 10, 2009.

[2] U.S. Energy Information Administration, "Annual energy outlook 2013 with projections to 2040," 2013.

[3] Statistisches Bundesamt, "Wirtschaftsbereich energie - erzeugung," Statistisches Bundesamt, Tech. Rep., 2013.

[4] NIST, "Roadmap for smart grid interoperability standards," *NIST special publication*, vol. 1108, 2010.

[5] J. Bruinenberg, L. Colton, E. Darmois, J. Dorn, J. Doyle, O. Elloumi, H. Englert, R. Forbes, J. Heiles, P. Hermans *et al.*, "Smart grid coordination group technical report reference architecture for the smart grid version 1.0 (draft) 2012-03-02," *CEN, CENELEC, ETSI, Tech. Rep*, 2012.

[6] J. Kok, C. Warmer, and I. Kamphuis, "Powermatcher: multiagent control in the electricity infrastructure," in *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*. ACM, 2005. doi: http://dx.doi.org/10.1145/1082473.1082807 pp. 75–82.

[7] J. Bergmann, C. Glomb, J. Gotz, J. Heuer, R. Kuntschke, and M. Winter, "Scalability of smart grid protocols: Protocols and their simulative evaluation for massively distributed ders," in *First IEEE International Conference on Smart Grid Communications (SmartGridComm)*. IEEE, 2010. doi: http://dx.doi.org/10.1109/SMARTGRID.2010.5622032 pp. 131 – 136.

[8] J. Kok, M. Scheepers, and I. Kamphuis, "Intelligence in electricity networks for embedding renewables and distributed generation," in *Intelligent infrastructures*. Springer, 2010, pp. 179–209.

[9] P. Palensky and D. Dietrich, "Demand side management: Demand response, intelligent energy systems, and smart loads," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 3, pp. 381–388, 2011. doi: http://dx.doi.org/10.1109/TII.2011.2158841

[10] D. Nestle, J. Ringelstein, and H. Waldschmidt, "Open energy gateway architecture for customers in the distribution grid," *Information Technology, Oldenbourg Verlag, Munich*, pp. 83–88, 2010.

[11] M. A. Piette, G. Ghatikar, S. Kiliccote, P. Palensky, C. McParland, E. Koch, and D. Hennage, "Open automated demand response communications specification (version 1.0)."

[12] T. Dethlefs and W. Renz, "A distributed registry for service-based energy management systems," in *39th Annual Conference of the IEEE Industrial Electronics Society, IECON*. IEEE, 2013. doi: http://dx.doi.org/10.1109/IECON.2013.6699896 pp. 4710–4714.

[13] C. W. Gellings, "The concept of demand-side management for electric utilities," *Proceedings of the IEEE*, vol. 73, no. 10, pp. 1468–1470, 1985. doi: http://dx.doi.org/0.1109/PROC.1985.13318

[14] Federal Energy Regulatory Commission, *A National Assessment of Demand Response Potential*, 2009.

[15] D. Huber, Z. Taylor, and S. Knudsen, "Environmental impacts of smart grid," National Energy Technology Laboratory, US-Department of Energy, Tech. Rep., 2011.

[16] T. Linnenberg, I. Wior, S. Schreiber, and A. Fay, "A market-based multi-agent-system for decentralized power and grid control," in *IEEE EFTA 2011*, 2011. doi: http://dx.doi.org/0.1109/ETFA.2011.6059126

[17] O. Lünsdorf and M. Sonnenschein, "Lastadaption von haushaltsgeräten durch verbundsteuerung," in *Tagungsband zum 3. Statusseminar des FEN*. Forschungsverbund Energie Niedersachsen, 09 2009.

[18] S. Ramchurn, P. Vytelingum, and A. Rogers, "Agent-based control for decentralised demand side management in the smart grid," in *The Tenth International Conference on Autonomous Agents and Multiagent Systems AAMAS 2011 (2011)*, 2011, pp. 5–12.

[19] S. Beer, M. Sonnenschein, and H.-J. Appelrath, "Towards a self–organization mechanism for agent associations in electricity spot markets," *Informatik*, 2011.

[20] A. Kamper, *Dezentrales Lastmanagement zum Ausgleich kurzfristiger Abweichungen im Stromnetz*. KIT Scientific Publishing, 2010.

[21] T. Logenthiran, D. Srinivasan, and T. Z. Shun, "Demand side management in smart grid using heuristic optimization," *IEEE Transactions on Smart Grid*, vol. 3, no. 3, pp. 1244–1252, 2012. doi: http://dx.doi.org/10.1109/TSG.2012.2195686

[22] W.-D. Zheng and J.-D. Cai, "A multi-agent system for distributed energy resources control in microgrid," in *Critical Infrastructure (CRIS), 2010 5th International Conference on*, Sept 2010. doi: http://dx.doi.org/10.1109/CRIS.2010.5617485 pp. 1–5.

[23] T. Logenthiran, D. Srinivasan, and A. M. Khambadkone, "Multi-agent system for energy resource scheduling of integrated microgrids in a distributed system," *Electric Power Systems Research*, vol. 81, no. 1, pp. 138 – 148, 2011. doi: http://dx.doi.org/10.1016/j.epsr.2010.07.019

[24] M. S. Narkhede, S. Chatterji, and S. Ghosh, "Article: Multi-agent systems (mas) controlled smart grid - a review," *IJCA Proceedings on International Conference on Recent Trends in Engineering and Technology 2013*, vol. ICRTET, no. 4, pp. 12–17, May 2013, published by Foundation of Computer Science, New York, USA.

[25] M. Dorigo and L. M. Gambardella, "Ant colony system: A cooperative learning approach to the traveling salesman problem," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 53–66, 1997.

[26] B. Bullnheimer, G. Kotsis, and C. Strauß, "Parallelization strategies for the ant system," 1997.

[27] M. Dorigo, V. Maniezzo, and A. Colorni, "Ant system: optimization by a colony of cooperation agents," *IEEE Transactions on Systems Man and Cybernetics*, vol. 26, pp. 29 – 41, 1996. doi: http://dx.doi.org/10.1109/3477.484436

[28] R. Stamminger, G. Broil, C. Pakula, H. Jungbecker, M. Braun, I. Rüdenauer, and C. Wendker, "Synergy potential of smart appliances," *Report of the Smart-A project*, 2008.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1497–1504
DOI: 10.15439/2014F318
ACSIS, Vol. 2

Overview of Research Challenges towards Smart Grid Quality by Design

David Gešvindr
Masaryk University
Faculty of Informatics
Brno, Czech Republic
Email: gesvindr@mail.muni.cz

Barbora Buhnova
Masaryk University
Faculty of Informatics
Brno, Czech Republic
Email: buhnova@mail.muni.cz

Jan Rosecky
Masaryk University
Faculty of Informatics
Brno, Czech Republic
Email: j.rosecky@mail.muni.cz

***Abstract*—Power distribution systems worldwide are entering one of the biggest transformations since their inception. The Smart Grid has become a strategic term, which is associated with promising technology enhancement on one hand, but also high investment risks on the other hand. To maximize the former and minimize the latter, power distribution companies are searching for methods and tools to simulate and compare different Smart Grid design alternatives and deployment scenarios.**

In this paper, we draw upon our involvement in the modelling and simulation of the Smart Grid in the Czech Republic, and elaborate possible directions of Smart Grid modelling and analysis research to reach the "quality by design" before the actual Smart Grid realization. The discussion then details a specific design challenge together with its context within the design of the Czech Smart Grid.

I. Introduction

THE CURRENT generation of power distribution networks faces new requirements that stimulate discussion on its future transformation. Some of the strongest triggers that the current power grid fails to keep pace with are the transition to a distributed power generation [1] and demand response optimization [2], whose both rely on mature information exchange and control. This stimulated the integration of the power distribution and information technology domains into the concept of a Smart Grid. The Smart Grid is an electricity network that can cost-efficiently integrate the behavior and actions of all users connected to it—generators, consumers and those that do both—in order to ensure economically efficient, sustainable power system with low losses and high levels of quality and security of supply [3]. The Smart Grid relies on an Advanced Metering Infrastructure (AMI) [4], [2] with a bi-directional communication channel which allows the control of electricity demand at customer level by directly or indirectly (e.g. by on-line adjusting electricity prices) controlling household appliances [2].

The deployment of Smart Grids is becoming a strategic act for many countries, forced by their legislation [5]. European Commission calls in its mandate M411 [6], released in 2009, for the standardization in the area of Smart Metering. The standards should enable interoperability of utility meters (water, gas, electricity, heat), and allow mass production of Smart Meter devices at the competitive European market. The mandate is aligned with the OPEN meter project [7], which ended in June 2011 and delivered open and public standards for smart metering, and consequent OpenNode project [8] with similar endevour regarding low-voltage monitoring.

The transition to the Smart Grid represents heavy investments (in orders of hundreds of millions of Euros) into technology that was never before deployed in such a large scale. To minimize investment risks and maximize benefits, power distribution companies are seeking new ways of Smart Grid modelling and simulation to evaluate and compare different Smart Grid architectures and deployment scenarios. This initiatives aim in each national context at a cost- and quality-effective validated solution, before mass investments into Smart Grid deployment.

Goal of the paper: In this paper, we draw upon our involvement in the modelling and simulation of the envisioned Smart Grid in the Czech Republic, and elaborate a discussion on the possible directions of Smart Grid modelling and analysis to reach the "quality by design" before the actual Smart Grid deployment.

Paper structure: The paper is organized as follows: Section II overviews the history, context and key terminology of the field; Section III discusses the specific research questions related to the modelling, simulation and analysis towards Smart Grid design; Section IV elaborates a specific research question concerned with the communication strategy discussed in the specific context of the Czech Smart Grid design; and Section V concludes the paper.

Related work: Although numerous overview papers have been published in the Smart Grid domain, the discussion of research questions and challenges towards mature Smart Grid design is not receiving systematic attention. The existing surveys instead focus on the goals, impact, benefits, risks and undelaying technology principles (architecture, communication technologies) of the Smart Grid, usually influenced by national research and deployment [4], [9], [2], [10]. The source of the information are hence the research papers elaborating individual research questions in isolation, such as for instance the papers on multi-agent approaches in a distributed Smart Grid [11], [12], [13], security-driven Smart Grid design [14], [15], [16], survivability analysis and Smart Grid self-repair [11], [17], or demand response optimization [18].

978-83-60810-58-3/$25.00

II. Background

One of the first large network utilities was built by Thomas Alva Edison in 1882 in New York. Edison promoted direct current for electric power distribution, despite the disadvantages of this approach, namely in short distance between a power plant and its customers (2.4 km was the limit of effective distance), which meant that small power plants needed to be built in customer areas. New research in Europe and America has brought electric motors and transformers working with alternating current and with the help of Edisons major opponent George Westinghouse, the previously used direct current was replaced with the alternating current, which persisted till today. Alternating current could be efficiently transmitted over long distance and hence allowed power plants to be built outside inhabited areas, closer to input resources.

The process of power transmission nowadays involves step-up transformers that transform electricity into very high voltage to decrease transmission losses. Electricity is then transmitted over hundreds of kilometers of transmission network. Before distribution to end customers, the voltage is decreased by a series of step-down transformers, which act as gates between high voltage *power transmission* and low voltage *power distribution networks*.

In the Czech Republic the *power transmission network* consists of power plants, first-level substations and high-voltage power lines. The first-level substations and power plants are connected using 400 kV and 220 kV lines. There is high flow of energy and therefore very high voltage is used to lower the electric current, which leads to lower losses during transmission. The first-level substations produce also 110 kV intended for power distributors.

The Czech *power distribution network* consists of 110 kV lines, second-level substations transforming 110 kV to 22 kV or 35 kV for connected third-level substations. Multiple third-level substations are situated in towns and produce well-known 230V/400V used by end customers.

The current generation of power transmission and distribution network has been designed for centralized electricity generation in a small number of large power plants. Nowadays, this concept is being replaced with a more distributed electricity generation thanks to an increasing amount of predominantly renewable energy sources. European Union supports renewable energy sources in its Energy and Climate Change Policy [20] with the goal to achieve 20% renewables share in the European energy mix until 2020.

The new approach, however, brings up serious issues: (1) Since the grid load changes along the day and since there is no efficient way to centrally store the surplus of generated electricity for later use or to rapidly increase the demand of electricity, power plants have to adjust their electricity production to meet the expected demand. Immediate production change is very complicated for most types of power plants, so both the consumption and the production of the renewables have to be predicted in advance, which brings additional risks to the whole process of electricity production and distribution. Moreover, thoughtless connection of many uncontrolled renewables (e.g. photovoltaic panels) may (2) damage the current grid or (3) cause significant technical losses, because the amount of generated electricity may in specific areas exceed its demand and overload the related part of the grid.

The first step towards effective management of power distribution is Smart Metering, which measures and collects power consumptions of individual customers. In particular, the deployment of Smart Metering can provide the following inputs for smart power distribution management:

a) Collection of power consumption data: Without deployment of Smart Meter units, the power consumption measurements are automatically collected only during electricity transmission in large substations. In the Czech Republic, moreover, only very large customers have similar devices already deployed. Collection of more data from individual consumers and all substations is a necessary condition of reliable predictions, which can help to locally balance the grid load.

b) Early fault detection: Current generation of power grids can only detect failures in substations that are connected via computer network to a control center that monitors the grid. It is also possible to detect broken high-voltage wires. But when it comes to individual customers, it does not provide means to detect power outages at individual customer level, because such an outage cannot be detected directly from the substation. In the Czech Republic, this type of outage is usually reported by the customer via a phone call at distributors service line. Nowadays, to detect this type of outage automatically, each monitored customer needs to be equipped with a device that periodically sends hearth beat signals to the substation. Smart meter measurements sent to a substation can substitute the hearth beat signaling.

The current state of the practice in power transmission and distribution implies numerous challenges for the Smart Grid design and deployment, many of which could be adressed with multi-agent approaches. The next sesion discusses these challenges and related research questions in detail.

III. Smart Grid modelling, simulation and analysis

To minimize the risks associated with the Smart Grid deployment and maximize its quality across multiple quality attributes, possible modelling and simulation techniques to evaluate and compare different Smart Grid design alternatives before the actual deployment are becoming of high interest to power distribution companies. This section summarizes the specific research questions towards Smart Grid design that we identified with our industrial partners, and aims at opening the discussion on the newly establishing research field of Smart Grid design support, to which the computer science, and multi-agent research in particular, can strongly contribute. The areas include:

Fig. 1. Map of the Czech transmission network, where red labels are the major power plants, red lines are 400 kV lines, purple dots are large substations and blue lines are 220 kV lines [19].

A. Prediction of the expected consumption
B. Load control
C. Energy backup and storage
D. Hardware infrastructure design
E. Smart grid topology design
F. Decision on the types of information to be exchanged and transferred
G. Evaluation of Smart Grid communication strategies
H. Quality assessment of different Smart Grid design options

A. *Prediction of the expected consumption*

Smart metering can significantly contribute to the creation of a new global power consumption model, recording power consumption even at an individual customer level. Such a model can be used for the prediction of expected electricity consumption. As mentioned in [2] a valid consumption-prediction model is much more complex than it used to be before, because of customer life style change. Also, a need to predict locally-dependent consumption and ability to react on lower electricity prices arises together with the requirement on local load-balancing.

We identified the following research questions associated with this topic:

- How to predict power consumption in both normal and crisis situations?
- What is the appropriate level of detail on which the consumption should be studied? How to cluster the Smart Grid architecture into consumption-relevant levels for analysis?
- What trends and technologies may significantly affect the validity of current prediction models? How to predict the consumption associated with the increasing use of electric automobiles?

B. *Load control*

The goal of the load control is to lower the gap between the base load and the daily peak load in each locality. This can be achieved by shifting power usage to off-peak hours, in particular to balance production of local renewables. There are multiple advantages of this approach: (1) It decreases the maximum power generation and transition capacity required to avoid blackouts; (2) It avoids extensive stopping and starting of power generating units by shaping the power usage to remain relatively constant over time [2]; (3) It helps avoiding electricity overflow from low voltage to high voltage and related problems mentioned above. This can be achieved either manually by motivating the customers to shift their power usage via variable pricing models or automatically by remote power distributor signaling that controls customer equipment.

In the Czech Republic, a combination of both mechanisms is used for households with water boilers which are controlled remotely via HDO signals[1]. The signals are broadcasted together with a code of the controlled area, so that only HDO switches in the targeted area are affected. Nevertheless, since these groups are fixed, one cannot smoothly control the load-balance, hence this solution works only partially and cannot be considered sufficient for the uprising problems with local renewable energy sources.

[1]HDO is a shortcut that stands for “Massive Remote Control” in Czech.

There are algorithms that can optimize the power demand at the customer level based on the electricity price. One of them is described and validated in [18]. However, to validate such a demand response model in a global context, we need to step-up in the hierarchy and model many interconnected customers in the grid and mainly the behavior of the distributor, who sends impulses (in form of real-time price or direct control of appliances) to increase or decrease the demand.

The research questions that could be answered with the help of a complex demand response model are:

- Are the current communication channels suitable for reliable control of customer appliances or real-time delivery of the price? How does the system behave in case of a broken communication line?
- How effective is the proposed demand response model in terms of lowering the gap between the base load and the peak load? What is its response time?
- What is the overall reliability and safety of the proposed solution? What happens if the load control mechanism misbehaves? E.g. intelligent agents start to spread invalid information to other agents in the grid which leads towards a snowball effect? What protection needs to be implemented to ensure grid survivability in such a case?

C. Energy backup and storage

One of the consequences of high difference between the base and peak load in current power grids can be the need for large-scale (to balance the whole grid) and small-scale (to balance individual areas) energy storage. As for the large-scale energy storage, according to Electric Power Research Institute the pumped-storage hydroelectricity provides more than 99% of bulk storage capacity worldwide [21] with the efficiency that varies between 70% and 80%.

Another approach to energy storage is the agent based micro-storage introduced by Vytelingum et al. [22]. This approach models the use of small capacity batteries (4 kWh) deployed in households and accompanied with an intelligent agent, which may be an integral part of a Smart Meter. The agent buys electric energy when it is cheap and stores this energy in batteries for its use during peak time when energy price increases. According to the study, this system can achieve savings of up to 13% for an average customer using a 4 kWh battery. Moreover, in case of power outage, the system can work as an energy backup. Other approaches involve electromobile batteries or local hydrogen production.

The research questions associated with this domain are:

- How to predict required energy storage capacity in case of successful Smart Grid deployment?
- Shall households employ local small-capacity energy storage? Will their deployment affect the overall demand-response strategy? Should they be strictly controlled by the utility provider or should they autonomously react on incentives (lower energy prices)?
- How will the demand response problem be affected by growing share of electro vehicles which are charged mostly over night or during working hours? Should be the chargers controlled remotely? How to balance the need of the user to charge the vehicle and the need of the grid to govern the demand? How to predict electric vehicle usage and its recharge power consumption?
- How to solve the problem of energy surplus when the demand response mechanism fails?

D. Hardware infrastructure design

The goal at least in the Czech Republic is to minimize hardware infrastructure investments and construction costs, and hence reuse as many existing infrastructure elements as possible when transiting to the Smart Grid. The suitability of every reused component should be individually assessed together with its parameters such as component dependencies, reliability, replacement costs and other associated limitations comparing to modern alternatives. Such an evaluation process, which is in fact multi-objective optimization, may be very slow and complex when done manually. Therefore, it is beneficial to employ a Smart Grid model with all relevant components modelled together with their dependencies and parameters. The model can be then used for semi-automatic or automatic evaluation of different versions of hardware components and their associated risks and benefits. The automation of this process may lead to more complex analysis of the model in a larger scale, which may identify previously hidden solutions that can be easily validated with running simulation on top of the grid models.

This approach is suitable for the modelling and validation of Smart Meter features, communication elements, and data concentrators. Even control and data-processing software can be evaluated in this way.

The most relevant research questions are:

- What design strategies can be used to maximize the quality of the Smart Grid under the constraint of legacy infrastructure?
- What hardware infrastructure changes will imply the highest ratio of quality increase to implementation costs (i.e. quality/cost)?

E. Smart grid topology design

The overall Smart Grid quality and behavior depends not only on the hardware components, but also on their organization into the grid topology. The topology, which determines the logical architecture of the grid, can take advantage of different levels of centralization (e.g. customer data from a given region managed through a single substation), hierarchical structure (e.g. substations governed by parent substations) and redundancy (to minimize the critical chain effects and maximize grid survivability in case of local outage).

The relevant research questions towards the logical architecture of a Smart Grid are the following:

- What shall be the level of centralization within the Smart Grid?

- Shall the substations and other grid components be organized in a hierarchical manner? If yes, how many levels should be used?
- What level of homogeneity/heterogeneity (with respect to the smart logic within the grid nodes) is most beneficial?
- What are the critical chains of the Smart Grid with respect to the consequences of their outage?
- What is the level of grid robustness and survivability? What are the most cost-effective topology improvements that significantly increase robustness and survivability? What is the expected benefit of self-repair mechanisms comparing to their implementation costs?

F. Decision on the types of information to be exchanged and transferred

The concept of a Smart Grid relies on the transfer of various information between grid components. The Advanced Metering Infrastructure (AMI) needs to transfer measurement results collected from Smart Meters for processing purposes, substations need to report their state to a parent substation in case of hierarchical infrastructure or broadcast their state to other substations in case of multi-agent processing.

To minimalize the traffic among Smart Grid components, the information value of different types of data needs to be understood and used to drive the identification of needless data, whose transfer should be limited. On the other hand, the critical information types shall be prioritized in the communication.

The relevant research questions include:

- How to identify and model importance of collected measurements for later processing in the Smart Grid? How this process can be automated?
- What is the ideal collection frequency of measurements from Smart Meters to satisfy the legislation needs, ensure sufficient precision of prediction models and help load balancing on one hand and lower the communication line load on the other hand?
- How to efficiently identify dependent information for producing complex measurements across more elements in the grid?
- What is the acceptable age of particular type of information for a given stakeholder?

G. Evaluation of Smart Grid communication strategies

Modelling of different communication strategies in the Smart Grid is essential for the evaluation of grid behavior in different situations. Identification of a suitable communication strategy is often hindered with a set of constraints that must be fulfilled. For instance, one of the primary restrictions in the Czech Republic is the employment of existing communication infrastructure, because significant rebuild of the infrastructure would unacceptably increase the Smart Grid deployment costs. When following the physical structure of the grid described above, we need to transfer measurements from Smart Meter units placed at customers electrical connections to a datacenter for processing, and control commands back to the elements of the grid. The problem is that a Smart Meter cannot be usually connected directly to the computer network to send all required data securely over the Internet.

Employment of the Internet connection of the customer for direct connection of the Smart Meter to the Internet would be a solution, but in terms of its availability, reliability and the legal complexity, it is merely impossible to deploy it nationwide.

There are two available groups of technologies that can be used in Smart Grids to transfer data between the elements of the grid—data transfer over power lines and wireless data transfer. Possible use and issues associated with these technologies in the Czech Smart Grid are examined in Section IV.

Relevant research questions for this area are:

- Can network modeling and network simulation tools be used in large scale as Smart Grid modeling and validation requires? What optimizations need to be done prior their use?
- What level of precision in a network simulation is required for trustworthy validation of the communication strategy?
- What communication channels and communication strategies are suitable for control channel? What are the requirements that control channels need to meet?
- Is it better to transfer raw measurements or pre-processed data? Where does the processing take place?
- How can we increase the reliability of the line? How can we get cost-effective line redundancy?

H. Quality assessment of different Smart Grid design options

The quality evaluation, prediction and optimization during Smart Grid design are critical cross-cutting concerns to all the research issues discussed above. The Smart Grid quality is typically understood in terms of the grid reliability, security and efficiency, and considered along with its environmental and energy sustainability and cost.

The reliability engineering in the Smart Grid domain aims at incorporating autonomous control actions to enhance reliability by increasing resiliency against component failures and natural disasters, and by minimizing frequency and magnitude of power outages subject to regulatory policies, operating requirements, equipment limitations, and customer preferences [23]. This includes the techniques for automated failure detection, isolation and restoration (FDIR), and terms such as grid survivability and robustness.

The security requirements for a Smart Grid are mainly the authenticity, integrity, and privacy of metering data, along with the safety and availability of the grid [24]. The Smart Grid design shall aim at resiliency against malicious attacks through better physical security and cybersecurity, to protect the infrastructure, energy asset and data of the grid.

The efficiency of a Smart Grid refers to the infrastructure utilization on both the energy-transfer and data-transfer level. The new research questions emerge mainly in the context of data-transfer efficiency, which is specifically concerned with

the level of utilization of communication channels and latency of data delivery.

The research questions concerned with the quality assessment and optimization include:

- Which of two or more Smart Grid design options is better with respect to a specific quality attribute?
- How shall a specific Smart Grid design alternative be changed to optimize a specific quality attribute (under given constraints, such as the cost)? E.g. to optimize the time needed for failure detection, isolation and grid restoration.
- What is the quality trade-off of different quality-related strategies? E.g. what is the security improvement and efficiency degradation of a specific security strategy?
- What is the context (characteristics of the settings) in which a specific quality-improving strategy yields highest benefits?
- Can the Smart Grid reach a specific unsafe condition (i.e. a state violating safety specification)?
- What are the critical components of the Smart Grid with respect to a specific quality attribute? E.g. what are the most critical components with respect to grid security or survivability? What communication links have the highest impact on data-transfer latency?
- What is the level of satisfaction of individual stakeholders, including the energy distributors, customers and public organizations?

Following on from these research questions the next section describes in detail various communication strategies that are being currently evaluated by Czech power distribution companies for potential deployment in their Smart Grid projects.

IV. Situational study of the communication strategy for the Czech Smart Grid

Together with our industrial partner, the Mycroft Mind company, which is a leading expert in the field of Smart Grid simulation in the Czech Republic, we are researching new ways of Smart Grid modelling, simulation and analysis that can be integrated into their analytical tool called Grid Mind.

One of the biggest challenges in Smart Grid deployment in the Czech Republic is the communication strategy. The Advanced Metering Infrastructure (AMI) produces waste amount of data that needs to be transferred to a datacenter of the power distributor for processing. Every day the grid of the CEZ power distributor is expected to produce more than 340 Million measurements. The key concern is that the current communication channels will not be able to transfer the measurements in time. Another challenge is to process the measured data at real time, which is necessary for effective demand response optimization. Last, there is a concern that the control channels will not be able to deliver the control commands with the necessary level of reliability.

Our colleagues at the Faculty of Informatics, Masaryk University, Brno, were involved in the research examining what transfer characteristics can be expected with different versions of GPRS. We now look at this problem in a broader way. Together with Mycroft Mind we are building a modelling and simulation environment that is planned to be used for modelling, simulation and validation of various communication strategies in the Smart Grid, whose key characteristics are elaborated in the rest of this section.

There are two available groups of technologies that can be used in a Smart Grid to transfer data between the elements of the grid—data transfer over *power lines* and *wireless* data transfer.

The Power Line Carrier (PLC) technology stands for a family of communication protocols and technologies that transfer data over *power lines*. Different PLC technologies vary in frequency ranges used as transport medium for data: NPL (Narrow Power Line) uses frequencies between 3 KHz and 148.5 KHz, BPL (Broadband Power Line) uses higher frequency range between 2 MHz and 50 MHz. The quality of the data transfer is significantly dependent on the properties of the power line, such as the path attenuation and noise in case of medium voltage lines connecting primary to secondary substations. The number of parameters that need to be evaluated increases when low voltage lines that connect secondary substations and Smart Meter devices are used. More detailed characteristics of PLC technologies can be found in [25].

In the low voltage network there are two types of devices: central low voltage unit (CLU) and remote low voltage unit (RLU). CLU is located at low voltage substations usually with a few hundreds of connected customers. RLU is built into a Smart Meter and may act also as a repeater in order to extend the action radius of the PLC signal. The PLC communication medium is composed of sub networks defined by CLU. To avoid collisions, access to PLC medium has to be controlled, therefore CLU controls all the traffic and RLU are allowed to communicate only in case they respond to CLU's request.

PLC was chosen as a suitable technology for centralized collection of measurements from Smart Meters. The benefit is that it uses existing power lines which decreases deployment costs and provides sufficient communication channel for collection of power consumption measurements. The deployment of this technology is currently being validated using almost 40,000 Smart Meter units deployed in four Czech towns [26]. The problem is how to transfer the collected data from substations.

The latter group of technologies usable for data transfers in the Smart Grid is *wireless* transfer. There are many available wireless communication technologies that can be effectively used for measurements transfer to a data center for processing. We discuss some of the technologies usable over large distance and explain their advantages and disadvantages below.

a) 2G GPRS/EDGE: In the Czech Republic, the coverage of public cellular networks is approaching 99% of area. 2G cellular networks provide data services in the speed of kbps or tens kbps. These data services are optimally tailored for automated meter readings. Mobile operators offer data-only SIM cards for industrial use with lower communication priority. No investments into network infrastructure are needed

for cellular network-based AMI solutions. Transferring data over 2G cellular network from each individual Smart Meter unit would not be cost effective in terms of running costs paid to mobile operator, therefore we propose to use this technology to transfer collected data from substation. It is necessary to model such a situation and validate if the transfer speed of 2G networks is sufficient.

b) 3G UMTS: 3G cellular networks are extending the functionality of traditional 2G networks, allowing data transfers at significantly higher speeds. With HSPA extension peak data rates up to 14 Mbit/s in the downlink and 5.75 Mbit/s in the uplink are supported. Data rates of 3G networks are fully sufficient for AMI purposes. In the Czech Republic, most of the 3G coverage is in densely populated areas. According to the data provided by the Czech Telecommunication Office, the mobile operator with the best 3G signal coverage covered at the end of year 2012 only 48% of the Czech Republic [27]. Substations in uncovered areas will have to rely on slower 2G networks.

c) 4G LTE: LTE cellular network rapidly increases transfer speeds up to 100 Mbit/s. Mobile operators started to deploy this communication technology in the Czech Republic at the end of 2013. The Vodafone mobile operator selected this technology as a complement to their 3G network which will not be extended anymore. They would like to primarily upgrade 2G EDGE connection in places uncovered by HSPA to cover 93% of the Czech population with 3G or 4G connectivity until December 2014. Such a rapid deployment of LTE as a replacement of 2G network should be considered in evaluation of Smart Grid communication model in case of 2G networks insufficient speed.

d) WLAN/WiFi: Broadband wireless technologies based on the IEEE 802.11 standard have found broad acceptance worldwide for wireless area networking. The use of WLANs for AMR has been frequently reported. In the US, several cities (e.g. Santa Clara, CA) already use WLANs to automatically collect readings from electricity, gas, and water meters [28]. Using 802.11s technology supporting mesh may provide a technically attractive solution for AMR [28]. Use of public WLAN in municipalities, if available, may solve problems with connection availability at substations, but considerable number of substations is out of reach of WLANs. Using WLAN as primary connection technology for measurements collection is less reliable than PLC.

e) WiMAX: Worldwide Inter-operability for Microwave Access is a wireless broadband access technology which has been designed to provide transfer speeds up to 72 Mbit/s in both directions. Communication range varies between few kilometers in case of non-line of sight conditions and 50 km in case of line of sight conditions. Deployment of a dedicated WiMAX network solely for the purpose of an AMI likely cannot be justified. In those areas where public WiMAX services are available (e.g. municipal networks), WiMAX may principally be an option [28].

f) ZigBee: ZigBee standards-based protocols provide infrastructure for wireless sensor network applications. ZigBee is capable of creating a mesh network, which provides additional capabilities—increased redundancy, self-configuration and self-healing. Situation of building nationwide dedicated network using ZigBee is the same like in case of WiMAX. Possible use of ZigBee wireless network in terms of Smart Metering is connection of individual sensors where the power line does not meet quality requirements of PLC based protocol. Another use may be peer-to-peer connection between substations to increase redundancy or capacity of Internet connection (e.g. EDGE).

g) Satellite systems: Satellite system are suitable for sending measurements in areas where no suitable terrestrial infrastructure is available. Today, there is a number of satellite operators offering data transfer services. For instance Iridium service, which uses low earth orbit satellites or SpaceChecker that uses geostationary satellites. Deployment in areas covered by other communication infrastructures needs to be evaluated, because costs are expected to be significantly higher than use of other terrestrial communication technologies.

The infrastructure that is currently examined in the Czech Republic combines PLC technology to transfer measurements from Smart Meters to a substation. Each substation is collecting data usually from a few hundreds Smart Meters. The data concentrator unit, which is deployed at substations, can process data locally which opens new possibilities in building more decentralized solutions. Each data concentrator is equipped with a cellular network modem, in the Czech Republic usually EDGE technology, and sends data to a data center. This approach brings significant improvement in terms of connection costs reduction but increases the demands on speed and reliability of Internet connection available at substations, because more collected data needs to be transferred from a single point.

We were provided with the operational measurements of data transfer speed and latency of currently used GPRS communication channel in the test deployment of Smart Grid in the Czech Republic. The used low-priority industrial SIM cards can utilize only 10% of the communication capacity of a mobile base transceiver station (BTS), therefore under heavy load they operate only at the speed of few Kbits per second. Provided measurements show problems with connection handling across different levels of the network stack. If there is a network connection that is not used for a longer period of time, this connection is disconnected by the BTS at the physical level, but the TCP connection remains open. Any later attempt to transfer data via this broken TCP connection results in a communication failure and a need to reopen the connection which takes a few seconds to establish. The discussed problems illustrate the need to build a valid simulation model of the network communication in the Smart Grid. To be able to handle the high complexity of this model, we propose together with our industrial partner Mycroft Mind decomposition of this model into three interconnected simulation models. The

first model describes the simplified behavior of the physical layer, the second model describes the simplified version of the transport protocol and the third model describes used application protocols. The Grid Mind simulation environment utilizes the discrete event simulation and proposed models to evaluate different communication channels, protocols and communication strategies.

Constraint in a form of available Internet connection at the locality of the substation is one of the key problems that needs to be evaluated using correct modelling and simulation approaches.

V. Conclusions

To minimize the risks associated with the Smart Grid deployment and maximize its quality across multiple quality attributes, possible modelling and simulation techniques to evaluate and compare different Smart Grid design alternatives before the actual deployment are becoming of high interest. In this paper, we elaborated a discussion on the possible directions of Smart Grid modelling and analysis to reach the "quality by design", and presented specific research questions towards Smart Grid design that we identified with our industrial partners. One of the research questions, concerned with the communication strategy, was then discussed in detail and linked to the specific situation in the Czech Republic. The communication strategy evaluation is at the same time our ongoing research aim, which will be in the future extended in the direction of the research questions discussed in this paper.

Acknowledgement

We would like to thank our industrial partners, especially the Mycroft Mind company and its head Filip Prochazka, for the fruitful discussions and valuable inputs that made this paper possible.

References

[1] J. Bremer and M. Sonnenschein, "A distributed greedy algorithm for constraint-based scheduling of energy resources," in *Proc. of FedCSIS'12*. IEEE, 2012. ISBN 978-1-4673-0708-6 pp. 109–115.

[2] Y. Simmhan, S. Aman, B. Cao, M. Giakkoupis, A. Kumbhare, Q. Zhou, D. Paul, C. Fern, A. Sharma, and V. Prasanna, "An informatics approach to demand response optimization in smart grids," Computer Science Department, University of Southern California, Tech. Rep., 2011.

[3] EU, "Standardization mandate to european standardisation organisations (ESOs) to support european smart grid deployment," March 2011, last retrieved 2014-04-21. [Online]. Available: http://ec.europa.eu/energy/gas_electricity/smartgrids/doc/2011_03_01_mandate_m490_en.pdf

[4] X. Fang, S. Misra, G. Xue, and D. Yang, "Smart grid – the new and improved power grid: A survey," *Communications Surveys Tutorials, IEEE*, vol. 14, no. 4, pp. 944–980, 2012. doi: 10.1109/SURV.2011.101911.00087

[5] EU, "Single market for gas & electricity: Smart grids," July 2013, last retrieved 2014-04-21. [Online]. Available: http://ec.europa.eu/energy/gas_electricity/smartgrids/smartgrids_en.htm

[6] EU, "Standardisation mandate to CEN, CENELEC and ETSI in the field of measuring instruments for the development of an open architecture for utility meters involving communication protocols enabling interoperability," March 2009, last retrieved 2014-04-21. [Online]. Available: http://ec.europa.eu/energy/gas_electricity/smartgrids/doc/2009_03_12_mandate_m441_en.pdf

[7] OPEN meter, July 2012, last retrieved 2014-04-21. [Online]. Available: http://www.openmeter.com/

[8] M. Alberto, R. Soriano, J. Gotz, R. Mosshammer, N. Espejo, F. Lemenager, and R. Bachiller, "OpenNode: A smart secondary substation node and its integration in a distribution grid of the future," in *Proc. of FedCSIS'12*. IEEE, 2012. ISBN 978-1-4673-0708-6 pp. 1277–1284.

[9] R. Hassan and G. Radman, "Survey on smart grid," in *Proc. of IEEE SoutheastCon'10*, 2010. doi: 10.1109/SECON.2010.5453886 pp. 210–213.

[10] W. Wang, Y. Xu, and M. Khanna, "A survey on the communication architectures in smart grid," *Computer Networks*, vol. 55, no. 15, pp. 3604–3629, 2011. doi: 10.1016/j.comnet.2011.07.010

[11] S. Amin and B. Wollenberg, "Toward a smart grid: power delivery for the 21st century," *Power and Energy Magazine, IEEE*, vol. 3, no. 5, pp. 34–41, 2005. doi: 10.1109/MPAE.2005.1507024

[12] M. Pipattanasomporn, H. Feroze, and S. Rahman, "Multi-agent systems in a distributed smart grid: Design and implementation," in *Proc. of PSCE'09*, 2009. doi: 10.1109/PSCE.2009.4840087 pp. 1–8.

[13] S. D. Ramchurn, P. Vytelingum, A. Rogers, and N. Jennings, "Agent-based control for decentralised demand side management in the smart grid," in *Proc. of AAMAS'11*. ACM, 2011, pp. 5–12.

[14] P. McDaniel and S. McLaughlin, "Security and privacy challenges in the smart grid," *Security Privacy, IEEE*, vol. 7, no. 3, pp. 75–77, 2009. doi: 10.1109/MSP.2009.76

[15] H. Khurana, M. Hadley, N. Lu, and D. Frincke, "Smart-grid security issues," *Security Privacy, IEEE*, vol. 8, no. 1, pp. 81–85, 2010. doi: 10.1109/MSP.2010.49

[16] A. Metke and R. Ekl, "Security technology for smart grid networks," *IEEE Transactions on Smart Grid*, vol. 1, no. 1, pp. 99–107, 2010. doi: 10.1109/TSG.2010.2046347

[17] D. S. Menasché, R. M. Meri Leão, E. de Souza e Silva, A. Avritzer, S. Suresh, K. Trivedi, R. A. Marie, L. Happe, and A. Koziolek, "Survivability analysis of power distribution in smart grids with active and reactive power modeling," *SIGMETRICS Perform. Eval. Rev.*, vol. 40, no. 3, pp. 53–57, Jan. 2012. doi: 10.1145/2425248.2425260

[18] A. Conejo, J. Morales, and L. Baringo, "Real-time demand response model," *Smart Grid, IEEE Transactions on*, vol. 1, no. 3, pp. 236–242, 2010. doi: 10.1109/TSG.2010.2078843

[19] CEPS, "Schéma rozvodné sítě v ČR (schema of Czech transmission network)," 2013, last retrieved 2014-04-21. [Online]. Available: http://www.ceps.cz/CZE/Cinnosti/Technicka-infrastruktura/PublishingImages/Mapa_siti_CZ.PNG

[20] EU, "Communication from the commision to the european council and the european parlament - an energy policy for europe," January 2007, last retrieved 2014-04-21. [Online]. Available: https://www.europia.eu/Content/Default.asp?PageID=412&DocID=13922

[21] EPRI, "Electricity energy storage technology options," December 2010, last retrieved 2014-04-21. [Online]. Available: http://www.epri.com/abstracts/Pages/ProductAbstract.aspx?ProductId=000000000001020676

[22] P. Vytelingum, T. D. Voice, S. D. Ramchurn, A. Rogers, and N. R. Jennings, "Agent-based micro-storage management for the smart grid," in *Proc. of AAMAS '10*, 2010, pp. 39–46.

[23] K. Moslehi and R. Kumar, "A reliability perspective of the smart grid," *IEEE Trans. on Smart Grid*, vol. 1, no. 1, pp. 57–64, June 2010. doi: 10.1109/TSG.2010.2046346

[24] D. von Oheimb, "It security architecture approaches for smart metering and smart grid," in *Proc. of SmartGridSec'12*, ser. LNCS, vol. 7823. Springer, 2013. doi: 10.1007/978-3-642-38030-3_1 pp. 1–25.

[25] OPEN meter, "Description of current state-of-the-art of technology and protocols - description of state-of-the-art plc-based access technology," May 2009, last retrieved 2014-04-21. [Online]. Available: http://www.openmeter.com/files/deliverables/OPEN-Meter\%20WP2\%20D2.1\%20part2\%20v2.3.pdf

[26] ČEZ, "Čez smart grids inteligentní měření," January 2014, last retrieved 2014-04-21. [Online]. Available: http://www.futuremotion.cz/smartgrids/cs/index.html

[27] CTU, "Výroční zpráva českého telekomunikačního úřadu za rok 2012 (annual report of the czech telecommunication office for 2012)," April 2013, last retrieved 2014-04-21. [Online]. Available: http://www.ctu.cz/cs/download/vyrocni_zpravy/vyrocni_zprava_ctu_2012.pdf

[28] OPEN meter, "Description of current state-of-the-art technologies and protocols - general overview of state-of-the-art technological alternatives," June 2009, last retrieved 2014-04-21. [Online]. Available: http://www.openmeter.com/files/deliverables/OPEN-Meter\%20WP2\%20D2.1\%20part1\%20v3.0.pdf

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1505–1514
DOI: 10.15439/2014F76
ACSIS, Vol. 2

Conjoint Dynamic Aggregation and Scheduling Methods for Dynamic Virtual Power Plants

Astrid Nießе*, Sebastian Beer*, Jörg Bremer†, Christian Hinrichs†, Ontje Lünsdorf* and Michael Sonnenschein†
*R&D Division Energy, OFFIS – Institute for Information Technology, Escherweg 2, D-26121 Oldenburg, Germany
Email: {astrid.niesse, sebastian.beer, ontje.luensdorf}@offis.de
†Environmental Informatics, Department of Computing Science, University of Oldenburg, D-26111 Oldenburg, Germany
Email: {joerg.bremer, christian.hinrichs, michael.sonnenschein}@uni-oldenburg.de

***Abstract*—The increasing pervasion of information and communication technology (ICT) in energy systems allows for the development of new control concepts on all voltage levels. In the distribution grid, this development is accompanied by a still increasing penetration with distributed energy resources like photovoltaic (PV) plants, wind turbines or small scale combined heat and power (CHP) plants. Combined with shiftable loads and electrical storage, these energy units set up a new flexibility potential in the distribution grid that can be tapped with ICT-based control following the long-term goal of substituting conventional power generation. In this contribution, we propose an architectural model and algorithms for the self-organization of these distributed energy units within dynamic virtual power plants (DVPP) along with first results from a feasibility study of the integrated process chain from market-driven DVPP formation to product delivery.**

***Index Terms*—Smart Grid, Virtual Power Plant, Agent-Based Control, Self-Organization.**

I. Introduction

DISTRIBUTED energy resources like photovoltaic (PV) plants, wind turbines or small scale combined heat and power (CHP) plants entered the energy market in many European countries, especially Germany, with the financial security of guaranteed electrical feed-in tariffs. With their share in the market still rising, a concept is needed to integrate them into the very same regarding both real power and ancillary services to reduce subsidy dependence and follow the goals as defined by the European Commission.

Virtual power plants are a well-known concept for the aggregation of distributed energy resources (DER) to deliver both energy products and ancillary services [1]. Besides the control of generation by distributed energy resources like e. g. photovoltaic plants, shiftable loads like heat pumps, water boilers or air conditioners can be controlled to adapt the load profile regarding different optimization targets. Electrical storage may additionally be a new player in this scene, delivering even more flexibility for the optimized use of distributed generation. To address these three aspects, generation, load and storage, we will refer to distributed energy units (DEU) for the rest of this paper.

Parts of this work have been funded by the Lower Saxony Ministry of Science and Culture through the 'Niedersächsisches Vorab' grant programme (grant ZN 2764) within the project cluster Smart Nord.

In this contribution, we present an architectural model and algorithms for conjoint distributed aggregation algorithms using flexibility modelling and distributed scheduling heuristics. We present the evaluation environment that will be used for the evaluation of these conjoint processes within dynamic virtual power plants (DVPP) for the use case of active power delivery on current energy markets like the European Power Exchange (EPEX SPOT), along with first results from a feasibility study implementing these processes. In developing the integrated process shown here, we followed the Smart Grid Algorithm Engineering (SGAE) approach described in [2].

To tap the full flexibility potential of all energy units in the distribution grid we set up the following domain-driven paradigms for DVPPs (cf. [3]):

- Distributed energy units have to trade their services on markets, for both active power products, and ancillary services (as far as possible; see e. g. [4] for the position of the German Federal Network Agency regarding this topic).
- To dynamically adapt to current power system operational states and handle the vast amount of DEU in the distribution grid, an approach based on self-organization principles is used. By this means, characteristics like robustness, scalability and adaptivity of the overall system should be gained.
- DVPPs should be set up on a per-product base, thus allowing for optimal aggregation of energy units regarding the products needed. The paradigm of a dynamic VPP with respect to the product obligation is completely different from current virtual power plant concepts. It has to be evaluated, if more flexibility can be extracted from the distribution grid with such a highly dynamic approach.
- The potential of DVPPs for power system control lies in their units' flexibility. Therefore a generic representation of these flexibilities is needed, building the foundation for all DVPP mechanisms concerned with DEU scheduling.
- For active power delivery on energy markets, the operation of DEUs is controlled using operation schedules for all different types of units. The resolution of the DEUs' operation schedules should reflect current schedule resolutions by indicating mean active power values

978-83-60810-58-3/$25.00

for each 15 min. time interval. This is different to the current handling of renewable energy sources – current systems work with prognoses and use schedules only for controllable generating electricity units.
- To deliver ancillary services with locality constraints (like voltage control), DVPPs have to be able to reflect the grid topology. Therefore grid topology should be an optional parameter in the aggregation process and within the operation of DVPPs.

Within this context, the objective of this contribution is to introduce a seamless process chain for day-ahead based active power provision by means of DVPPs. We present an integrated multi-agent system (MAS) realizing the aggregation algorithm, the scheduling heuristic as well as the flexibility modelling used for DVPP management. For this, we start with an overview on the state of the art regarding distributed control in energy systems in Section II and show why this control scheme is appropriate for DEU interaction on energy markets. In Section III we introduce the use case of active power delivery on day-ahead markets in detail. In Section IV the algorithms for the aggregation of agents to dynamic VPPs are explained, showing how grid topology is a guidance in this process without yielding hierachically restrained static aggregation schemes. The scheduling of DEU is depicted in Section V. In Section VI the generation of surrogate models to represent DEU flexibilities is explained. The integrated agent model with respect to the generation and usage of this surrogate model is shown in Section VII. The evaluation architecture and first results from a feasibility study are presented and discussed in Section VIII. We finish this contribution with a conclusion and an outlook on future work in Section IX.

II. Distributed Control in Energy Systems

The operational management of energy systems involves a number of complex tasks ranging from technical aspects like supervisory control and data acquisition (SCADA) to organizational measures performed by business management systems (BMS). These are coupled within an energy management system (EMS) based on information and communication technology (ICT). Traditionally, the EMS is implemented as a centralized control system. However, given the increasing share of DEUs in the distribution grid today, the evolution of the classical, rather static (from an architectural point of view) power system to a dynamic, continuously reconfiguring system of individual decision makers endangers the feasibility of such centralized control schemes. In the seminal work of Wu et al. [5], the need for decentralized control has been identified as follows: "Control centers today are in the transitional stage from the centralized architecture of yesterday to the distributed architecture of tomorrow. [...] To summarize, in a competitive environment, economic decisions are made by market participants individually and system-wide reliability is achieved through coordination among parties belonging to different companies, thus the paradigm has shifted from centralized to decentralized decision making." In line with this vision, the International Energy Agency (IEA) describes a possible transition to decentralized control in three steps [6]:

1) **Accommodation.** Distributed generation is accommodated into the current market with the right price signals. Centralized control of the networks remains in place.
2) **Decentralization.** The share of DG increases. Virtual utilities optimize the services of decentralized providers through the use of common communications systems. Monitoring and control by local utilities is still required.
3) **Dispersal.** Distributed power takes over the electricity market. Microgrids and power parks effectively meet their own supply with limited recourse to grid-based electricity. Distribution operates more like a coordinating agent between separate systems rather than controller of the system.

The concept of a *virtual utility* mentioned therein was introduced in the late nineties and describes a "[...] flexible collaboration of independent, market-driven entities that provide efficient energy service demanded by consumers [...]." [7] Virtual power plants (VPP) have been studied extensively as a derivation from this concept with a number of successful realizations [8]. Additionally, different operational targets have been defined and implemented for VPPs, like aggregating energy (commercial VPPs) or delivering system services (technical VPPs) [1]. These VPP concepts form a basis for the decentralization stage in the transition path above. However, such VPPs usually focus on the long-term aggregation of generators (and sometimes storages) only and are each still operated in a centralized manner. For an implementation of the dispersal stage in the transition path, a more flexible concept is required. In the last years, a significant body of research emerged on this topic. For instance, [9] surveys the use of agent-based control methods for power engineering applications. Exemplary applications can be found in [10], [11], [12]. Finally, a research agenda in this context was proposed recently in [13].

In contrast to the work referenced above, the concept of DVPPs explicitly takes the current market situation into account for the process of forming aggregations of DEUs: DVPPs form with respect to concrete products at an energy market, and will dissolve after delivering a product. Additionally, fully distributed control algorithms are being used, as will be shown in the following sections, building the foundation for the dispersal stage in the mentioned transition path. A preliminary description of the concept including a detailed differentiation from related approaches was given in [3].

III. Dynamic Virtual Power Plants

To introduce the concept of dynamic virtual power plants and show which tasks have to be performed by the software agents, we refer to the use case of active power products traded on the day-ahead power market, where product trading is based on an auction mechanism as described in [3] (see Fig. 1). From the market perspective, three different phases have to be distinguished. In the first phase, bids can be placed in the so-called order book for predefined product types. Once the

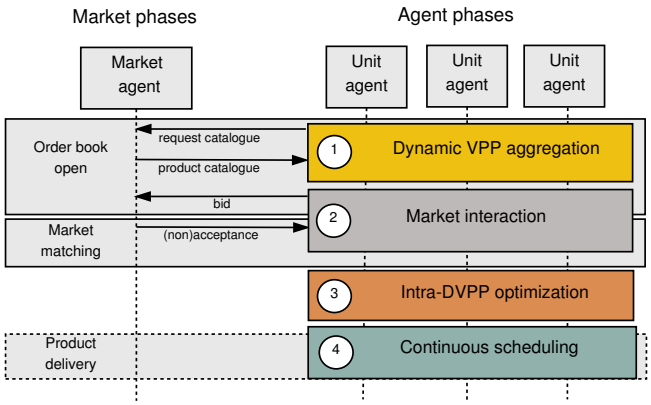


Fig. 1. Simplified use case of a dynamic VPP delivering energy products on a day-ahead market

order book is closed (e. g. at 12 a.m. for active power products traded day-ahead in Germany at EPEX SPOT) a matching mechanism clears supply and demand bids to set up the market price. In the last phase, these products have to be delivered, but no distinct market actions are entangled with this phase: the surveillance of product delivery and associated actions for balancing are subject to balancing group management.

To implement this process with regard to DVPPs, unit agents are set up to represent distinct DEUs within a multi-agent system. Four sequential phases within these unit agents are needed for active power delivery on day-ahead markets, as can be seen on the right hand side of Fig. 1:

1) **Dynamic VPP aggregation:** First, energy units have to be appropriately aggregated to DVPPs with the goal to deliver active power products. Grid topology has to be an optional parameter in this phase.
2) **Market interaction:** In the second phase, DVPPs place their active power products on the market by means of a *representative* agent for each respective DVPP and are informed about acceptance after market matching. Thus, after market matching the units' obligations regarding their power contributions are known.
3) **Intra-DVPP optimization:** Within a third phase, an intra-DVPP optimization is performed, taking into account these obligations and updated prognoses regarding the units' operational states.
4) **Continuous scheduling:** The last phase is concerned with continuous energy scheduling to ensure product fulfillment. In case of an incident endangering product delivery a rescheduling of the units has to be performed.

In Fig. 2 the same process is shown from a more detailed perspective regarding the software agents. An exemplary energy unit is shown on the left hand side that is controlled by a unit agent. To the right of this unit agent, one additional unit agent is shown with less details as example for all other unit agents. Last, a market interaction agent is shown. Details on the different agent roles during DVPP setup are given in Section IV.

In the first phase of dynamic VPP aggregation, the agent *unit_agent_1* identifies relevant market products via an interaction with the market agent. With this information, it starts the VPP aggregation process. The result of this process is a (product-specific) DVPP consisting of a set of unit agents, with a designated *representative* and a DVPP schedule mapping the DEUs to operation schedules in such a way that the product can be fulfilled.

In the next phase (market interaction), the *representative* bids at the market. After market matching, it is informed about the product to be delivered. The *representative* communicates the needed contributions to all other unit agents within the DVPP. A unit agent might have proposed active power delivery of his unit in several DVPPs (e. g. for adjacent hourly time intervals, i. e. different power products). As all obligations are known to all unit agents within the DVPP after market matching, an optimization can use remaining flexibilities. For all DEU within the DVPP, updated forecasts and measurements can be used to optimize product delivery in this step, before configuring the units with these optimized schedules.

All unit agents have to follow the same task in the last phase, from unit schedule configuration until the product delivery is finished: They have to ensure the delivery of the DVPP active power product. Therefore, the unit agents continuously (e. g. on a minute base) check the unit's operational state and check it for schedule compliance. If a unit is not following the desired schedule and if the overall DVPP active power contribution will not fulfill the defined product as a consequence, a rescheduling is performed within the DVPP agents.

In the following sections we will focus on the algorithmic details of the aforementioned steps.

IV. Dynamic Aggregation

The problem of dividing the unit agents into several DVPPs can be generally described as coalition structure generation (CSG) problem. Goal of CSG is to find an optimal partition of a given set of agents A, referred to as coalition structure (CS). The elements of a coalition structure are coalitions (C) and can be evaluated using a value function $v(C)$, where the value of a coalition structure, $V(CS)$, is calculated as the sum of values of all comprised coalitions. Goal of a CSG algorithm is to maximize $V(CS)$. There are different algorithmic solutions for CSG problems, including dynamic programming, anytime optimal solution strategies and heuristic approaches [14]. The dynamic aggregation process described in the following provides a heuristic solution to the CSG problem. For a detailed description of the considered setting see also [15].

The dynamic aggregation of energy units takes place within a defined market area which is represented by a power grid comprising a set of connected DEUs. Each DEU is supervised by a unit agent as described in the previous section. General goal of the aggregation process is an optimized provision of active power on a global level. To this end, agents are generally able to cooperate in order to form coalitions and aggregate the capabilities of their supervised DEU. The purpose of each coalition is the provision of a day-ahead active power

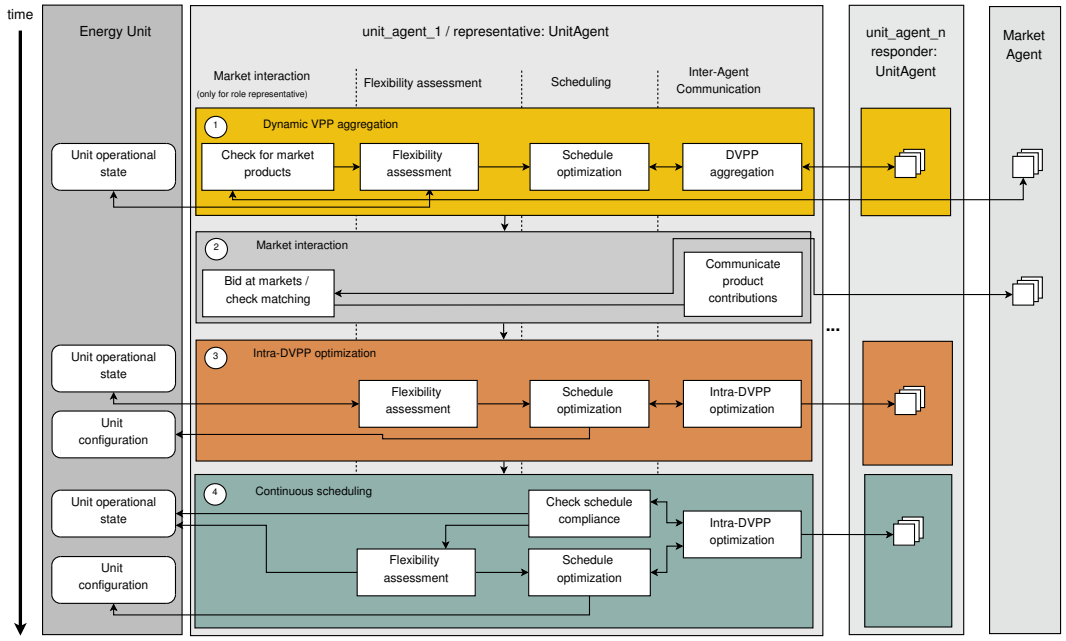


Fig. 2. Agent tasks for DVPP setup, market interaction and product delivery

product which is intended to be supplied within a corresponding product horizon reflecting the time of fulfillment. Thus, the process of DEU aggregation directly corresponds to the one of forming coalitions of supervising unit agents. Each agent contributes to an aspired power product by providing an amount of electrical energy according to the operational potentials of its energy unit. As optimization criteria, agents take the aspired target product as well as the respective costs of provision into account.

In accordance with the requirements and paradigms described in Section I, our proposed mechanism for unit aggregation is based on the principle of self-organization where agents form coalitions without external control of a superior instance. Moreover, aggregation takes place in a fully distributed and temporally flexible fashion, meaning that the organizational binding resulting from common product procurement is restricted to the provision of a provided product only and coalitions dissolve after their fulfillment. The mechanism is an iterative process, each cycle consisting of four activities which are carried out by each agent with the goal of forming coalitions with other unit agents and thus aggregating respective operational capabilities of their supervised energy units:

1) **Product portfolio generation:** In the course of the first activity an agent creates an individual product portfolio comprising a set of target products which it is generally willing to trade on the market. These target products satisfy all operational constraints of its supervised unit (like minimum operation time) as well as the product constraints obliged by the market (like maximum price/kWh). Generally, in this step several markets of same or different kind (like active power or reserve control markets) could be integrated into the decision making process in order to optimize benefit. However, our current work is restricted to a single day-ahead active power market only.
2) **Neighbourhood formation:** Given the set of target products as well as a distance function quantifying physical distance between DEUs in the grid, agents start forming neighborhoods which comprise potential cooperation partners for forming coalitions given the constraint that their supervised units are located within a specified range of physical proximity. This second activity allows unit agents to initially reduce communication and computation costs in the course of the actual aggregation process. Moreover, by taking the grid topology into account, coalitions are generally able to provide grid-sensitive power products within a specified area of the grid and thus to procure respective system services like redispatch capacities for congestion management.
3) **Coalition formation:** As third activity, agents start unit aggregation by forming coalitions within their afore defined neighborhoods in order to collectively fulfill common target products. In case an agent is not able to find a suitable coalition within its current neighborhood, it iteratively goes back to the second step and extends the scope of its neighborhood in order to include more potential cooperation partners.
4) **Payoff division:** Finally, after unit aggregation has finished and a coalition's product was accepted on the market (e. g. after clearance on an exchange), agents enter the last task and divide the payoff received from common product fulfillment among each other based on agreed criteria like contributed energy amount or reliability of procurement. To allow a fair division, the

payoff is distributed based on game-theoretic concepts [16].

V. Scheduling of Units in Planning and Operation

The scheduling of energy units within DVPPs can be regarded as 0-1 multiple-choice combinatorial optimization problem. In this type of problems, multiple sets or classes of elements (i. e. feasible operation schedules in our use case) are given, from which each exactly one element has to be chosen to form a solution. The goal is to find a solution that minimizes (or maximizes) a given objective function (e. g. power product fulfillment).

Many problems solved by multi-agent systems are modeled as distributed constraint optimization problems (DCOP). According to [17], in a DCOP, a number of independent agents each control the state of (a subset of) the variables in the system, with the joint goal of maximizing the global reward for satisfying constraints. If global constraints affect a larger subset of agents and the problem should be solved in a distributed manner though, classical DCOP methods are not feasible. Hence, in our concept, we use the Combinatorial Optimization Heuristic for Distributed Agents (COHDA) for all scheduling aspects throughout the process (see [18] for details including an overview on and discussion of alternative solutions for this problem).

A. Intra-DVPP optimization

In the considered use case of day-ahead power provision, there will be a significant time span between market matching and product delivery (e. g. 12 hours for the EPEX SPOT in Germany). This available time span will be used for an internal optimization within a DVPP prior to product delivery. Here, updated prognoses and measurements for each DEU can be utilized in order to tap the full potential of the unit's flexibility, which allows for optimizing the unit's schedule with respect to accuracy and reliability. For example, the operational state of a combined heat and power plant in combination with new measurements of e. g. the water temperature for an attached hot water storage are used to recalculate the unit's flexibility (the modelling of a units' flexibility is described in Section VI). With this new information, a rescheduling for the units within a DVPP is performed: The obligation of a DVPP is given in the form of a power product that comprises an active power profile for a defined planning horizon (e. g. the next 24 hours) as a series of constant amounts of active power (e. g. hourly intervals, as is common for the EPEX SPOT). The optimization goal in this phase is to find a schedule for each DEU of the DVPP such that the sum of all schedules matches the power product as close as possible. However, each unit agent must be permitted to decide itself which schedule it contributes. This way, economically or ecologically rooted soft constraints can be taken into account as secondary optimization goals while preserving privacy and autonomy of the participating units. Thus we employ the self-organizing heuristic COHDA as described in [19], [18] for this task as follows.

The key concept of COHDA is an asynchronous iterative approximate best-response behavior, where each unit agent reacts to updated information from other agents by adapting its own selected schedule with respect to the power product. In order to reduce the communication overhead of this distributed optimization problem, the unit agents are placed in an artificial communication topology (e. g. a *small world* topology), such that each unit agent is connected to a non-empty subset of other unit agents. To compensate for the resulting non-global view on the system, each $unit\ agent_i$ collects two distinct sets of information: on the one hand the believed current configuration γ_i of the system (that is, the believed current schedules of all unit agents), and on the other hand the best known combination γ_i^* of schedules with respect to the power product it has encountered so far. Recall that an agent initially only knows its own flexibilities, and the difficulty of the problem is given by the distributed nature of the system in contrast to the task of finding a common allocation of schedules. Thus, the agents coordinate via message exchange. Beginning with the *representative* of the DVPP, each $unit\ agent_i$ executes the following three steps, cf. [18]:

1) **(`update`)** When a $unit\ agent_i$ receives information from one of its neighbors (say, $unit\ agent_j$), it imports these information (γ_j and γ_j^*) into its own knowledge base by updating γ_i and, if better, replacing γ_i^* with γ_j^*.
2) **(`choose`)** If γ_i or γ_i^* has been modified in the previous step, the agent adapts its own schedule according to the newly received information, while taking its own local objectives into account. If it is not able to improve the believed current system configuration γ_i, the configuration γ_i^* will be taken instead. The latter causes $unit\ agent_i$ to revert its current schedule to the one stored in γ_i^* (note that γ_i^* contains a schedule for each agent in the system and $unit\ agent_i$ takes its own of course).
3) **(`publish`)** If γ_i or γ_i^* has been modified in one of the previous steps, the agent finally publishes its knowledge base (γ_i, including its own selected schedule, and γ_i^*) to its neighbors. Local objectives are not published to other agents, thus maintaining privacy.

The algorithm terminates when for all agents γ and γ^* are identical. At this point, γ^* is the final solution of the heuristic and contains exactly one schedule for each unit in the DVPP. With this information, the unit agent configures its respective DEU by setting the schedule.

B. Continuous scheduling

Once the internal optimization has finished, operation schedules for each DEU are known to the unit agents. These schedules can now be transferred to the respective DEU of each unit agent as set values.[1] However, incidents of several types may have rendered the chosen operation schedules

[1] Appropriate communication technology choice and information modeling is not subject of this paper. For example, data can be modelled using state-of-the-art international standards and protocols like OPC UA and CIM [20].

infeasible, like DEU breakdown, updated forecasts or the operation of the DEU for unforeseen services like system services. Therefore a rescheduling is needed in those cases, where the summed deviations of the DVPP's energy units enhinder product fulfillment. To detect this behavior, the unit agents continuously monitor their DEU. If product fulfillment cannot be guaranteed anymore, rescheduling is triggered. The scheduling heuristic COHDA described in Section V-A is used for this application as well, but additional constraints and optimization criteria have to be considered beside the target product to be delivered:

- **Local DEU contraints:** As the DEUs are already in execution of a given operation schedule, reconfiguration of the unit should take care of the DEU's current operational state. In the algorithmic framework presented here, surrogate models are used to cover this task: A simulation model is initialised with the current operational state of the DEU to deliver feasible sample schedules. Thus local constraints are covered by each operation schedule retrieved from this surrogate model.
- **Robustness:** The schedules generated during internal optimization may still hold severe uncertainty regarding their feasibility. With the product delivery period approaching or even started, robustness of a DEU for a chosen schedule becomes more important, as a repeated rescheduling by the agents and resulting reconfiguration of the DEU may result in suboptimal overall system performance. Therefore the weighting of robustness may increases over time depending on the specific facets of robustness important in the context of DEU scheduling, i. e. soft constraints within the operation of DEU and power grid feasibility margins.
- **Cost:** As long as there is still enough time left until product delivery, the cost of the schedules is the most important optimization criterion. A bad robustness value can be compensated by rescheduling. When product delivery has started though or not enough time is left for rescheduling, robustness can outbalance the costs if product delivery is threatened and thus other costs (e. g. for balancing energy) would severely cut the DVPP profits. Therefore the weighting of costs decreases over time.

The optimization function for continuous rescheduling therefore has to be formulated as time-dependent optimization function, where these factors are convexly combined and given hard constraints like product fulfillment and other criteria (e. g. power grid related criteria) are taken as side conditions. The details on this are subject to current work.

VI. Representing flexibilities with surrogate models

Real world scheduling problems often face nonlinear constraints. This set of constraints defines the shape of a region within the search space (a hypercube defined by operation parameter limits) that contains all feasible solutions. This feasible region might be arbitrary shaped or discontinuous and defines the region where to pick feasible solutions from. Several techniques for handling constraints during optimization have been developed. Nevertheless, almost all are concerned with special cases of non-linear programs or require a priori knowledge of the problem structure in order to be properly adapted [21]. A good overview can be found in [22] or [23].

At the same time, support vector machines and related approaches have been shown to have excellent performance when trained as classifiers for multiple, especially real world problems. Tax and Duin developed the support vector domain description as a one-class support vector classifier that is capable of modeling the region that is defined by some given training data [24]. We adapted this concept for integrating constraints into optimization in a way that allows for efficiently navigating the feasible region. The basic idea is to construct a mapping from the whole, unconstrained domain of the problem (the hypercube) to the feasible region to be able to automatically repair infeasible solutions during optimization. In this way, the scheduling problem is transferred into an unconstrained one by mapping any arbitrary solution onto a nearby feasible one.

Information about the flexibility of a DEU, i. e. the capability to alter energy production or consumption, is indispensable for coordinating processes within a DVPP. Planning for a product specific adaption of operations demands for a detailed model of a unit's scope of action. Taking into account all feasible alterations of operation, flexibility can be represented as the set of realizable (operable without violating any constraint) schedules. Unfortunately, a full assessment of this set is in general intractable. Depending on the time resolution of the schedule and possible operational settings of a unit, the number of theoretically realizable schedules can be in the order of some 10^{100}. Each unit has to obey individual technical, economic or user defined constraints in their operation that restrict the set of feasible schedules resulting in an individually shaped feasible region. Thus, the search space that defines feasible solutions of each unit forms an individually shaped feasible region. A mathematical model of the flexibility has to be derived repeatedly on demand as it depends on the current setting (e. g. operation state) and on recent forecasts (e. g. on thermal demand). Furthermore, because a DVPP continuously re-organizes in our approach, a mathematical optimization model for a DVPP cannot be determined statically in advance. Thus, with a newly formed DVPP the model for scheduling has to be re-built according to the participating units and their individual current flexibility. Hence, we use surrogate models for a unit's flexibility as proposed in [25]. The core of the model for the set of feasible schedules is a one-class support vector classifier trained with a set of operable example schedules. This flexibility model works as follows: Given a set of sample schedules, a description of the inherent structure of the feasible region of a unit is derived. After mapping the data to a high dimensional feature space by means of an appropriate kernel, the smallest enclosing ball in this feature space is determined. When mapping back this ball to data space, the pre-image of the ball forms a set of contours (not

necessarily connected) enclosing the given data sample (in our case: the feasible schedules of the respective DEU). The result of this procedure is a decision function that in general allows deciding on an arbitrary data point whether it belongs to the same region that contains the other data or not. In our use case, it allows testing a schedule wether it can be operated by the DEU or not.

But, for a controlled construction of solutions we have to go one step further, as we want to have a means for a goal-oriented search that allows us to systemetically find feasible schedules. In this way, we need a means that guides any algorithm where in the search space to look for feasible schedules. The advantage of our model is that it allows to generate a decoder that transforms the problem of distributed active power planning into an unconstrained one [26]. In general, a decoder is a constraint handling technique that imposes a relationship between feasibility and decoder solutions in order to give an algorithm hints on how to construct a feasible solution [23]. The flexibility of a unit is represented as pre-image of a high-dimensional ball. This representation has some advantageous properties. Although the pre-image might be some arbitrary shaped non-continuous blob in $\mathbb{R}^d$, the high-dimensional representation is still a ball and thus geometrically easier to handle. The relation is as follows: If a schedule can be operated without violating any constraint, it lies inside the feasible region. Thus, it is inside the pre-image (that represents the feasible region) of the ball and thus its image in the high-dimensional representation lies inside the ball. An infeasible schedule lies outside the feasible region and thus its image lies outside the ball. Additionally, we know some relations: the center of the ball, the distance of the image from the center and the radius of the ball. Hence, we can move the image of an infeasible schedule along the difference vector towards the center until it touches the ball. Finally, we calculate the pre-image of the moved image and get a schedule at the boundary of the feasible region: a repaired schedule that is now feasible. We do not need an explicit mathematical description of the feasible region or of the constraints to do this. Working with this decoder concept comprises two successive stages:

1) **A model/decoder training phase:** During the training phase (flexibility assessment) the decoder is build out of an set of example schedules derived from a simulation model of the unit.
2) **A successive planning phase:** Once the model is built, it can be (re-)used for assessing the feasibility of arbitrary schedules and for systematically generating feasible schedules with the decoder.

The latter is the main use case for the flexibility model: Whenever a prospective schedule is generated as a candidate solution to a specific unit, the decoder is used to convert this schedule into a similar schedule that is guaranteed to be operable by this DEU. The feasible version is used for scheduling. At the same time, performance indicators characterizing individual schedules with respect to different optimization goals are automatically preserved with this method [27]. Thus, after mapping a schedule (even with wrong or no associated performance indicators) to a feasible one, evaluation with respect to multiple criteria is possible. For distributed problem solving, the decoder can serve as a substitute for an often (particularly with regard to a fully automated generation in dynamic environments) hardly derivable mathematical model of a unit. The flexibility model automatically derives a means for generating feasible solutions from an unknown (to the agent) technical model. Hence, in our use case the flexibility model allows the agent for always working with operable schedules and thus with feasible solutions during coalition formation and scheduling without a need for a unit specific agent implementation.

VII. Agent model

In order to manage the repeatedly executed tasks of the four phases (DVPP aggregation, market interaction, intra-DVPP optimization, and continuous scheduling) the automaton depicted in Fig. 3 is guiding each agent through the process. Once started, the agent first executes the flexibility assessment task by querying the current state of the controlled unit and simulating possible flexibilities with its parameterization. The flexibility model and the decoder are built and provided for reuse in successive tasks. If the agent is not yet part of a DVPP, the agent takes part in the aggregation process that as a result assigns the agent to a newly formed DVPP. After forming the DVPP each agent participates in a continuous optimization process that aims at assigning a schedule to each agent's unit such that correct delivery of the product is ensured as reliable and at the same time as efficiently as possible. To do this the described optimization process is executed. The first execution is started at latest directly before product delivery. In case of an event that invalidates the current flexibility model due to changed assumptions or technical problems, the flexibility assessment has to be started again. With this new flexibility model, the optimization process may then be executed for rescheduling as a reaction to the event. The rescheduling may be triggered without a new flexibility assessment in case it is triggered due to the invalidation of another agent's flexibility model. After product delivery the DVPP's existence comes to an end and the cycle starts again.

VIII. Feasibility study

A. Simulation Environment

The purpose of distributed control concepts as described in this work is to realize an agent-based control of distributed energy units according to a current market situation, the current energy unit's state and the power grid's operational state. As a consequence, purpose of the evaluation system is to evaluate the effect of these distributed control concepts on the energy units and (for some applications) the power grid. Therefore a Smart Grid simulation has to be performed, following requirements regarding the reuse of existing models, a convenient and easy to reuse scenario specification, a well-defined API to real-world components and a synchronization concept and implementation between the multi-agent system

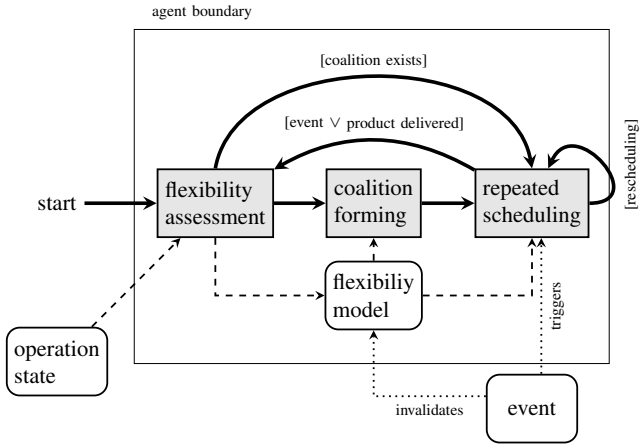


Fig. 3. Flow chart of the agent model with control flow (solid), data flow (dashed) and trigger (dotted)

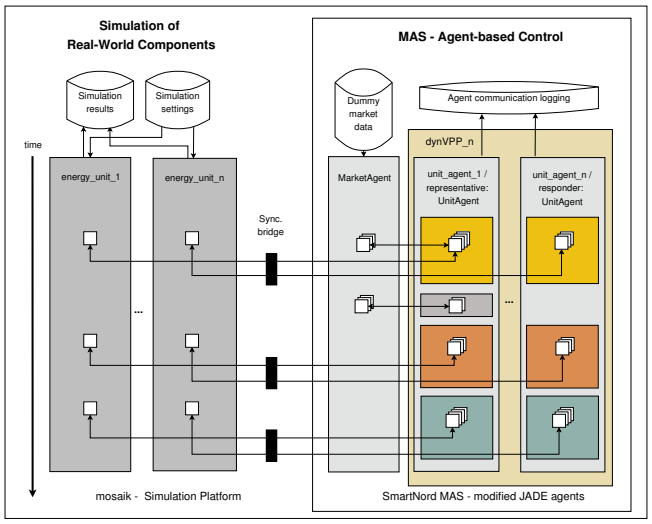


Fig. 4. Evaluation system with unit simulation using mosaik, multi-agent system for control of the VPP and synchronisation agent.

and the Smart Grid simulation. We use mosaik [28] as Smart Grid simulation framework that was built to meet these requirements. The chosen Smart Grid simulation framework has to be coupled with the distributed control technology framework chosen. In our case, we chose JADE [29] with some modifications.[2]

In Fig. 4 an overview on the evaluation system is given. Smart Grid simulation and agent-based control are completely separated. On the left hand side, the Smart Grid simulation with all simulation models for DEU is shown. On the right hand side, the MAS is shown. Within this part of the MAS, the agent-based control of the simulated real-world components (that is DEU in the use case chosen here) is realized. During simulation, the unit agents have to access the simulated energy units to set schedules and retrieve current measurement values. This access is realized using a synchronization bridge (depicted in in the middle of Fig. 4) offering a JADE interface. Besides data transfer, the bridge handles the sychronisation between Smart Grid simulation and MAS. The agents cannot distinguish being run in a simulation environment or in a real-world application. Within mosaik, the bridge allows to handle the MAS as an additional simulation component. In our system, dummy market data serve as input for the market agent to define products and realize the market matching once the order book is closed. The output data of both Smart Grid simulation and multi-agent system are stored in two HDF5-databases.

[2]As JADE schedules events in real-time, parts of the agent framework JADE had to be rewritten to realize this synchronization, thus allowing to run JADE as a MAS-simulation.

B. Experimental setup and results

To give an illustrative example for the integrated process chain within a feasibility study, we setup a scenario with 38 combined heat and power plants combined with thermal storage connected to households in a low-voltage grid. As product to be realized by the DVPPs we defined a product of 25 kWh from 2 p.m. to 3 p.m. on January 2nd, 2013, i. e. 15 min. time intervals 56 to 59 on day 1. The simulation was run from January 1st, 2013 (day 0) to January 2nd, 2013 (day 1) to cover the day-ahead market use case defined in Section III. The Smart Grid simulation was run with a stepsize of one minute, taking into account the weather conditions on the chosen simulation days.

We expected the following phases when running the MAS in the coupled simulation with mosaik: (1) Day-ahead flexibility assessment for all units and DVPP formation for defined product for day 1, and (2) Intra-day pre-delivery flexibility assessment for DVPP units and rescheduling of units for day 1.

We started the process of DVPP setup at 0 a.m. on day 0. Reassessment of flexibilities and rescheduling was started at 0 a.m. at day 1. In this setup, grid topology is reflected within DVPP setup as defined by the neighbourhoods (cf. Section IV).

In Fig. 5 the sample schedules for the initial flexibility assessment are shown for one energy unit ($unit_1$). For each sample schedule, the mean power value is plotted for all 96 operation schedule intervals. The distribution of power values over time is quite uniform. In the surrogate model trained with these samples (cf. Section VI), very different operation schedules can be found for the product horizon starting at interval 56 (2 p.m.). During DVPP formation, energy units aggregate to coalitions if their potential contributions fit the product needed (cf. Section IV). In the example given here units should aggregate to deliver a 25 kWh hourly product. For reasons of clarity, we only illustrate the product setup of one DVPP from this scenario. For the example DVPP chosen, 5 unit agents jointly deliver the defined product. On the left hand side of Fig. 6 the active power contributions of all energy units within the DVPP are shown as stacked chart over time for the product horizon (2 – 3 p.m.). The energy that would be delivered by the DVPP following these schedules is depicted on the right hand side. The product target of 25 kWh is not reached due to the tolerance settings within coalition

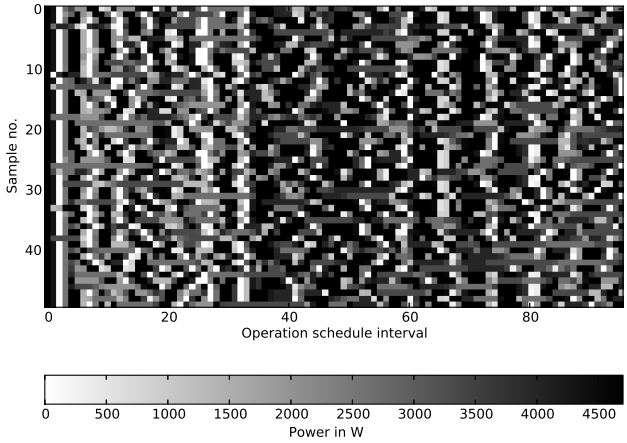


Fig. 5. Day-ahead sample schedules for $unit_1$

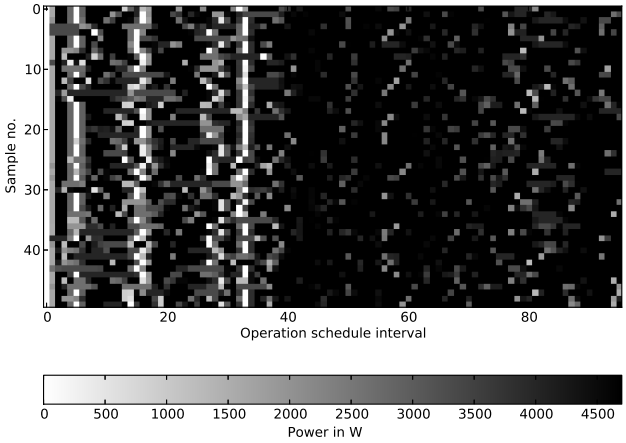


Fig. 7. Intra-day sample schedules for $unit_1$

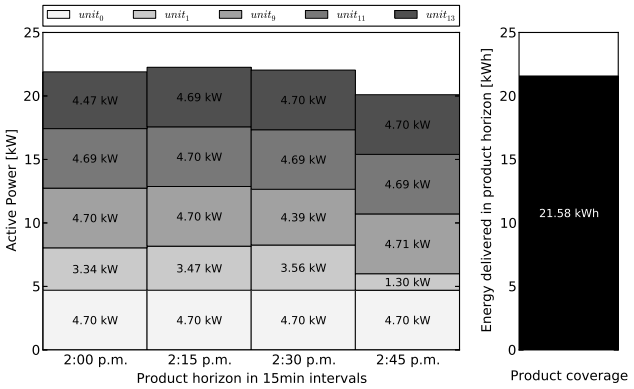


Fig. 6. Cluster schedule and product coverage after DVPP setup

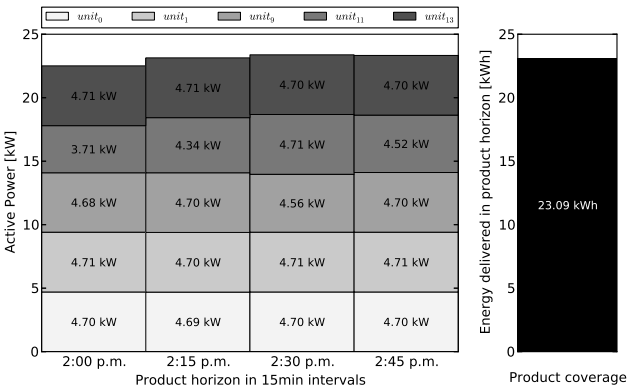


Fig. 8. Cluster schedule and product coverage after rescheduling

formation; these can be adapted individually for each product.[3]

A second flexibility assessment was triggered at 0 p.m. on day 1. Fig. 7 shows the samples taken from the simulation of the unit at that simulation time. If compared with the initial samples (Fig. 5), a much darker region can be detected at the product horizon (time intervals 56 – 59). Obviously, flexibilities in unit operation have narrowed down (e. g. due to the current capacity of the thermal storage or updated weather prognoses). The surrogate model trained with these samples will not produce schedules with power values lower than 3000 W in the chosen time horizon. We now take a closer look on the results of the planning phase using the surrogate models trained by the sample schedules from the second flexibility assessment.

In Fig. 8 the schedules after rescheduling using COHDA (cf. Section V-A) are shown. As can be seen, the product is now fulfilled better (23.09 kWh). One reason for this might be the changed surrogate model of $unit_1$: In the original plan, this unit has been scheduled with a contribution of 1.3 kW for time interval 59 (2:45 p.m.). This active power range cannot be retrieved from the surrogate model trained with the samples shown in Fig. 7, as an active power value lower 3000 W is not within the set of feasible schedules. Therefore, a new solution has been found by COHDA following both the updated flexibilities (modelled within the surrogate models of the energy units) and the target product tolerance margins.

[3]The evaluation of an optimal setting of these margins is part of a later evaluation of the overall system.

IX. Conclusion

In this paper we presented an agent-based control method for dynamic virtual power plants self-adapting their unit set and operational plan to be able to trade products on a power market. The three coordination steps of aggregation, schedule optimization, and continuous scheduling for DVPP control are integrated into the behavior of unit agents. These agents interact with each other and with the power market. Aggregation as well as scheduling of DVPPs is based on self-organization methods to achieve adaptivity to a changing set of units and new products traded on the market. So, DVPPs are self-coordinating, self-optimizing, and – within certain limits – self-healing. A cross-sectional technology for representation of the units' flexibilities is provided by a

surrogate model allowing to integrate new types of units easily into the coordination mechanism of DVPPs.

A simulation-based demonstrative example of the interaction of the coordination steps of a DVPP has been given in this contribution. Based on the current state of knowledge, DVPPs are a very promising approach to exploit flexibilities of decentralized units in a future power grid to support the integration of renewable energy resources. More and particularly more complex scenarios have to be studied to evaluate the performance, stability and dynamics of this control method. This will be a significant goal in our ongoing project cluster Smart Nord.

ACKNOWLEDGMENT

We thank Steffen Schütte and Stefan Scherfke for the technical support in setting up the coupled simulation environment using mosaik.

REFERENCES

[1] O. Abarrategui, J. Marti, and A. Gonzalez, "Constructing the Active European Power Grid," in *Proceedings of WCPEE09*, Cairo, 2009, pp. 1–4.

[2] A. Nieße, M. Tröschel, and M. Sonnenschein, "Designing Dependable and Sustainable Smart Grids – How to Apply Algorithm Engineering to Distributed Control in Power Systems," *Environmental Modelling & Software*, 2013. doi: 10.1016/j.envsoft.2013.12.003

[3] A. Nieße, S. Lehnhoff, M. Tröschel, M. Uslar, C. Wissing, H.-J. Appelrath, and M. Sonnenschein, "Market-based self-organized provision of active power and ancillary services: An agent-based approach for smart distribution grids," in *Complexity in Engineering (COMPENG)*, 2012. doi: 10.1109/CompEng.2012.6242953

[4] Bundesnetzagentur, "'Smart Grid' and 'Smart Market'," 2012, [Online; accessed 2014-01-24]. [Online]. Available: http://www.bundesnetzagentur.de/SharedDocs/Downloads/DE/Sachgebiete/Energie/Unternehmen_Institutionen/NetzzugangUndMesswesen/SmartGridEckpunktepapier/SmartGridPapier_EN.pdf

[5] F. Wu, K. Moslehi, and A. Bose, "Power system control centers: Past, present, and future," *Proceedings of the IEEE*, vol. 93, no. 11, pp. 1890–1908, 2005. doi: 10.1109/JPROC.2005.857499

[6] International Energy Agency, *Distributed Generation in Liberalised Electricity Markets*. OECD Publishing, 2002. ISBN 978-9-2641-7597-6

[7] S. Awerbuch and A. M. Preston, Eds., *The Virtual Utility: Accounting, Technology & Competitive Aspects of the Emerging Industry*, ser. Topics in Regulatory Economics and Policy. Kluwer Academic Publishers, 1997, vol. 26. ISBN 0-7923-9902-1

[8] D. Coll-Mayor, R. Picos, and E. Garciá-Moreno, "State of the art of the virtual utility: the smart distributed generation network," *International Journal of Energy Research*, vol. 28, no. 1, pp. 65–80, 2004. doi: 10.1002/er.951

[9] S. McArthur, E. Davidson, V. Catterson, A. Dimeas, N. Hatziargyriou, F. Ponci, and T. Funabashi, "Multi-agent systems for power engineering applications – Part I: Concepts, approaches, and technical challenges," *IEEE Transactions on Power Systems*, vol. 22, no. 4, pp. 1743–1752, 2007. doi: 10.1109/TPWRS.2007.908471

[10] R. R. Negenborn, Z. Lukszo, and H. Hellendoorn, Eds., *Intelligent Infrastructures*, ser. Intelligent Systems, Control and Automation: Science and Engineering. Springer, 2010, vol. 42. ISBN 978-90-481-3597-4

[11] S. D. Ramchurn, P. Vytelingum, A. Rogers, and N. R. Jennings, "Agent-based homeostatic control for green energy in the smart grid," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 4, pp. 35:1–35:28, Jul. 2011. doi: 10.1145/1989734.1989739

[12] G. Anders, F. Siefert, J.-P. Steghöfer, H. Seebach, F. Nafz, and W. Reif, "Structuring and Controlling Distributed Power Sources by Autonomous Virtual Power Plants," in *IEEE Power and Energy Student Summit (PESS 2010)*. IEEE Power & Energy Society, 2010.

[13] S. D. Ramchurn, P. Vytelingum, A. Rogers, and N. R. Jennings, "Putting the 'smarts' into the smart grid: A grand challenge for artificial intelligence," *Commun. ACM*, vol. 55, no. 4, pp. 86–97, Apr. 2012. doi: 10.1145/2133806.2133825

[14] T. Michalak, J. Sroka, T. Rahwan, M. Wooldridge, P. McBurney, and N. R. Jennings, "A distributed algorithm for anytime coalition structure generation," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1*, 2010. doi: 10.1.1.153.8211 pp. 1007–1014.

[15] S. Beer and H.-J. A. Appelrath, "A formal model for agent-based coalition formation in electricity markets," in *2013 4th IEEE/PES Innovative Smart Grid Technologies Europe (ISGT EUROPE)*, 2013. doi: 10.1109/ISGTEurope.2013.6695442 pp. 1–5.

[16] S. S. Fatima, M. Wooldridge, and N. R. Jennings, "A Linear Approximation Method for the Shapley Value," *Artificial Intelligence*, vol. 172, no. 14, pp. 1673 – 1699, 2008. doi: 10.1016/j.artint.2008.05.003

[17] A. C. Chapman, A. Rogers, N. R. Jennings, and D. S. Leslie, "A unifying framework for iterative approximate best-response algorithms for distributed constraint optimization problems." *Knowledge Eng. Review*, vol. 26, no. 4, pp. 411–444, 2011. doi: 10.1017/S0269888911000178

[18] C. Hinrichs, M. Sonnenschein, and S. Lehnhoff, "Evaluation of a self-organizing heuristic for interdependent distributed search spaces," in *ICAART 2013 – Proceedings of the 5th International Conference on Agents and Artificial Intelligence*, J. Filipe and A. Fred, Eds., vol. 1. Barcelona, Spain: SciTePress, 2013. doi: 10.5220/0004227000250034. ISBN 978-989-8565-38-9 pp. 25–34.

[19] C. Hinrichs, J. Bremer, and M. Sonnenschein, "Distributed Hybrid Constraint Handling in Large Scale Virtual Power Plants," in *2013 4th IEEE/PES Innovative Smart Grid Technologies Europe (ISGT EUROPE)*. IEEE Power & Energy Society, 2013. doi: 10.1109/ISGTEurope.2013.6695312

[20] S. Rohjans and M. Specht, *OPC UA: An Automation Standard for Future Smart Grids*, ser. Standardization in Smart Grids. Springer, 2013, ch. 12. ISBN 978-3-642-34916-4

[21] Z. Michalewicz and M. Schoenauer, "Evolutionary algorithms for constrained parameter optimization problems," *Evol. Comput.*, vol. 4, pp. 1–32, March 1996. doi: 10.1162/evco.1996.4.1.1

[22] C. A. Coello Coello, "Theoretical and numerical constraint-handling techniques used with evolutionary algorithms: a survey of the state of the art," *Computer Methods in Applied Mechanics and Engineering*, vol. 191, no. 11-12, pp. 1245–1287, Jan. 2002. doi: 10.1016/S0045-7825(01)00323-1

[23] O. Kramer, "A review of constraint-handling techniques for evolution strategies," *Appl. Comp. Intell. Soft Comput.*, vol. 2010, pp. 1–19, January 2010. doi: 10.1155/2010/185063

[24] D. M. J. Tax and R. P. W. Duin, "Data Domain Description using Support Vectors," in *European Symposium on Artificial Neural Networks – ESANN*, 1999, pp. 251–256.

[25] J. Bremer, B. Rapp, and M. Sonnenschein, "Encoding distributed Search Spaces for Virtual Power Plants," in *IEEE Symposium Series in Computational Intelligence 2011 (SSCI 2011)*, Paris, France, 4 2011. doi: 10.1109/CIASG.2011.5953329

[26] J. Bremer and M. Sonnenschein, "Constraint-handling for optimization with support vector surrogate models – a novel decoder approach," in *ICAART 2013 – Proceedings of the 5th International Conference on Agents and Artificial Intelligence*, J. Filipe and A. Fred, Eds., vol. 2. Barcelona, Spain: SciTePress, 2013. doi: 10.5220/0004241100910100. ISBN 978-989-8565-38-9 pp. 91–105.

[27] ——, "Automatic Reconstruction of Performance Indicators from Support Vector based Search Space Models in Distributed Real Power Planning Scenarios," in *Informatik 2013, 43. Jahrestagung der Gesellschaft für Informatik e.V. (GI)*, ser. LNI, M. Horbach, Ed., vol. 220. Koblenz: GI, 2013. ISBN 978-3-88579-614-5 pp. 1441–1454.

[28] S. Schütte, S. Scherfke, and M. Tröschel, "Mosaik: A Framework for Modular Simulation of Active Components in Smart Grids," in *1st Int. Workshop on Smart Grid Modeling & Simulation*. Brussels: IEEE, 2011. doi: 10.1109/SGMS.2011.6089027

[29] Telecom Italia S.p.A., "JADE (Java Agent Development Framework)," 2014, [accessed 2014-06-15]. [Online]. Available: http://jade.tilab.com/

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1515–1523
DOI: 10.15439/2014F215
ACSIS, Vol. 2

El Farol Bar problem, Potluck problem and electric energy balancing – on the importance of communication

Weronika Radziszewska
Systems Research Institute
Polish Academy of Sciences
Warsaw, Poland
Email:Weronika.
Radziszewska@ibspan.waw.pl

Ryszard Kowalczyk
Swinburne University of Technology
Melbourne, Australia
Email: rkowalczyk@swin.edu.au

Zbigniew Nahorski
Systems Research Institute
Polish Academy of Sciences
Warsaw, Poland
Email: Zbigniew.
Nahorski@ibspan.waw.pl

***Abstract*—Power balancing is an important issue when microgrids in island mode are considered. It requires active, real-time decision making to minimise the imbalances. El Farol Bar and Potluck problems are artificial problems designed to challenge the decision making process in a situation of limited information. They are theoretical, ill-defined problems designed to show that rational decision making in a situation of scarce information can give worse results than non-rational behavior. Potluck problem can be compared to electric power balancing in microgrids. But due to reduced complexity and constraints of theoretical problems and differences in goal functions the approach has to be in each case different. The study of these differences in this paper shows the importance of exchange of information between participants (or agents); it also suggest what type of information are necessary in what conditions. The amount of information that needs to be exchanged depends on the relation between participants (agents): the less cooperating their relation is, the less information they are willing to exchange.**

I. Introduction

THE POWER grids enter a new era, where the ecological impact and power security are playing an important role in their development. The introduction of relatively cheap micro sources has created a concept of microgrids: a localized part of the power network that usually is equipped with micro power sources and can be disconnected from the external power grid (operate in island mode). With new technologies, new challenges appeared, and among them management of power in the microgids. This task is difficult due to a relatively large changeability of production and consumption of power in small scale grids. The problem of power balancing became especially important for the development of microgrids operating in island mode. In such case, imbalances can lead to an ineffective system, to waste of power or even to damages of the devices. This problem is a subject of many research, where various methods have been developed for balancing and for minimizing the imbalances.

The power management can focus on the consumption side, and is called the demand side management [1] or demand response [2]. The idea behind it is to convince the power users to adjust their behavior patterns according to the energy situation in the grid or to allow some information system to do it automatically. An alternative is the supply side management, which requires managing the operating point of power sources, including switching on and off power production devices, to compensate for non-controllable production from renewable power sources. In this work, only supply side management will be considered.

Enumula and Rao described the Potluck problem in [3] and pointed out its similarity to the balancing of electric power. Both problems deal with equalizing supply and demand in an iterated way, where the supply can be only approximately assessed and balancing depends on decisions of many suppliers. The theoretical Potluck problem is claimed to be a generalization of El Farol Bar problem described in [4], where it is shown that, when there is lack of information, the inductive reasoning gives better results than rational reasoning. The theoretical Potluck problem is very constrained and does not allow any flow of information between participants, which make the problem ill-defined. An inductive reasoning based algorithm provides an acceptable solution to the Potluck problem, in the sense that the solution oscillates around the desired outcome. For energy balancing, such a solution is not good enough: it implies that the system is constantly experiencing imbalances. Relaxing the constraints of the Potluck problem allows for a more exact solution to be found using a rational method; the analysis of which constraints need to be relaxed in the Potluck problem provide valuable information to develop a scheme in which balancing in a microgrid can be solved.

In this article, a comparison of the balancing problem, the El Farol Bar problem and Potluck problem is presented. Although it is claimed that they are similar, there are major differences that require different approach to each of them. In section II a description of the power balancing and its importance is presented. Section III explains the theoretical El Farol and Potluck problems. In section IV a comparison of the problems is presented. Section V deals with real life approaches to the power balancing issue. The final section VI concludes the paper.

978-83-60810-58-3/$25.00

II. BALANCING ELECTRIC ENERGY IN MICROGRIDS

High voltage grids, unlike low-voltage microgrids, have different properties. Large grids require an approach where power flow dynamics are considered, as the power in large grids have to travel large distances and the power loses are significant. Such networks have also much more inertia and delays in changing the operation levels. When a node of such grid is not balanced (there is either an overproduction or a deficit of power) it affects all connected nodes and a number of nodes have to compensate, as the power disturbance can travel through the network, even to distant nodes. Microgrids are smaller and react faster to control requests, so they are manageable in an easier way, without considering the effect of distance, induction of power in the grid or power losses.

A microgrid is a part of an electric grid that potentially can be disconnected. Very often, microgrids are equipped with power sources (such as e.g. gas microturbines, micro wind turbines, photovoltaic panels) or power storage (e.g. batteries, flywheels) and can work in an island mode. In the island mode operation, in a microgrid it is necessary to balance the production and consumption of power to maintain the quality factors of the electric current; these guarantee the safety of the devices in the microgrid. A symbolic picture of microgrid elements is presented in Fig. 1. If a microgrid has an abundance of power production, then energy has to be wasted and/or production limited; in case of shortage of energy, some of the consuming devices should be switched off according to importance or preference. The difficulty is that a decision has to be made and it should follow the constantly changing conditions in the grid. The faster the decision is made, the less power is wasted and the safer the devices are. Power produced by some renewable sources (especially micro sources, which might lack the ability to manage their operating point) fluctuates dynamically due to sudden changes in e.g. wind and solar irradiance. Predictors, to some extent, can forecast the production and help minimizing the imbalances, but the predictions are not perfect. Consumption of energy is also very changeable and often unpredictable, especially in small microgrids, where a single device can make a noticeable difference in overall power usage. That means that the actions of a single human being can make a noticeable disruption from a typical daily power usage profile.

Balancing should make the amount produced ($s(t_k) = \int\limits_{t \in t_k} s(t)dt$), for a given time ($t_k$), equal to the amount that can be consumed ($d(t_k) = \int\limits_{t \in t_k} d(t)dt$) at that time. The real energy balancing is a continuous process, but from the operational point of view it can be quantified to a number of short time periods t.

$$\sum_{i=0}^{n} s_i(t_k) = \sum_{j=0}^{m} d_j(t_k) + L(t_k),\ t_k \in T \qquad (1)$$

where $n \in N$ is the number of active producers and $m \in M$ is the number of active consumers. The losses of power during transmission ($L(t_k)$) are not considered: they are relatively small, their amount depends on network structure and their absence does not influence the theoretical solution, but allows for a simplification of the model.

Balancing is possible due to the existence of controllable devices (their operation point can be changed by the energy management system) and the ability of switching off or on a part of the consumption. In most real life installations, a microgrid is connected to an external power grid, which can provide or absorb a large amount of power. In large power grids, a constant reserve of production power is kept in order to cover occurring imbalances. Due to the high costs of this type of installation, small microgrids usually do not have such a reserve.

The presence of the power storage units, e.g. batteries, can facilitate the balancing, as they provide a time and power buffer for the management system. Power storage units are generally much faster than controllable power sources when it concerns changing the amount of taken or given energy. Extremely fast operating storage units such as flywheels can smooth the sudden peaks of power and compensate for short power loses. Large enough capacity of power storage devices in the microgrid can solve a lot of issues, even completely eliminating the imbalances. Detailed analysis of influence of power storage can be found in [5]. Storing the power unfortunately results in loses of power, a high cost of installation of storage units and, in some cases, a necessity of replacing them relatively often. In microgrids, large capacities of storage units are not common due to high costs. The most frequently considered devices are batteries, flywheels and superconductors.

The problem of balancing a microgrid is of interest to many research teams. General architectures of energy management systems might be found in [6], [7] and [8]. Details of the algorithm of the market based short-time balancing can be found in [9].

A microgrid in general can consist of producers, consumers and prosumers. Each of these can be controllable or uncontrollable. Uncontrollable devices are those which are not manageable by computer system, to this category are included most of the power consuming devices and small renewable power sources (in which power production depends on weather conditions). The balancing problem reverts to a decision problem of setting the operation point of controllable devices in the microgrid, so that supply and demand equal according to equation (1). To simplify a model, all uncontrollable devices can be aggregated to a single value: this value is either 0 (perfect balance of uncontrollable devices), positive (behaves as producer) or negative (behaves as consumer). Here, the aggregated device is assumed to be a consumer, in order to avoid a situation of overproduction by uncontrollable producers. Such situation needs special actions (e.g. removing an uncontrollable producer, wasting power, etc.), which are not common situations in the typical balancing problem.

The power sources have physical limitations: a minimal and maximal operating point, a time necessary for changing the operation point, etc. Managing a controllable power source

means deciding if the device will be active in the next time period ($s_i(t_k)$), and if so, determine the amount of power it will provide.

Limited information about the amount of consumption and production is a problem for balancing. The consumption is changeable in time: it is the sum of consumption of a number of small devices, where each of them may have different usage patterns. While patterns and cycles, such as daily or weekly, are usually visible in the amount of power usage, the exact amount can only be roughly predicted.

The production is the sum of the decisions of all producers which operate in the microgrid. They may or may not know how many producers are present in total and decide to participate in balancing, but in any case they can only estimate the amount of power produced by all other sources. The information shared between producers is a property of a used scheme, which can depend on the level of cooperation, the size of the microgrid or other factors, such as cost of power production, ownership, regulations, etc. In microgrids with one owner, there can be full cooperation with full flow of information; allowing for central balancing to be used. When the competition of producers is present, the flow of information may be constrained to the minimal level that is necessary for the process. When a microgrid is really small, it is helpful to know the physical limitations of the units to predict the amount of produced power.

Other factors also play important roles. When for instance the cost of operation of a power source is considered, some sources are more profitable than others. Preference to a source can be modeled by introducing the cost, which does not necessarily match the real world cost, but allows for a preference ordering that can be dynamically adjusted. Such value in market schemes is sufficient for achieving effective balancing as presented in [9]. In microgrids operating in the synchronous mode (with connection to an external power provider) the costs and profits usually dominate the decision about the production – the balancing is not the main goal of the system but becomes a constraint. In extreme cases, a certain level of imbalance might be tolerated even though it means cutting off selected consumers from the grid. Such a solution is not considered in this article. Here the microgrid is considered to be in an island mode only.

Amount of public information and what information are being exchanged is an important problem. For various reasons, as e.g. safety, competition, willingness to make profit, the producers tend to keep certain information private. The lack of information exchange can make it impossible to perform balancing. Such situation was considered by Arthur in [4], where a method to deal with such ill-defined problem is suggested. The extension of this problem, the Potluck, considers the supply and demand equalization with almost lack of information exchange.

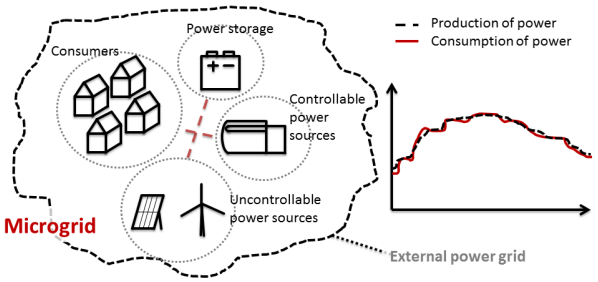


Fig. 1. Schema of the microgrid.

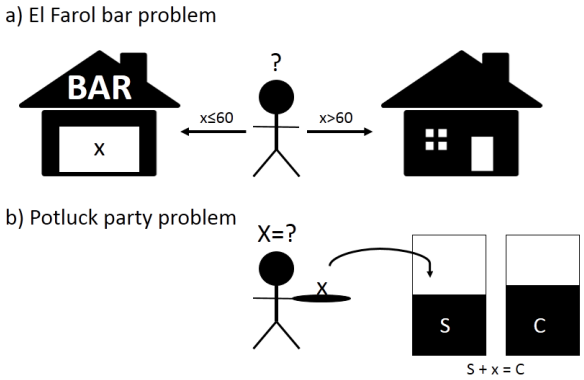


Fig. 2. Graphical representation of El Farol bar problem (a) and Potluck problem (b), S is a total supply and C is the value of consumption.

III. El Farol and Potluck Problem

A. Agents and communication

To achieve common understanding of the term agent in this article, an explanation is required. In the computer science, an agent is an autonomous programmable unit that interacts with its environment and tries to fulfill given goals. Agent can be also understood as a representative of an entity, that takes decisions to achieve given goals. These two definitions are similar – both underline the autonomous decision making feature. In this article, the term agent is used in more general sense – as an entity that can make decisions. In this context, any exchange of information can be called communication, without further description on how such communication is made.

B. El Farol Bar Problem

The El Farol Bar problem (or Santa Fe Bar Problem) was introduced by Arthur in 1994 [4]. The problem was inspired by a real bar in Santa Fe, which was very popular during Thursday nights. But if too many people decided to go to the bar to enjoy the music, it was too crowded. Arthur defined the problem as follows. If there are not more than 60 people in the bar, the people inside enjoy being there. Otherwise they feel better at home. This problem is illustrated in Fig. 2(a). So a participant is considered a winner if she/he goes to the bar while it is not crowded, or if she/he stays at home when the bar is crowded. In the El Farol Bar problem, the participants' goal

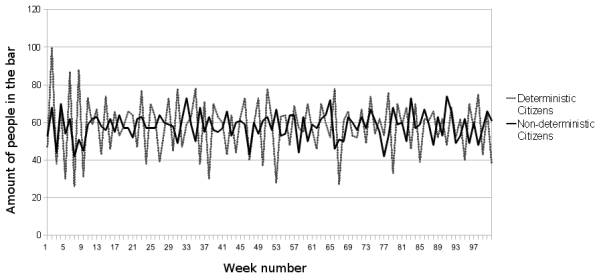


Fig. 3. Attendance to the El Farol bar of deterministic and non-deterministic citizens.

is to win as many times as possible, where the goal function $g_i(t)$ of i-th participant in the t-th night is defined by:

$$g_i(t) = \begin{cases} 1 & \text{if (go to the bar and the bar not crowded)} \\ & \text{or (not go and the bar is crowded)} \\ 0 & \text{if (not go and the bar not crowded)} \\ & \text{or (go and the bar is crowded)} \end{cases} \tag{2}$$

The participants do not know how many of them are in the city, they are not communicating with each other and they have no idea what other people want to do. The only information available to them is the historic attendance: each participant knows how many people were in the bar during the last weeks. In a such defined problem there is no win-win solution - when in the bar is 60 people, the ones remaining at home loose, when there is 61 people in the bar, this 61 people loose.

In this situation, there is not enough data to make a deductive, rational decision, which makes the problem ill-defined. In [4], an inductive reasoning scheme is proposed. This is an idea taken from psychology: people are very often facing ill-defined problems and humans cope with this situation by looking for patterns and similarities in other situations. If a person would be asked the reason for going to the bar, possible answers could be: 'last week it was empty so this week it will be the same', 'last week it was full, so this week it will be empty' or just 'because I want to go'. From the game theory and mathematical analysis point of view these answers are not reasonable, but due to lack of information they are as good as any other. Humans often do not analyze all possible actions deeply, but make shortcuts and take non-optimal decisions, sometimes due to undefined reasons. It is logical from the evolution point of view, as taking decisions fast has been more crucial for survival, than being indecisive and not performing any actions. Arthur (in [4]) assumed that each person has its own way of predicting the attendance in a bar – they have a set of simple predictors. The predicted attendance might be: an average of the last few weeks; the same as last week; the same as 3 days ago (cycle detector) or assumption that the bar will always be half empty. Each person also knows the attendance from few last weeks. So a going or not-going decision depends on the known history of attendance and ones own hypothesis. What is more, participants can choose their predictors from a pool, according to the success rate of a considered predictor (how many times it gives a good advice). Surprisingly, the simulations show that the attendance in the bar is oscillating around the chosen maximal comfortable number of the participants in the bar. An example of the bar attendance in this problem is presented in Fig. 3. Arthur called it inductive reasoning method and defined it as follows. When there is a lack of knowledge to take a reasonable decision, one should use simple models that worked best in the past, and after each iteration of the process evaluate the models.

An interesting feature of this approach is that starting from some defined conditions and following totally deterministic rules, the outcome is a sequence of attendance that resembles a stochastic process. The amount of people in the bar is oscillating around 60. It is explained by the fact that citizens choose the predictors that performed best, this creates a self-regulating system where the number 60 is a natural attractor.

It is worth to notice that if the citizens know how many of them are in the city, they can solve the problem quite easy by coming to the bar in cycles. This solution was described in [10].

C. Potluck Problem

The Potluck problem was described and defined in [3]. A potluck is a party where every guest brings some food for everyone to eat. If the food is in an excess, the guests feel uncomfortable, as their food has to be thrown away. On the other hand, if there is a deficiency of the food, guests are hungry and unhappy. The perfect situation would be to have the exact amount of food, but the appetites of the guests depend on many factors and can vary between parties. So, without communication guests have to guess the total amount of food they have to bring, not knowing what strategy other guests will adapt, see Fig. 2(b) for an illustration.

In the Potluck problem, the goal function is to balance the supply and demand. Assuming that demand is something out of control, the goal function of l-th guest can be defined as, see [3]:

$$g_l(t) = \begin{cases} 1 & \text{if } \sum_{i=0}^{n} s_i(t) = \sum_{j=0}^{m} d_j(t) \\ 0 & \text{if } \sum_{i=0}^{n} s_i(t) \neq \sum_{j=0}^{m} d_j(t) \end{cases} \tag{3}$$

The notations are explained in Table I. A guest is in the winning position when the sum of supply is equal to the sum of demand. But a guest has no means to communicate with other guests to inquire about the amount of food they plan to bring or the food they want to eat. This lack of information makes the problem ill-defined, where the rational reasoning does not help in winning of any of the guests. Enumula and Rao ([3]) define rational reasoning as applying the best strategy in given situation, that is with the assumption that the consumption level will be the same in future as in the last time. This show that it leads to increased oscillation of supply. If all guests make this assumption, they will take similar decisions, which will lead to an exaggerated change in the supply of food and the balance is never reached. The way to prevent it is by introducing different strategies for each

of the participants, hoping that at least to some extent the undersupply and oversupply will balance.

A method to deal with this problem is presented in [3]. It is a non-rational approach similar to the inductive reasoning described in [4]. As was mentioned before, rationality, according to Enumula and Rao [3], is to apply the best strategy according to the present knowledge. In the Potluck problem, the rational action is to act as if the supply has not changed since the last party. The non-rational approach is to allow participants to take an action that is assuming a certain change in the future supply (usually not explained by analysis of the problem). Participants have a set of simple predictors with assigned weights, that forecast the level of consumption. The decision is made on the basis of a weighted sum of predictors response (weighted majority algorithm). After each party, the predictors are evaluated and weights are adjusted accordingly. Enumula and Rao [3] called it a non-rational learning algorithm. In the cited article prediction of supply side is not considered.

Enumula and Rao claim in [3] that the Potluck problem is a generalization of the El Farol Bar one. But actually the points of view of decision-makers and the goal functions are in both problems different. The personal goal function in the El Farol Bar problem is given by equation (2). It is clearly an egoistic goal, which does not consider the well-being of other participants. Decisions of a participant are influenced by the actions of others, which can be interpreted as influencing the decision-maker, but it is not done intentionally. A participant has no intention to make the bar full or not, because in both situations there is a possibility of winning. Arthur [4] underlined that participants are independent agents, that are following their goals, even not being aware how many of participants are in the system.

In the Potluck problem, the goal function is to balance supply and demand, as described by equation (3). The goal can be defined as a global goal function which means that it is a type of a social welfare function. Unlike the El Farol Bar problem, it does not consider a personal gain or loss, but the sum: all participants win or lose. In the El Farol problem, if the bar is crowded, the people in the bar loose, but the people that did not go to the bar win. An analogy to social welfare in the El Farol Bar problem would be the situation where the citizens are trying to reach 60 people in the bar every week, and in case when there are more or less attendees everyone looses.

The question arises if these problems are really equivalent although the goals of the participants are different. The methods of approaching them are similar, but the problems complexity is changing in them when some exchange of information is introduced. In case of a personal goal, adding knowledge about the decisions of others (by communication) only introduces complications in decision making: the agent has to actively make effort to be in the winning position. To clarify this statement a following scenario can be considered: there are 100 participants in the El Farol bar problem, but just 99 of them has some media of communicating their decisions, e.g. announcing it on the social network. None of them knows

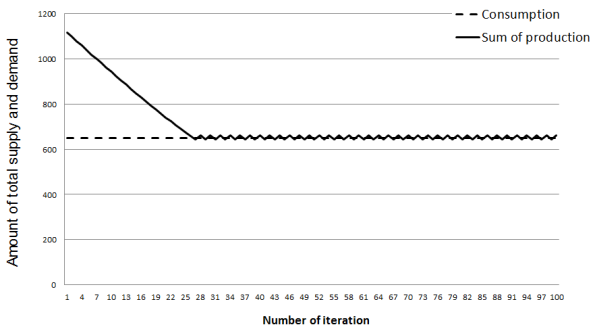


Fig. 4. Potluck problem simulation with linear consumption, with n=100.

what the 100th participant will decide, and this participant does not know the decisions of others. By communicating each other, the participants can agree to perform a schema that will ensure fair amount of winnings for each agent. They can agree that 59 of them are going to the bar and 40 staying home (the agents that are going to the bar can change every week, which would mean introducing going to the bar in cycles). This is a solution where the winner group is the largest, independent of the decision made by the isolated person. But it requires of participants to make concessions for some kind of social fairness. In the original problem agents are assumed to be myopic and egoistic, which does not allow them to cooperate. So, seeing the situation, participants staying home will be willing to change their decision. If this happens, the situation will change again and the decisions of the participant will also change. That would trigger a set of changes that would lead to an apparently chaotic behavior. A stopping condition may be applied, e.g. it might be the time (an hour of going to the bar is defined) or the number of decision changes. When the decision making process is closed the number of citizens in the bar is very likely to be not optimal. The outcome will show pseudo stochastic oscillations around the number of 60 people going to the bar, even when almost everything is known. In the Potluck problem the behavior in this scenario is different. Information about the amount of food brought by 99 out of 100 people suggest their predicted consumption level and all participants try to minimize the error of prediction. After a number of iteration the 99 participants can predict the production level of the 100th participant and consider his decision. Imbalance is still present, but the oscillations are relatively smaller. Socially aware agents are more likely to cooperate, make concessions and negotiate their decisions. Introduction of communication to the problem makes it possible to reason rationally.

IV. From Potluck to power balancing

Table I presents a comparison of the theoretical El Farol Bar problem, the Potluck problem and a practical problem of power balancing. The problems seem very similar, but a quick analysis shows main differences which cause that distinct methods of facing these problems have to be considered. The theoretical problems are very simplified and constrained. The

TABLE I
COMPARISON BETWEEN PROBLEMS OF EL FAROL, POTLUCK AND POWER BALANCING.

Symbol	El Farol Problem	Potluck Problem	Energy balancing
T	set of weeks	set of weeks	set of time periods
N	set of participants	set of guests	set of suppliers
M	equal to 1	set of consumers	set of energy consumers
$s_i(t)$	binary decision of going or not	amount of brought food by i-th guest	amount of energy produced by i-th supplier
$d_j(t)$	constant	amount of food expected by consumer j	amount of energy demanded by consumer j in time t
$S(t) = \sum_{i=0}^{N} s_i(t)$	total attendance in the bar in a week t	amount of food brought to the party	amount of energy produced by all suppliers
$D(t) = \sum_{j=0}^{M} d_j(t)$	constant	total demand of food	total consumption
$P_i(t)$	prediction of the attendance to the bar in a week t	prediction of the amount of total consumption in a week t	prediction of total consumption in time t

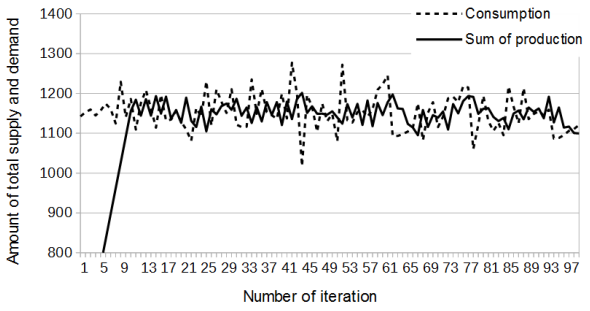


Fig. 5. Potluck problem simulation with random consumption, with n=100.

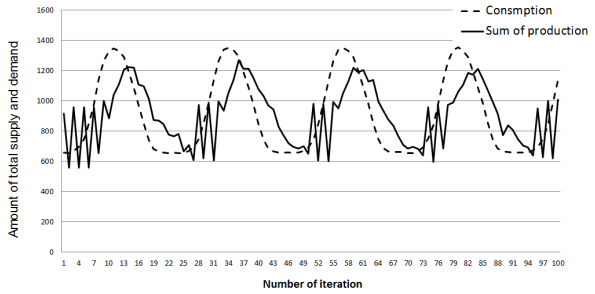


Fig. 6. Potluck problem simulation with sinusoidal consumption, with n=100.

most limiting constraint is that agents are banned from exchanging information. The Potluck problem can be expanded with additional constraints that resemble physical limitations that are present in the power balancing problem (e.g. minimal operating point, maximal operating point, latency of operating point change, etc.), but these do not significantly change the problem considered: it is still an ill-defined decision problem, the additional constraints do not simplify nor complicate it.

In the Potluck problem, lack of information about power production is equally problematic as lack of knowledge about its consumption. Tests have been made using a non-rational learning algorithm with different consumption patterns: the performance of the algorithm with a random consumption is shown in Fig. 5, with a fast changing sinusoid consumption in Fig. 6 and with a linear consumption in Fig. 4. Several categories of predictors have been used in the calculations:

- average demand over the last k periods,
- randomly chosen value of demand from the last k periods,
- choosing the demand from $t - k$ period, this predictors are cycle detector, it can detect cycles of 2, 3, 5 periods,
- mirror image around the average of the last k periods,
- the same as the previous period,
- trend over the last k periods,
- median of the last k periods,
- weighted solution over the last k periods - the random k weights are chosen: $w_1, w_2, \ldots, w_k$, where $\sum_{i=1}^{k} w_i = 1$ and the prediction is calculated as: $\sum_{i=1}^{k} D(t-i)w_i$.
- the smallest value of demand out of the two last periods,
- the larger value of demand out of the two last periods.

It is clear that a less changeable consumption makes it easier to reduce imbalances, as predictors work better. However, even for linear consumption agents could not fully balance demand and supply. The oscillations are still visible.

The reason for this is that suppliers use the same algorithm and therefore take similar decisions based on the same information, which in turn leads to overcompensation. This is logical, as a supplier has no knowledge of other suppliers and tries to solve the imbalance by itself. In some situations (e.g. the oscillations in the linear case) the result can be improved by introducing additional conditions to the agents' logic, but such specialization would decrease overall performance in the general case. A better solution is to allow for communication between the suppliers.

There are many ways in such communication can be introduced. The simplest case considers publishing information for all agents. One way to prevent big oscillations is to limit the number of suppliers that can change their decision, such that not all suppliers will react to the imbalance. This automatically limits the total change. It can be achieved by introducing tokens to tell a supplier that it is allowed to change its output, and publishing which agents have the tokens in the given iteration. Preferring certain suppliers over others becomes a

matter of a central body that has to decide how the tokens are distributed. Another way of preventing big oscillations is by introducing direct communication between the suppliers. This can lead to bilateral and multilateral negotiations, which permit for rational reasoning. In a similar way, a solution with an ordering of suppliers can be introduced, and the agents higher in hierarchy would be privileged to change their supply. In both last approaches, preference of a supplier can be decided by all the suppliers, using the information they share, without a central decision body.

Realistically, a rotation of the suppliers should be introduced, based on various factors, such as e.g. the cost of supply. To dynamically adjust the ordering, a market scheme can be adopted: prices will introduce a certain order. Exchanging information about the price and defining a cost of imbalance (the bigger the difference between supply and demand the higher the cost) is a simple market based scheme for balancing. Considering such approach requires concessions from participants, but also allows for rational decision making and leads to almost perfect balancing.

The goal function in real life power balancing is much more complex than in the artificial, theoretical problems. Comparing the goal functions (equations (1) and (3)) can give impression that they are the same. But in many of the described models the criterion is to maximize the profit or minimize the cost of producing energy ([9], [11]), where achieving balance is just one of the conditions. Often only these microgrids are considered that have connection to the external power network. Such a reserve (external network is not a constraining factor in this case: it can supply or receive any amount of power) is ensuring that the balance can always be achieved, which facilitates the decision making. In such conditions, the problem of balance is not the primary one and the goal function focus usually on profitability of the power production. The concept of a microgrid is fairly new. Pointing out that it can generate a revenue can motivate further development of this technology and construction of microgrids. When the island mode operation of the microgrid is considered, the power balancing becomes crucial for safety of the devices and the network itself.

In the Potluck and the El Farol problems decisions must be taken in discrete time intervals and need many iterations to allow the learning algorithms to adjust the predictors. After each iteration the outcome is calculated – the amount of people in the bar or amount of food on the party. The power balancing problem is in reality a continuous process, but it is often quantified to allow computer algorithms to cope with. The shorter the quantification time, the more small changes can be balanced, leading to smaller loses and better security of the grid. However, shortening of the balancing time has also its limits; as the change of the operation point of the devices requires time. Different devices have varied times of reactivity regarding their operation point change, which makes it impossible to derive the optimal minimal length of a time period in general. It can be approximated when the real set of devices that are installed in the defined microgrid is known. At present, for energy management system, the time periods may be 10 minutes, 5 minutes, but seldom less than 1 minute. Of course, the minimal physical time depends on the set of devices, but that can be evaluated only experimentally.

V. Methods of power balancing

The problem of power balancing is slightly different on each level of the power grid. Balancing power in the high voltage network can benefit from big aggregation of consumption. There the daily and weekly cycles dominate [12] and the inertia of the grid is much larger. In microgrids, the changes in consumptions still have visible cycles, but the random behavior plays a bigger role and the inertia of devices is smaller. This requires fast decision making regarding changing the operation point of sources and consumers in the grid. That poses a computation challenge, especially when the number of nodes is large and an energy management system has to balance the energy in all nodes, considering also all the physical limitation of the devices within a defined time period.

Effective balancing requires some kind of communication or a schema of cooperation between the producers of energy. The most straightforward schema is the centralized management: it is then possible to have one predictor of demand (e.g. that which gives the smallest errors), based on which the plan for production is made and the system distributes the power production. Centralized systems offer possibility of optimal production distribution [11], possibly considering multicriteria decision making. Centralized systems unfortunately have a number of different disadvantages: sensitivity to central controller failure, poor scalability, and requirement of full control over the sources. Full control may not be a problem in microgrids with a single owner, but may be unacceptable in a general situation. A centralized system might also not be able to consider specific preferences of the source owners or might give unacceptable results when a source owner happens to actively make decisions on its own (although that should not happen in a well designed system).

Non-centralized solutions have been also developed and showed promising results. Agent-based power balancing systems are quite popular approach. Due to the intrinsic characteristics of the agents, these system are distributed. A classification of different energy management schemes for agent-based systems can be found in [13]. Agents can represent single devices, nodes in the power grid, subsets of nodes or even single microgrids. The presented categories of management schemes are central-hierarchical control structure, distributed-hierarchical control structure, and decentralized control structure (peer-to-peer relation). The hierarchical organization of agents introduces an order and defines agent's functions in optimization and decision making. This can speed up the processing of the data, by dividing and distributing the tasks for calculation. The hierarchy can handle power distribution in a similar way as centralized systems. Completely decentralized control structures are extremely robust to failures and can quickly adapt to changing conditions, but because there is a larger exchange of data and negotiation, such systems tend

to operate slower, which might be the source of additional imbalances.

The last group of control systems are the ones based on market structures. The market is the central element of the balancing process, but the participants decide what kind of offer is placed on the market. In such approaches, money and cost functions play the role of ordering the power from most desired sources (i.e. cheapest and most efficient) down to the sources that are used only in emergency (i.e. more expensive power systems). Presentation of market based energy control systems can be found in [6], [9], [14].

VI. Conclusion

The power balancing problem in microgrids is especially important when the island mode operation is considered.

There are many methods that approach this problem with success, by using well known concepts such as markets. However, question arises whether other and perhaps simpler schemes might also be successful and more performent. This problem is adressed in the paper.

Practical power balancing has a lot of different limitations, which are very difficult to model: profitability, physical limitations, long term contracts for power, interaction between sources, delays in changing operation point, etc. Despite of this, theoretical models of the grid can give many useful information about the behavior of the real world system in an extremely controlled environment. The Potluck problem is very simplistic, but it touches the core of balancing: distributed decision making, in a situation of insufficient knowledge about demand and supply. Unfortunately, the problem is ill-defined and there are simply not sufficient data to create a rational-learning algorithm. The inductive learning and non-rational approaches give results that are oscillating around the ideal solution which is acceptable for theoretical problems but is unacceptable in power balancing.

In the Potluck problem, it is clear that the uncertainty is on both supply and demand side. The power balancing problem is different, as the demand can be predicted and a quick adjustment of the supply can be done in order to compensate the errors of the prediction. The inertia of the devices defines a time in which this adjustment has to be made: the faster the managing algorithm is, the smaller are the imbalances that occur, and the better it is for the grid devices.

The Potluck problem further shows that, when demand is constant and known to the suppliers, autonomous decisions made by the suppliers can actually result in a big oscillation of supply. Having the same set of information (even true and exact) cannot help if all suppliers are taking rational decisions, but do not consider decisions of other suppliers.

This also applies to the power balancing. Certain willingness to cooperate and to make concessions is needed to assure that the balancing works, with minimal oscillations. Reactiveness to the current situation and to other participants turns out to be more important than historic data about power usage and production. As shown in the experiments, knowledge about behavior of other producers is more important than historical knowledge of consumers. More accurate predictions help to improve the situation, but as human behavior is quite unpredictable, not much improvement can be made. The Potluck problem can be solved with a rational solution, if certain constraints are relaxed. Relaxing the constraint pertaining information sharing between suppliers, which we consider to be in a form of communication, is sufficient. In this situation, the Potluck problem closer resembles the power balancing one, where communication or sharing of information between producers is also sufficient to achieve a balancing within the time in which the devices need to react.

Moving from traditional grids to microgrids in the island mode is a huge modification and poses new technological challenges, not only in the physical structure, but also in the management and underlying schemes. Communication between controllable producers is anyway becoming a standard in modern power grids. This evolution gives enough information to develop a management scheme for a microgrid in the island mode, where proper balancing can be achieved within the limits of available production and consumption.

Acknowledgment

The research of W. Radziszewska was supported by the Foundation for Polish Science under International PhD Projects in Intelligent Computing. Project financed from The European Union within the Innovative Economy Operational Programme 2007-2013 and European Regional Development Fund.

References

[1] R. Palma-Behnke, C. Benavides, E. Aranda, J. Llanos, and D. Saez, "Energy management system for a renewable based microgrid with a demand side management mechanism," in *Computational Intelligence Applications In Smart Grid (CIASG), 2011 IEEE Symposium on*. IEEE, 2011, pp. 1–8.

[2] V. Balijepalli, V. Pradhan, S. Khaparde, and R. M. Shereef, "Review of demand response under smart grid paradigm," in *Innovative Smart Grid Technologies - India (ISGT India), 2011 IEEE PES*, Dec 2011, pp. 236–243.

[3] P. K. Enumula and S. Rao, "The potluck problem," *CoRR*, vol. abs/0809.2136, 2008.

[4] W. B. Arthur, "Complexity in economic theory: Inductive reasoning and bounded rationality," *The American Economic Review*, vol. 84, no. 2, pp. 406–411, May 1994.

[5] P. Vytelingum, T. D. Voice, S. D. Ramchurn, A. Rogers, and N. R. Jennings, "Theoretical and practical foundations of large-scale agent-based micro-storage in the smart grid," *J. Artif. Int. Res.*, vol. 42, no. 1, pp. 765–813, Sep. 2011.

[6] H. Vogt, H. Weiss, P. Spiess, and A. Karduck, "Market-based prosumer participation in the smart grid," in *4th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*. IEEE, 2010, pp. 592–597.

[7] W. Radziszewska, Z. Nahorski, M. Parol, and P. Pałka, "Intelligent computations in an agent-based prosumer-type electric microgrid control system," in *Issues and Challenges of Intelligent Systems and Computational Intelligence*, ser. Studies in Computational Intelligence, L. T. Kóczy, C. R. Pozna, and J. Kacprzyk, Eds. Springer International Publishing, 2014, vol. 530, pp. 293–312. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-03206-1_20

[8] S. Ramchurn, P. Vytelingum, A. Rogers, and N. Jennings, "Putting the 'smarts' into the smart grid: a grand challenge for artifitial intelligence," *Communications of ACM*, vol. 55, no. 4, pp. 86–97, 2012.

[9] P. Pałka, W. Radziszewska, and Z. Nahorski, "Balancing electric power in a microgrid via programmable agents auctions," *Control and Cybernetics*, vol. 4, no. 41, pp. 777–797, 2012.

[10] J. R. Leady, "If nobody is going there anymore because it's too crowded, then who is going? experimental evidence of learning and imitation in the el farol coordination game," august 2007. [Online]. Available: http://ssrn.com/abstract=1275537orhttp://dx.doi.org/10.2139/ssrn.1275537

[11] A. Tsikalakis and N. Hatziargyriou, "Centralized control for optimizing microgrids operation," *Energy Conversion, IEEE Transactions on*, vol. 23, no. 1, pp. 241–248, March 2008.

[12] A. Lovins, M. Odum, J. Rowe, and J. Rowe, *Reinventing Fire: Bold Business Solutions for the New Energy Era*. Chelsea Green Publishing Company, 2011. [Online]. Available: http://books.google.com.au/books?id=ZQVZxsGFjnAC

[13] G. Rohbogner, S. Fey, U. Hahnel, P. Benoit, and B. Wille-Haussmann, "What the term agent stands for in the smart grid definition of agents and multi-agent systems from an engineer's perspective," in *Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on*, Sept 2012, pp. 1301–1305.

[14] M. Vasirani and S. Ossowski, "A collaborative model for participatory load management in the smart grid," in *Proc. 1st Intnl. Conf. on Agreement Technologies*. CEUR, 2012, pp. 57–70.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1525–1527
DOI: 10.15439/2014F354
ACSIS, Vol. 2

Don't step on the Distribution's Tail!

Investigating the impact of random fluctuations on efficient resource utilization

Fabrice Saffre
BT Research and Innovation
Ipswich, United Kingdom
fabrice.saffre@bt.com

Hanno Hildmann
NEC Laboratories Europe
Heidelberg, Germany
hanno.hildmann@neclab.eu

***Abstract*—It is sometimes assumed that the total amount of a resource being consumed is the key consideration when attempting to devise the most efficient management strategy. We explore a case in which the rate and timing of resource utilization are also susceptible to impact on performance and illustrate how bringing random fluctuations under control can help maximize efficiency in a simple client-server scenario. The role of self-organization and distributed control methods in achieving this goal is briefly discussed.**

***Index Terms*—random fluctuations; resource management; performance optimization; client-server architecture**

I. Introduction

MUCH has been said about demand side management [1], the so called smart grid [2] and about intelligent infrastructure [3]. Intuitive results are often generalised and the important little exceptions are sometimes overlooked. This short paper aims to invite the reader to follow the authors on a interesting, empirical result based investigation.

We simulate a server farm under a capping constraint, as it is commonly discussed in the literature when e.g. considering the use of renewable energies to power some infrastructure. By comparing two scenarios and assuming a cost for the deviations from the norm for both, we can compare the impact either strategy has on the cost efficiency of the operation. This led us to a result which we consider worth sharing.

II. Background

Efficient resource management is fast becoming one of the most critical features of almost any human activity. Whether this is the case seems clear to us, why this is the case we would rather leave to a philosophical debate. On a very abstract level, this is arguably because we have been so successful as a species that we have reached a limit above which the planet is no longer capable to sustain our wasteful habits. There was a time when we could afford to be oblivious of how much water, food, energy etc. we were using, simply because our impact was negligible, the environment was able to replenish the stocks of whatever resource we needed as quickly as it was consumed. But is no longer the case, and has not been the case for some time.

A. The problem

For better or worse, we have entered an age in which we will have to act more responsibly, to use only what is needed when it is needed, under penalty of seeing our global society descend into conflicts and service disruption as we are forced to compete ever more aggressively for dwindling supplies. The *we* here can be understood as the societal *all of us*, but also as the managerial *our company* or even the personal and cost aware *I* in every day live.

The most visible aspect of this requirement is that we must seek to limit the total amount of a resource that is consumed in the process of achieving a certain goal. There are countless examples of this today, from designing more fuel-efficient vehicles to engineering crops that require less water or nutrients. There is however another, perhaps less obvious angle to this quest for efficiency: notwithstanding how much of a resource is consumed in total, when and at which rate it is being used can also have a critical impact on its availability (and cost). While this seems an obvious fact, it is often overlooked, or lost in abstraction.

B. Example

A good example of such a resource is renewable power. Imagine that a solar plant producing 1 MW for 10 hours a day is used to power a community that consumes a total of 10 MWh over 24 hours. If some power is used at night or if the load ever exceeds the available output from the solar plant during the day, even if the aggregate supply as well as consumption total for the day is still exactly identical (i.e. 10 MWh), some provision for this "mismatch" between supply and demand will need to be made, in the form of an additional power source or storage facility. On the contrary, if the demand could be exactly mapped to the output of the solar plant over time, then it could be met locally and efficiently (with fewer losses). This is of course widely discussed in the recent literature, and techniques to shape the load so as to better approximate the available supply are collectively known as "Demand-Side Management" or DSM [4].

There is a slightly different but related and very common special case of this problem: what if the average demand for a resource is constant and known but instantaneous consumption fluctuates randomly and unpredictably? If a facility (or supply) is dimensioned so as to accommodate the average requirement, then it will necessarily be sometimes under-, sometimes over-used. Moreover, depending on the characteristics (amplitude, symmetry, . . .) of the fluctuations, the time spent in either state (over-utilization and under-utilization), as well as the deviation from the average, may vary greatly.

978-83-60810-58-3/$25.00

C. The issue with statistics

In practice, it is often assumed that, because fluctuations statistically cancel each other out over time, this is no cause for concern and so not a relevant field of study. But this view is contradicted by the realization that the rate and timing of resource utilization can impact on efficiency (as illustrated by the above example). Similarly, the commonly held view that in a large enough population of consumers, deviations from the global average will be negligible is also a simplistic one. Indeed, there is no guarantee that the costs incurred by such deviations grow linearly. If they don't, then even small deviations can have a significant impact and large ones, although extremely rare, could have a disastrous effect.

D. Aim of this paper

In this paper, we experiment with a set of conceptual tools designed to quantitatively measure the influence of random fluctuations on performance in a simple client-server scenario. Specifically, we compare the case in which no upper bound is imposed on the number of active servers to that in which there is such a constraint. In effect, we propose a cost/benefit analysis of two strategies: cutting the upper-end tail of the distribution (which creates execution delays) versus "stepping on it" (which may incur extra costs).

III. Model and simulations

A. The model

We used Monte Carlo simulations to approximate the dynamics of a group of servers. On every time-step, all identical servers have a fixed probability `P` to receive a new job, the duration of which is comprised between 1 and (2 $\times$ `avg` 1) time-steps (where `avg` is the average duration of a job). `P` is chosen in such a way that, statistically, slightly less than 1 out of 8 servers is expected to be busy at any time.

On every time-step, the number of servers in the "busy" state is recorded. In the default scenario (no constraints), this value is allowed to exceed the 1:8 ratio. In the other scenario, no more than 1 out of 8 servers are allowed to be active simultaneously. If random fluctuations in job arrivals / duration would cause the system to exceed this limit, the execution of a corresponding number of jobs (randomly selected) is temporarily suspended (i.e. the servers processing them are put on stand-by).

In the default scenario, performance degradation is measured by the number of server-time-steps falling above the 1:8 target and incurring extra costs. In the constrained scenario, it is the cumulative delay (total number of time-steps) suffered as a result of imposing a cap on the number of active servers (QoS (quality of service) penalty).

B. Simulations

Data-centers were modelled to operate between 1000 and 4000 servers (by increments of 500) and were simulated for 512 time-steps. The average job duration was 8 time-steps (i.e. flat distribution between 1 and 15). There were 1000 realizations for every size and for each scenario, totaling 14000 independent simulation runs.

A control simulation (simulating batches up to 1024 time steps was run to confirm the results presented in the next section. The results from this investigation confirmed that the system was at, or very close to, its steady state at t=512.

IV. Results

A. Frequency distribution

Fig. 1 shows the frequency distribution of simulation outcome for 4 different server population sizes. As expected, for the "constrained" scenario, there is a sharp peak corresponding to the 1:8 ratio (125, 250, 375 and 500 servers respectively) since the whole tail of the distribution has "collapsed" onto this single value. Note that the remainder of the distribution closely follows the Gaussian profile found for the default scenario, apart from in the case of smaller populations, where capping also seems to negatively affect the height of the normal peak.

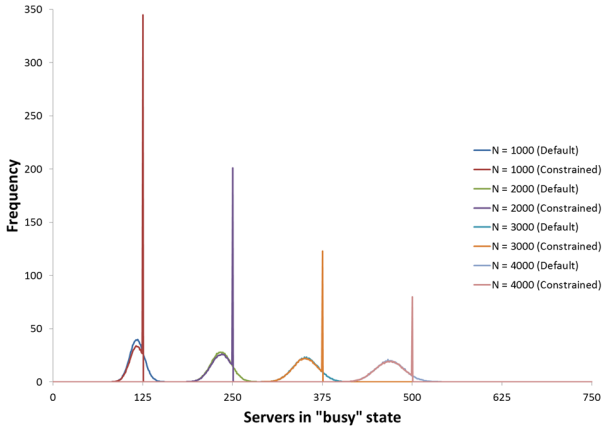


Fig. 1. Frequency distribution of system state as a function of the number of active servers, for 4 different population sizes and for the two possible scenarios ("Default" vs. "Constrained"). 1000 runs per set of parameter values.

The values reported in Fig. 1 consider only the values for the last 128 (out of 512) time-steps, when the system is at or very close to steady state. The motivation is that the initial values would be too much affected by the build up, but after $3/4^{th}$ of the steps we can assume that the overspill from previous steps has normalised (especially since the individual duration can not exceed 15 and we are not including the first 384 steps).

B. Aggregated QoS violation

Fig. 2 shows the evolution of the average delay (QoS penalty) incurred over all processed jobs, for increasing system size and for both scenarios. As expected, this value is unaffected by system size in the default case, where a delay only occurs when a job is submitted to a server before it has completed the execution of its predecessor. By contrast, in the constrained scenario, a QoS penalty is also incurred when the overall workload exceeds the processing capacity of $1/8^{th}$ of the population.

The main finding is that the average delay converges quickly for both cases as system size increases. This is due to the amount of time spent above the cut-off limit in the default scenario being inversely proportional to the number of servers involved; thus capping becomes proportionally less frequent in the constrained scenario (illustrated by the decreasing height of the sharp peak corresponding to 1:8 limit, see Fig. 1).

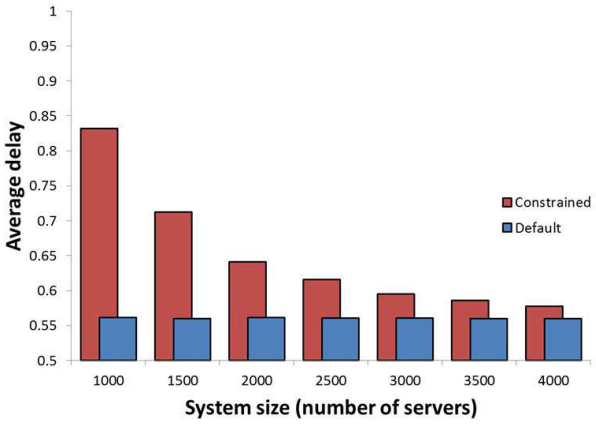


Fig. 2. Evolution of the average delay per job as a function of population size (workload proportional to the number of servers). The larger the system, the lower the QoS penalty resulting from enforcing the 1:8 cap.

C. Comparative QoS violations

Although to some extent these results are merely intuitive and can be anticipated from known statistical properties [5], they can also be interpreted in a different and more surprising way. For instance, the fact that, in the default scenario, the absolute number of server-time-steps falling inside the high-end tail of the frequency distribution shown on Fig. 1 (i.e. above the 1:8 limit) exhibits a maximum (see Fig. 3) clearly suggests that the capping strategy would yield most benefits within a finite range of system sizes.

D. Potential benefit analysis

Table I illustrates how the potential benefits of using the capping strategy would vary as a function of system size, under the arbitrary assumption that every server-time-step over the 1:8 target incurs an extra £0.01 cost (e.g. because of higher server rental cost) and every time-step delay incurs a £0.01 penalty fee (e.g. as compensation for breaching the service-level agreement).

TABLE I
POTENTIAL SAVINGS FROM APPLYING A CAPPING STRATEGY (ARBITRARY PENALTY COSTS). NOTE THE PRESENCE OF A MAXIMUM FOR N = 3000.

N	Cost Breakdown		
	Cost Differential (default-constraint)	*QoS Penalty (constraint-default)*	*Net Savings*
1000	£33,792	£21,251	£12,541
2000	£41,238	£12,687	£28,552
3000	£43,052	£8,160	£34,892
4000	£39,187	£5,562	£33,625

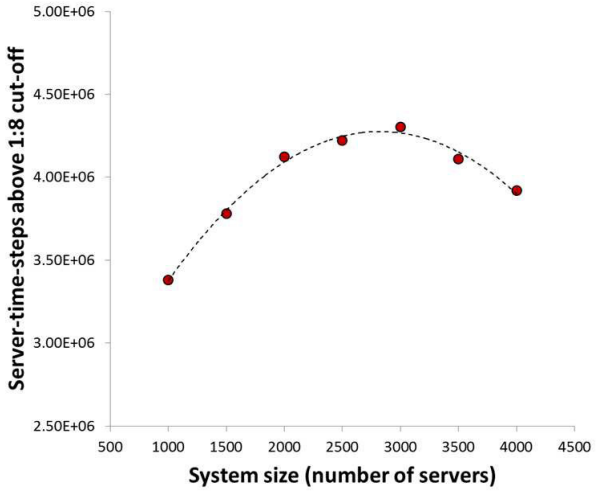


Fig. 3. Total (absolute) number of server-time-steps falling above the 1:8 cut-off limit in the default scenario, as a function of population size.

V. CONCLUSIONS

These preliminary findings confirm that there are realistic applications for which limiting the total amount of a resource being consumed over a period of time is not enough. Preventing the instantaneous load from exceeding a predetermined value can be beneficial and contribute to improving efficiency.

Moreover, relying on statistical effects, either over time or over a large population, to ensure a smooth demand profile may be a high-risk strategy as it is possible that even small or rare deviations from the average incur a severe penalty.

In this context, it seems important to investigate control methods that would permit to regulate the aggregated demand from a population of resource consumers so as to prevent them from exceeding a certain target. Because a central mechanism for achieving such a goal may not be feasible (due to scalability problems, limits on information availability or even ownership boundaries), we argue that a form of collective intelligence capable of supporting self-management is required. The investigation of a candidate technology for implementing such a distributed resource controller will be the subject of future work.

REFERENCES

[1] A. A. Garcia, "Demand side management integration issues a case history," *Power Systems, IEEE Transactions on*, vol. 2, no. 3, pp. 772 –778, Aug. 1987.

[2] T. T. Kim and H. V. Poor, "Scheduling power consumption with price uncertainty," *Smart Grid, IEEE Transactions on*, vol. PP, no. 99, p. 1, 2011.

[3] F. Saffre, H. Hildmann, and S. Nicolas., "The adaptive grid." The Institute of Telecommunications Professionals (ITP) - ITP Journal, vol. 6, no. 3, pp. 31–38, 2012.

[4] G. Strbac, "Demand-side management: benefits and challenges," *Energy Policy*, vol. 36, pp. 4419–4426, 2008.

[5] D. Bini, G. Latouche, and B. Meini, *Numerical Methods for Structured Markov Chains*, ser. Numerical Mathematics and Scientific Computation, 2005.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1529–1538
DOI: 10.15439/2014F49
ACSIS, Vol. 2

Synthesised Constraint Models for Distributed Energy Management

Alexander Schiendorfer, Jan-Philipp Steghöfer, Wolfgang Reif
Institute for Software & Systems Engineering, Augsburg University, Germany
Email: {alexander.schiendorfer, steghoefer, reif}@informatik.uni-augsburg.de

***Abstract*—Resource allocation is a task frequently encountered in energy management systems such as the coordination of power generators in a virtual power plant (unit commitment). Standard solutions require fixed parametrised optimisation models that the participants have to stick to without leaving room for tailored behaviour or individual preferences. We present a modelling methodology that allows organisations to specify optimisation goals independently of concrete participants and participants to craft more detailed models and state individual preferences. While considerable efforts have been spent on devising efficient control algorithms and detailed physical models in power management systems, practical aspects of unifying several heterogeneous models for optimisation have been widely ignored – a gap we aim to close. As a by-product, we give a formulation of warm and cold start-up times for power plants that improves existing power plant models. The concepts are detailed with the load-distribution problem faced in virtual power plants and evaluated on several random instances where we observe that a significant number of soft constraints of individual actors can be satisfied if considered.**

I. Constraint Optimisation Problems in Power Systems

RESOURCE allocation and scheduling are difficult problems that occur frequently in energy systems, be it the coordination of power generation [1], demand-side management, or building control software. In a producer-based view, supply needs to meet the demand as accurately as possible in order to guarantee stability and avoid costs incurred by corrective measures. Similarly, consumers may try to find cost-minimising schedules for processes required throughout a day with respect to time-dependent energy prices. Current initiatives[1] are based on the assumption that *groups* of prosumers (i.e., energy producers and/or consumers) can form and team up to achieve better prices or production rates for their participants. We also adopt the notion of *agents*, indicating that the prosumers are in principle autonomous entities, even if they surrender the decision about their power output to the group.

A straightforward solution (see, e.g., [2], [3], [4], [5]) to this resource allocation problem is to model the decision making process (e.g., distributing the load in a virtual power plant (VPP) or scheduling energy-consuming domestic processes in a consumer coalition) as a mathematical optimisation problem such as a mixed integer program (MIP), a linear program (LP) or as a constraint satisfaction and optimisation problem (CSOP) as done by industrial distributed energy management tools such as Siemens DEMS [6] or PLEXOS Integrated Energy Model [7]. DEMS is used, e.g., by the municipal utility of the city of Munich for controlling a VPP [8]. In essence, the problem is specified in terms of (decision) variables, their associated domains, and constraints that regulate which assignments are valid. The task accomplished by the respective solvers is then to assign values to all variables such that no constraint is violated and an optimisation objective is minimised (or maximised).

Typically, such tools (DEMS in particular) offer a predefined range of agent types such as energy generators, storages, or controllable loads. Users may then specify the topology of their energy system to calculate optimized power schedules. A concrete power generator is thus essentially represented by *one tuple* in a data repository containing the parameters defining its behaviour. Consequently, the provided models constitute a static *one-for-all* solution that needs to encompass all supported characteristics of power generators, including, e.g., time-dependent properties such as inertia.

Clearly, power generators show varying characteristics such as change rates, cool or warm start-up times or power boundaries depending on, e.g., the power plant type or manufacturer. *Parametrised models* as described above cannot support this variety. At some point the model has to be fixed for all participants and individual variables necessary to model a certain constraint cannot be added. To overcome this limitation, we suggest to synthesise an optimisation problem from several *individual models*. Such *synthesised models* allow for individual preferences (typically in the form of knowledge acquired by power plant operators such as economically optimal production ranges or limited ramp-up or -down of a generator) and separate modelling of the organisational optimisation problem and physical models of individual participants – properties that are attractive for organisations as more clients can be served as well as for individual participants as they can influence the assigned plans. This methodology is not only nice to have in multi-agent systems, where optimisation problems result from a combination of several sub-problems – it is *necessary*.

Our contribution leads to a methodology that offers:

1) support for heterogeneous prosumers requiring specific sets of variables;
2) isolated modelling of physical components;
3) clean separation of the organisational aspects such as

[1] cf. https://www.energiekosten-stop.at/ for consumer alliances or http://www.swm.de/geschaeftskunden/effizienz-umwelt/virtuelles-kraftwerk.html for virtual power plants

978-83-60810-58-3/$25.00

objectives or fairness constraints from physical models;
4) incorporation of individual preferences into the optimisation routine of a coalition to increase and incentivise the participation.

We exemplify model synthesis with the problem of creating schedules in a virtual power plant and show how to integrate custom behaviour in the form of cold and warm start-up times that are specific to certain power plant types as well as individual economical preferences. While we demonstrate the modelling and synthesis approach for a single organisation, these concepts can be incorporated to solve hierarchical resource allocation problems as described in [9], which focused on the abstraction of constraint models.

The paper is structured as follows: Sect. II introduces a formalism to express preferences with the help of constraint relationships while Sect. III shows the general approach to power plant scheduling within a virtual power plant. Both approaches are used in Sect. IV to synthesise individual models within a group of power plants. Sect. V then instantiates the general framework for use with IBM ILOG CPLEX. Our experimental results and the findings drawn from them are discussed in Sect. VI. We conclude the paper with a discussion of future research directions.

II. Constraint Programming with Constraint Relationships

As we suggest a modelling methodology for constraint satisfaction and optimization problems (CSOPs, see, e.g., [10]) to solve resource allocation problems, we briefly revisit the core model elements as well as the definition of constraint relationships [11] that we use to denote individual preferences of single agents. A CSOP $\zeta = \langle \mathcal{X}, \mathcal{D}, \mathcal{C}, f \rangle$ consists of a set of decision variables $\mathcal{X}$ that take values from the domain $\mathcal{D}$ where consistent assignments $\theta \in (\mathcal{X} \to \mathcal{D})$ are regulated by the set of constraints $\mathcal{C}$. We write $\theta \models c$ if the constraint c is satisfied by an assignment θ. The objective function $f : (\mathcal{X} \to \mathcal{D}) \to \mathbb{R}$ measures the quality of a solution and effectively imposes an ordering over the assignments where we seek the best one.

However, not all constraints need to be hard requirements and some may also be violated if no assignment simultaneously satisfies all constraints [12]. We call these constraints *soft* and denote them by $\mathcal{C}_s$ as opposed to hard constraints $\mathcal{C}_h$ with $\mathcal{C}_h \cup \mathcal{C}_s = \mathcal{C}$, $\mathcal{C}_h \cap \mathcal{C}_s = \emptyset$. A set of *constraint relationships* for the soft constraints C_s of a CSOP ζ is given by a binary asymmetric relation $\mathcal{R} \subseteq \mathcal{C}_s \times \mathcal{C}_s$ whose transitive closure $\mathcal{R}^+$ is a partial order relation. We write $c' \prec_{\mathcal{R}} c$ or $c \succ_{\mathcal{R}} c'$ iff $(c, c') \in \mathcal{R}$ to define c to be *more important* than c', analogously for $\mathcal{R}^+$. If $c' \prec_{\mathcal{R}} c$ we call c' a *direct predecessor*, if $c' \prec_{\mathcal{R}^+} c$ a *transitive predecessor* of c. Moreover, we refer to the *constraint relationship graph* as the directed graph spanned by $\langle \mathcal{C}_s, \mathcal{R} \rangle$. Figure 1 shows a toy example of a CSOP with constraint relationships.

The binary relation over soft constraints needs to be lifted to sets of soft constraints that are violated by an assignment. Such a *violation set* is denoted by capitalizing the letter used for the assignment; i.e., for some assignment $t \in (\mathcal{X} \to \mathcal{D})$ its violation set is $T = \{c \in C_s \mid t \not\models c\}$. We propose different dominance properties $p \in \{\text{SPD}, \text{TPD}\}$ where SPD (single-predecessor-dominance) indicates that one constraint may only dominate a single predecessor and TPD (transitive-predecessor-dominance) defines that a single constraint is more important than *all* of its predecessors (corresponding to the relative importance among constraints) – we refer to [11] for more details. We write $T \longrightarrow^p_R U$ to denote that the violation set T *worsens* to U with dominance level p and use the following rules:

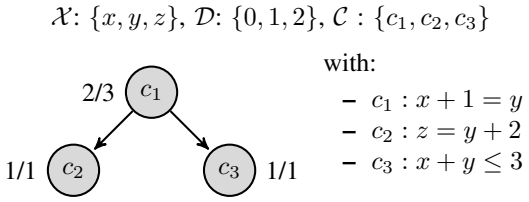


Fig. 1. Not all three constraints can be satisfied simultaneously, e.g. c_2 forces z to be 2 and y to be 0, conflicting with c_1, We can choose between solutions satisfying $\{c_1, c_3\}$ or $\{c_2, c_3\}$. Weights are given in the form "SPD / TPD".

$$T \longrightarrow^p_R T \uplus \{c\} \tag{W1}$$

$$\frac{T_1 \longrightarrow^p_R U_1 \quad T_2 \longrightarrow^p_R U_2}{T_1 \uplus T_2 \longrightarrow^p_R U_1 \uplus U_2} \tag{W2}$$

$$T \uplus \{c\} \longrightarrow^{\text{SPD}}_R T \uplus \{c'\} \quad \text{if } c \prec_R c' \tag{SPD}$$

$$T \uplus \{c_1, \ldots, c_k\} \longrightarrow^{\text{TPD}}_R T \uplus \{c'\} \quad \text{if } \forall c \in \{c_1, \ldots, c_k\} : c \prec_{R^+} c' \tag{TPD}$$

Finally, we define a partial order over solutions, denoted by $t >^p_R u$ and to be read as "t is better than u", using $T \ (\longrightarrow^p_R)^+ \ U$ (meaning repeated sequential application of the rules) for the selected $p \in \{\text{SPD}, \text{TPD}\}$. To use them in a CSOP, we calculate weights for the constraints that respect the partial order $>^p_R$ where we summarise the weights of violated constraints as the *penalty* of a solution:

$$w^{\text{SPD}}_R(c) = 1 + \max\{w^{\text{SPD}}_R(c') \mid c' \in \mathcal{C} : c \succ_R c'\}$$
$$w^{\text{TPD}}_R(c) = 1 + \sum_{c' \in \mathcal{C}_s : c \succ_R c'} (2 \cdot w^{\text{TPD}}_R(c') - 1)$$

Consequently, we define an objective based on $p : (\mathcal{X} \to \mathcal{D}) \to \mathbb{N}$ as:

$$\underset{\theta}{\text{minimize}} \ \ p(\theta) = \sum_{c \in \mathcal{C}_s, \theta \not\models c} w(c) \tag{1}$$

III. Scheduling Power Plants within a Virtual Power Plant

Our approach to synthesise individual models is exemplified with the problem of finding schedules for power plants in a virtual power plant that we described in [9] and [13]. This problem is also known as economic load dispatch (ELD) [14] or unit commitment (UC) [15]. In essence, the task is to

```
int lastSimStep = 10;
range TIMERANGE = 0..lastSimStep;
{string} PowerPlants = ...;

tuple PowerPlantData {
  float minimal; float maximal; float fixedRamp;
};

PowerPlantData plants[PowerPlants] = ...;
float demand[TIMERANGE] = ...;

dvar float+ production[PowerPlants][TIMERANGE];
dexpr float totalProduction[t in TIMERANGE] =
            (sum( p in PowerPlants ) (production[p][t]));

minimize
  sum ( t in TIMERANGE )
     abs(demand[t] - totalProduction[t]);

subject to {
  forall (p in PowerPlants, t in 0 .. (lastSimStep - 1)) {
    production[p][t] >= plants[p].minimal;
    production[p][t] <= plants[p].maximal;
    abs(production[p][t] - production[p][t+1])
        <= plants[p].fixedRamp;
  }
}
```

Listing 1. A minimalistic, *parametrised* model of a load distribution problem

distribute a given load (for a certain time window) to a set of power generators in such a way that their capacities as well as inertia between consecutive time steps are respected. We present this basic scheduling problem as a CSOP (which we will refine to accommodate additional model aspects):

$$\begin{aligned} \underset{P_t^a}{\text{minimize}} \quad & \sum_{t\in\mathcal{W}} |P_t - \mathcal{D}_t|\,,\ P_t = \sum_{a\in\mathcal{A}} P_t^a \\ \text{subject to} \quad & \forall a \in \mathcal{A}, \forall t \in \mathcal{W} : P^a_{\min} \le P_t^a \le P^a_{\max}, \\ & v^a_{\min}\left(P^a_{t-1}\right) \le P_t^a \le v^a_{\max}\left(P^a_{t-1}\right) \end{aligned} \tag{2}$$

where P_t^a are the decision variables representing the production of plant $a \in \mathcal{A}$ at time step $t \in \mathcal{W}$, the scheduling window. The demand is given as the vector $\mathcal{D}$ and basic properties about the constraints of each power plant include minimal and maximal production values, $P_{\min}$ and $P_{\max}$ and functions denoting minimal and maximal production given the current output $v^a_{\min}, v^a_{\max} : \mathbb{R} \to \mathbb{R}$ to incorporate inertia and model the ramping behaviour of power plants [16]. To make this example more concrete, we assume that a power plant is described by its production boundaries as well as a fixed ramp up rate between two consecutive time steps (taking the role of $v_{\min}$ and $v_{\max}$) depending on the nameplate capacity as, e.g., given in [17]. We can then formulate the optimisation problem from Eq. 2 in IBM's optimisation programming language (OPL) as in Listing 1. OPL is used by IBM ILOG CPLEX [18] which in turn can be employed by both DEMS and PLEXOS.

The listing shows the shortcomings of a *parametrised* model: All power plants are described by the *same* set of parameters and constraints are defined uniformly for all power plants and time steps. Hence, the possible variety is severely limited.

TABLE I
DIFFERENT COLD AND HOT START-UP TIMES FOR POWER PLANT TYPES. A COLD START OCCURS IF A PLANT IS DOWN FOR MORE THAN 48H, A HOT START IF IT IS DOWN FOR LESS THAN 8H. TAKEN FROM: [17], [21]

Plant type	Cold start-up (h)	Hot start-up (h)
Black coal	4 – 5	2
Brown coal	6 – 8	2 – 4
Gas turbine	0.5	0.25
Photothermal	4 – 5	2

In order to achieve feasible schedules, a number of different types in constraints are usually employed in power plant models. While the models presented within the scope of this work are far from complete and much more detailed models exist (see, e.g., [14], [15], [16]) these types are common and therefore representative. In addition, the more complex and heterogeneous models found in the literature only emphasise the need for a methodology to incorporate these diverse descriptions of physical and economic limitations with organisational aspects. The types of constraints considered here are:

Minimal up/down times: a generator is required to run (or be switched off) for a certain number of steps before being switched off (or turned on) [15].

Ramp up/down rates: usually, a fixed amount of production change between two consecutive time steps is assumed [19], especially in thermal power plants in which physical boundaries for heating and cooling the system have to be captured (as used in Listing 1). Alternatively, the possible change can be specified as a relative quantity denoting the percentage of the current output that a power plant can adapt.

Cold/warm start-up times: a power plant may need a certain number of time steps to ramp up from 0 to its minimal production level as modelled by [16]. However, for some plants, this start-up duration depends on the downtime as "cold starts" differ from "warm starts" (see Table I for sample values). We show how to formalise these start-up times in a MIP-framework using the transition system in Fig. 2. We consider the actual duration of a start-up as opposed to the costs which can be approximated with exponential functions [20].

Please bear in mind that *not all* power plants feature the *same* constraints. Consider, e.g., a power plant where the ramp up/down rates are high enough to regulate from minimal to maximal production in just one time step (say 15 minutes). Then, the model may not contain such a constraint [19]. If we want to consider this constraint in a parametrised model, *all* power plants would have to include model elements (e.g., ramp up rates). The problem becomes even more obvious when we consider minimal off times. As we demonstrate in Sect. IV-A, we need *additional* decision variables for the current off time

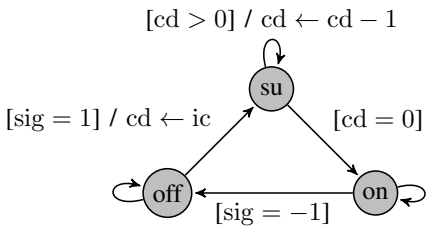


Fig. 2. Transition system to model adaptive start-up times depending on down time. The signal (sig) takes values from $\{-1, 0, 1\}$ where 1 shows a start-up, 0 is the default, and -1 triggers a shut-down. A countdown is initialised with ic as a result of a function of the down time and has the plant stay in the state *su* (start-up) for ic steps. A plant can only contribute in the state *on*. Expressions in brackets contain guards and other annotations at the transitions denote actions manipulating local variables.

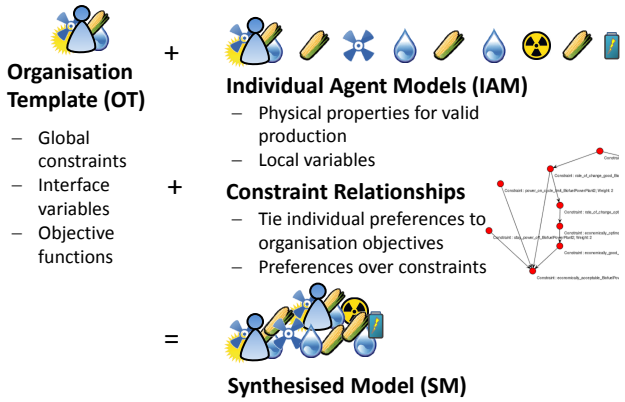


Fig. 3. Overview of the synthesis process

to adequately model these constraints. A parametrised model would again introduce these variables for *all* power plants including those that do not show minimal off times. For them, dummy values allowing to ignore certain constraints (e.g., a fixed ramp up rate of $P_{\max}$) would be required.

From an engineering perspective, it thus makes sense to model these aspects separately. Variants of these constraints can also be specified as preferences, if, e.g., a ramp up of 15 % is technically possible but it is more desirable to limit ramp up to 10 % to save expenses and material. We will use these exemplary constraints to create individual, heterogeneous power plant models to be used in synthesised models.

IV. Synthesis of Coalition Models

In light of the presented problems we can distinguish two aspects that are intermingled in Eq. 2:

Individual Agent Models (IAM) describe the properties of *one* agent representing a physical entity in terms of constraints for the available production. Constraints can be formulated depending on the internal state (being on/off, production levels etc.) independent of other agents. This model needs to be provided by the agent designer, e.g., the power plant manufacturer, possibly with customisations by its operator. An IAM defines the feasible production or consumption range of an agent but also regulates possible transitions between different time steps. Moreover, preferences (such as those avoiding high ramp-up rates) can be specified with constraint relationships to further constrain feasible schedules.

Organisational Templates (OT) represent the specific goals of an organisation or a coalition formed by the organisation. We consider organisations to be entities that exist independently of their specific agents and are therefore modelled separately [22]. The template captures the optimisation criterion and provides so-called *interface variables* that each IAM needs to incorporate. Additional (soft) constraints can impose policies of the organisation such as "prefer agents of type X" or "distribute resources in a fair manner".

Model synthesis is concerned with creating a CSOP from a set of agent models including their individual preferences and an organisational template as Fig. 3 shows. Aside from the generally required parameters of the individual agent models introduced in Listing 1 such as possible contributions, the power plant models exhibit varying characteristics regarding feasible schedules given by the constraints presented in Sect. III. Recall that to formalise these properties, we might need additional decision variables such as number of time steps a power plant has been switched off to capture the down time. As only *some* power plants may need these for modelling their start-up behaviour, we call them *local variables*. IAMs constitute a possible refinement of *option models* in the context of *Energy Agents*, a unifying framework for agent-based energy systems [23].

In our VPP example, the interface *decision* variables are the scheduled production values of all power plants over all time steps. Similarly, $P^a_{\min}$ and $P^a_{\max}$ need to be specified by each power plant a to calculate load percentages. Thus, the interface variables model the *homogeneous* parts of all considered agents as opposed to the *heterogeneity* introduced by custom behaviour in local variables: To accommodate minimal up/down times, we need variables $\top_t$ and $\bot_t$ that represent for how many time steps a power plant has been up or down, respectively, at time step t. For convenience, we further define $\nu_t \leftrightarrow P_t > 0$. A minimal up time of $\top^{\min}$ is expressed by the following constraint:

$$\forall t \in \mathcal{W} : \nu_t \wedge \neg\nu_{t+1} \rightarrow \top_t \geq \top^{\min} \tag{3}$$

For start-up times that depend on down times we need a countdown variable that regulates the time steps the power plant has to be in the start-up state shown in Fig. 2. We will further discuss how to formulate this constraint but first describe the general steps required for model synthesis.

A. Components of Model Synthesis

We first formally illustrate the ingredient models of our synthesis approach and give an implemented example using OPL in Sect. V. For doing so, we assume an organisation λ that controls a set of agents, $\mathcal{A}^\lambda$, along with their IAMs that consist of the interface variables $\mathcal{X}^\lambda$ and local variables

$\widehat{\mathcal{X}}^a$ for each agent a. In our example, modelling downtime limitations, $\mathcal{X}^\lambda = \{P_{\max}\} \cup \{P_t \mid t \in \mathcal{W}\}$ and $\widehat{\mathcal{X}}^a = \{P^*_{\min}, P^*_{\max}\} \cup \{\top_t, \bot_t, \nu_t \mid t \in \mathcal{W}\}$ would be a suitable choice. We need $P_{\max}$ to specify load factors later on and can technically distinguish $P_{\max}$ from P_t since the former are indeed constants and not decision variables as their value is fixed for a power generator. However, for ease of presentation we assume $P_{\max}$ to be just a decision variable with a domain consisting of only one value and treat all variables alike. The same holds for $P^*_{\min}$ and $P^*_{\max}$ that locally specify preferred ranges of operation. Then, the valid scope for all constraints within a single IAM of an agent a is $\mathcal{X}^\lambda \cup \widehat{\mathcal{X}}^a = \mathcal{X}^a$. The set of individual constraints $\mathcal{C}^a$ consists of both hard and soft constraints, $\mathcal{C}^a_h$ and $\mathcal{C}^a_s$ where $\mathcal{R}^a$ represents the constraint relationships defined over $\mathcal{C}^a_s$. We illustrate how to tie the auxiliary local variables in $\widehat{\mathcal{X}}^a$ to the interface variables with the vector denoting whether a plant is running:

$$\forall t \in \mathcal{W} : \nu_t \leftrightarrow P^0_t > 0$$

This expression helps in defining additional constraints and can be implemented by a decision expression in OPL directly. An individual preference for a certain production subrange $[P^*_{\min} P^*_{\max}]$ is specified by the soft constraint:

$$\forall t \in \mathcal{W} : P^*_{\min} \leq P^a_t \leq P^*_{\max}$$

To summarize, IAMs are specified as constraint satisfaction problems and represented as $\langle \mathcal{X}^a, \mathcal{D}^a, \mathcal{C}^a, \mathcal{R}^a \rangle$ with suitable choices of domains.

Next, consider the organisational template (OT). It provides at least the interface variables for *all* its subordinate agents – i.e., in contrast to a single IAM which specifies in terms of variables $x \in \mathcal{X}^\lambda$, the OT consists of a set of interface variables $\mathcal{X}^{\mathcal{A}^\lambda} = \{x^a \mid x \in \mathcal{X}^\lambda, a \in \mathcal{A}^\lambda\}$ for every agent a. These variables have to be provided in every IAM so the OT can use them for defining the organisational goal. Note that this would certainly not be true for local variables. It is not possible to specify, e.g., $\top^a_t$ indexed by a as it does not have to be defined in every model – contrary to a parametrised scheme. For unit commitment, the objective is taken from Eq. 2:

$$f = \underset{P^a_t}{\text{minimise}} \sum_{t \in \mathcal{W}} \left| P^\lambda_t - \mathcal{D}_t \right|,\ P^\lambda_t = \sum_{a \in \mathcal{A}} P^a_t$$

where $\mathcal{D}_t$ are constants representing the demand at time steps t and P^λ_t can either be seen as an additional decision variable in OT restricted by the constraint to be the sum of all agent productions or directly as a decision expression that is fully determined by the value of its associated decision variables (P^a_t). A soft fairness constraint for a VPP could be that no power plant should produce less than 40 % of its nominal capacity, $P_{\max}$, if possible. We call this organisational constraint oc1:

$$\text{oc1} : \forall a \in \mathcal{A}^\lambda, t \in \mathcal{W} : P^a_t \geq 0.4 \cdot P^a_{\max} \tag{4}$$

Similarly, we could impose a soft upper bound (oc2) for the load factor to prefer schedules that do not rely on few power generators providing all energy. All together, an organisation template is given by $\langle \mathcal{X}^{\mathcal{A}^\lambda}, \mathcal{D}^\lambda, \mathcal{C}^\lambda, \mathcal{R}^\lambda, f \rangle$ – the interface variables instantiated for each agent and their domains, a set of organisational hard and soft constraints with optional constraint relationships and an optimisation function f.

B. Steps of Model Synthesis

Synthesis takes a set of IAMs $\langle \mathcal{X}^a, \mathcal{C}^a \rangle$ for agents $a \in \mathcal{A}^\lambda$ and an organisation template to create the synthesised model (SM) for λ, a CSOP $\zeta = \langle \mathcal{X}, \mathcal{D}, \mathcal{C}, \mathcal{R}, f \rangle$, the construction of which we discuss step-by-step.

Step 1: First we make sure that local variables do not clash when combining several agents by renaming them to a unique identifier. A substitution operator $\{x_1 \mapsto x_2\}$ replaces all occurrences of the variable x_1 by the variable x_2 in a constraint or optimisation function. We then write x^a for a variable $x \in \mathcal{X}^a$. Note that this intentionally connects the interface variables $\mathcal{X}^\lambda$ used in an IAM with the counterpart in the OT, $\mathcal{X}^{\mathcal{A}^\lambda}$. The set of variables of ζ is then defined as $\mathcal{X} = \mathcal{X}^{\mathcal{A}^\lambda} \bigcup_{a \in \mathcal{A}^\lambda} \{x^a \mid x \in \widehat{\mathcal{X}}^a\}$ (furthermore, constraints keep their original domains to form $\mathcal{D}$).

The constraints $\mathcal{C}$ then consist of all hard and soft constraints of the organisational template and the constraints in all IAMs after substituting variables by their labelled counterparts. Hence:

$$\mathcal{C} = \mathcal{C}^\lambda \cup \bigcup_{a \in \mathcal{A}^\lambda} \{c\{x \mapsto x^a, \forall x \in \mathcal{X}^a\} | c \in \mathcal{C}^a\}$$

Constraint relationships for the synthesised model can be formed under the assumption that organisational constraints have a higher priority than individual preferences.

Step 2: As Fig. 4 shows, we then impose artificial edges that specify that all constraints in the OT are more important than all soft constraints of IAMs that do not have constraints that are deemed more important. However, this is a design decision as it might not be desirable for all problems to strictly prioritise organisational constraints. It should be noted further, that constraint relationships of different IAMs all have a disjoint scope – the respective $\mathcal{X}^a$ sets – and that the constraint relationships are subject to the variable renaming as well.

$$\mathcal{R} = \mathcal{R}^\lambda \cup \bigcup_{a \in \mathcal{A}^\lambda} \mathcal{R}^a \cup \{(x, y) \mid x \in \mathcal{C}^\lambda_s, y \in \mathcal{C}^a_s, \nexists z : z \succ_{\mathcal{R}^a} y\}$$

Step 3: Finally, as the optimisation function is defined in terms of the instantiated interface variables, $\mathcal{X}^{\mathcal{A}^\lambda}$, we can keep f as the objective of ζ. This yields two possibly conflicting objectives regarding the satisfaction of individual and organisational soft constraints versus meeting the original objective f. We could then use an existing multi-objective optimisation technique such as utopia search [24] but propose an alternative three- stage optimisation to strongly favour the organisational objective f while still optimising p if possible.

C. Multi-objective Optimisation with Constraint Relationships

During synthesis, constraint relationships are responsible for combining soft constraint priorities as well as establishing a

connection to organisational constraints. Fig. 4 shows how agents have different preferences on soft constraints that only affect their individual performance, e.g., *economically optimal* states that prescribe that production should better be in this range whereas other values are technically feasible. Optimizing the SM $\zeta = \langle \mathcal{X}, \mathcal{D}, \mathcal{C}, \mathcal{R}, f \rangle$ with respect to the original objective *and* soft constraint satisfaction (as defined by the penalty function p in Eq. 1) can be achieved as follows (w. l. o. g. we restrict ourselves to minimization problems):

1) Let $\hat{f}$ be the optimal result of ζ (with the original objective f).
2) Find an upper/lower bound f^* for the objective such that $f^* = \delta_f \times \hat{f}$ for some x, where $\delta_f \geq 1.0$ for a minimisation problem. Note that for a non-negative minimisation objective, $\delta_f \times \hat{f} = 0$ so the bound collapses to a single point, 0 such that any valid solution has to be optimal if at least one is.
3) Impose a constraint to restrict $f(\theta)$, $\theta \in (\mathcal{X} \rightarrow \mathcal{D})$. Let ζ' be the problem that minimises the violation of penalties with respect to the bound f^*: $\zeta' = \langle \mathcal{X}, \mathcal{D}, \mathcal{C} \cup \{c' : (f(\theta) \leq f^*\}, \mathcal{R}, p \rangle$.
4) Solve ζ' and let $\hat{p}$ be the minimal sum of penalties of violated constraints.
5) Impose a restriction on the sum of penalties to be less than or equal to $\hat{p} \times \delta_p$ and solve for the original objective: $\zeta'' = \langle \mathcal{X}, \mathcal{D}, \mathcal{C} \cup \{c'' : (p(\theta) \leq \hat{p}\}, \mathcal{R}, f \rangle$. This step is necessary as otherwise a solver might just lie within the tolerance of f^* even if better solutions in terms of f (having the same penalty sum) exist.

Appropriate choices for δ_f and δ_p have to be found specific to a problem but so do weights of a single combined objective function commonly employed in a global criterion method [24]. This approach can certainly be costly if the optimisation problems themselves are hard and time-consuming to solve – but if the coalitions are sufficiently small and organised (as, e.g., in a hierarchical system [9]) the benefits of finding solutions that are attractive to the individual participants outweigh the expensive optimisation runs. Moreover, the solutions found in previous steps can be used as starting points in subsequent runs to speed up the optimisation and other multi-objective optimisation approaches such as utopia search require several optimisation runs as well.

V. Implementing Synthesis in CPLEX

We exemplify the synthesis process with an example of a VPP consisting of three power plants with heterogeneous constraints modelling the requirements presented in Sect. III. For each power generator, we formulate the individual constraints and explain the usage of local and interface variables. The problem is comparable to the parametrised model presented in Listing 1. The models are directly presented in the Optimisation Programming Language (OPL) used in CPLEX [18] to provide a prototype even though our concepts do not rely on any specific CPLEX features such that the models could equivalently have been presented in a pseudo-code for CSOPs. Extensions to OPL regarding soft constraints are therefore implemented using the keyword *SOFT-CONSTRAINTS* in comments. We do not describe entire models (which are provided online[2]) but rather highlight the most important aspects.

A. Organisational Template

We start with the organisational template as it contains the core optimisation problem to be tackled. The first section indicates the identifier of the set consisting of the agents' identifiers (*plants* in our case) as well as generally needed constants such as the time series and load curve.

```
{string} plants = ...;
range TIMERANGE = 0..5;
float loadCurve[TIMERANGE] =
  [200.0, 250.0, 230.0, 247.0, 349.0, 551.0];
```

The set of *plants* needs to be filled with the identifiers of actual power plant models. *TIMERANGE* and *loadCurve* are furthermore examples of interface variables that individual agents may use to define their local variables and constraints. Additional decision variables and expressions common to all agents are described in the next section. *Production* for different time steps are the decision variables in this example and we require each agent to provide its maximal output (nameplate capacity) such that we are able to define the load factor of each plant. Note that these decision variables are indexed by the set of child agents ($\mathcal{X}^{\mathcal{A}^\lambda}$). Decision expressions primarily are syntactic tools to facilitate the formulation of optimization functions and constraints. They have to be fully determined by the value of the decision variables but offer to aggregate a set of decision variables using the *sum*, *min*, or *max* constructs. The presented expressions serve to provide a formulation of the constraint presented in Eq. 4.

```
float P_max[plants] = ...;
dvar float+ production[plants][TIMERANGE];
dexpr float totalProduction[t in TIMERANGE] =
  sum(p in plants) production[p][t];
dexpr float loadFactor[p in plants][t in TIMERANGE] =
  production[p][t] / P_max[p];
dexpr float minLoadFact[t in TIMERANGE] =
  min(p in plants) loadFactor[p][t];
dexpr float maxLoadFact[t in TIMERANGE] =
  max(p in plants) loadFactor[p][t];
```

Violation takes the sum of the absolute values of the deviation between the aggregated production and the load curve. It is the quantity that we aim to minimize.

```
dexpr float violation = sum(t in TIMERANGE)
  abs(totalProduction[t]-loadCurve[t]);

minimize violation;
```

We can explicitly request that all productions are below the nameplate capacity as a *hard* constraint. Furthermore, this organisational template provides for two soft constraints regarding distributions of load — denoted by $\mathbf{oc1}$ and $\mathbf{oc2}$. They indicate that the load factor of the individual agents should not

[2]Please refer to footnote 4

vary too much if possible thereby avoiding a highly skewed resource allocation. In particular, power plant operators might expect to have their generator contribute at least to a certain extent to generate revenue. Indifference between **oc1** and **oc2** is expressed by not modelling a relationship.

```
subject to {
  forall (t in TIMERANGE) {
    oc1: minLoadFact[t] >= 0.4;
    oc2: maxLoadFact[t] <= 0.6;
    forall (p in plants) {
      production[p][t] <= P_max[p];
    }
  }
};

/* SOFT-CONSTRAINTS
oc1
oc2 */
```

Compared to Listing 1, we only require a subset of the homogeneous aspects (the maximal power). However no restriction in terms of rates of change and other, inertia-based constraints are imposed as they are part of the individual agent models.

B. Individual Agent Models

We describe three types of power plant models consisting of the constraints presented in Sect. III. These are implemented independently from the organisational template but also contain definitions for the interface variables to be tested in advance. The first type, **A**, does not implement a warm and cold start-up and does not foresee minimal up and down times. Thus, it corresponds to a power plant that can be started up fast enough for the considered time step durations [19]:

```
float P_min = 50.0; float P_max = 100.0;

float rateOfChange = 0.15;
dvar float production[TIMERANGE];
dexpr int running[t in TIMERANGE] = !(production[t] == 0);

subject to {
  forall (t in 0..TIMERANGE) {
    running[t] => production[t] >= P_min;
    rate_of_change: (running[t] == 1) && (running[t+1] == 1)
     => abs(production[t] - production[t+1])
          <= production[t] * rateOfChange;
    c1: (running[t] == 1) && (running[t+1] == 1)
     => abs(production[t] - production[t+1])
          <= production[t] * 0.07;
    c2: (running[t] == 1) && (running[t+1] == 1)
     => abs(production[t] - production[t+1])
          <= production[t] * 0.10;
  }
};
/* SOFT-CONSTRAINTS
 c1 >> c2 */
```

During synthesis, the system automatically distinguishes between *P_max* being an interface variable and *P_min* being a local variable. Some power generators such as hydro power plants could be throttled down to no production at all, whereas others can enforce a minimal operation production as modelled in this example. The decision expression *running* helps syntactically to distinguish between these states. Note that here, *production* is not indexed over any agent set, so an individual agent model only has access to its own decision variables. We furthermore have two constraints reflecting a preference for small rates of change which aim for operational stability.

The second type, **B**, is used to express minimal uptimes [15]. We need additional decision variables to capture the number of time steps a plant is running consecutively at a particular time step. Based on these, we can decide whether a transition from on to off is feasible. And for the sake of the argument, assume that no relative rate of change but rather a fixed rate of change at every production level is given (we omit variables already discussed). We also assume three ranges of different economical preference that can be expressed with soft constraints:

```
int minUpTime = 2;
float fixedChange = 20;

dvar int+ consRunning[TIMERANGE];
 [...]
forall (t in TIMERANGE) {
  c1: production[t] >= 22 && production[t] <= 25;
  c2: production[t] >= 20 && production[t] <= 30;
  c3: production[t] >= 18 && production[t] <= 33;

  fixed_change: (running[t] == 1 && running[t+1] == 1) =>
    abs(production[t] - production[t+1]) <= fixedChange;
  cons_run: (running[t+1] == 1 &&
              consRunning[t+1] == (1 + consRunning[t])) ||
            (running[t+1] == 0 &&
              consRunning[t+1] == 0);

  min_up_time: (running[t] == 1 && running[t+1] == 0) =>
                (consRunning[t] - minUpTime) >= 0;
}
/* SOFT-CONSTRAINTS
 c1 >> c2
 c2 >> c3 */
```

Note that it becomes apparent that a parametrised model would now be severely limited. To support minimal up times within a MIP framework, we need those variables (or implement custom constraints and propagators for a constraint solver) but would have to offer these variables for *all* plants.

Finally, our third type, **C**, incorporates hot and cold start-up times based on the down time (with down time being defined analogously to up time in the previous model). In essence, we provide a MIP formulation for the transition system presented in Fig. 2 implementing start-up times that respect the data shown in Table I. We use a stepwise function that returns 2 time steps duration for down times of less than 3 and 4 time steps for longer down times. A decision variable *signal* stores when to initiate a start-up process and we need to enforce that those signals are only sent when a plant is in the appropriate state (e.g., sending a start-up signal only when in the *idle* state). Additionally, we also have three soft constraints (c1, c2, c3) regarding economical ranges and rates of change.

```
int IDLE = 0;
int STARTING = 1;
```

```
int STOPPING = 2;
int UP = 3;
stepFunction startUp = stepwise{ 2->3; 4 };

dvar int+ countdown[TIMERANGE];
dvar int+ powerPlantState[TIMERANGE] in 0..3;
dvar int+ consStopping[TIMERANGE];
dvar int+ consRunning[TIMERANGE];
dvar int+ state[TIMERANGE] in 0..3;
dvar int signal[TIMERANGE] in -1 .. 1;
 [...]
forall (t in TIMERANGE) {
 c1: production[t] >= 300.0 && production[t] <= 350.0;
 c2: production[t] >= 280.0 && production[t] <= 370.0;
 c3: abs(production[t] - production[t+1]) <= 20;

  signal_states : signal[t] == 1 => state[t] == IDLE &&
                  signal[t] == -1 => state[t] == UP;
 state[t] == IDLE =>
        ( state[t+1] == IDLE ||
        ( state[t+1] == STARTING &&
          signal[t] == 1 &&
         (countdown[t+1] == startUp(consStopping[t ]))));

 ( state[t] == STARTING && countdown[t] >= 1) =>
        ( state[t+1] == STARTING &&
          countdown[t+1] == countdown[t] - 1);
 ( state[t] == STARTING && countdown[t] == 0) =>
          state[t+1] == UP;
 state[t] == UP => (state[t+1] == UP ||
       ( state[t+1] == IDLE && signal[t] == -1));

}
/* SOFT-CONSTRAINTS
 c1 >> c2
 c1 >> c3 */
```

This model is significantly more detailed than the previous ones and indicates how future realistic models could be integrated to achieve more accurate schedules.

C. Synthesised Model

Combining these three individual agent models with the organisational template results in one CPLEX model where constraints, local variables and interface variables are replaced as described earlier. In contrast to the formal definition (where x would just be replaced by x^a), we can however distinguish interface from local variables in the synthesised model. This is due to the fact that it can come in handy to have interface variables indexed by plant identifiers (e.g., for defining the expression `totalProduction`). As we cannot assume local variables to be available for every individual agent, we add the plant identifier to the variable name, such `rateOfChange` would become `rateOfChange_a` for a plant `a`. The individual constraint relationship graphs are combined with the organisational template yielding a synthesised graph as depicted in Fig. 4. We give some snippets that show parts of the synthesised model.

```
{string} plants = {"b", "c", "a"};
float P_max[plants] = [35.0, 400.0, 100.0];
dexpr float totalProduction[t in TIMERANGE] =
  sum( p in plants ) production[p][t];
 [...]
```

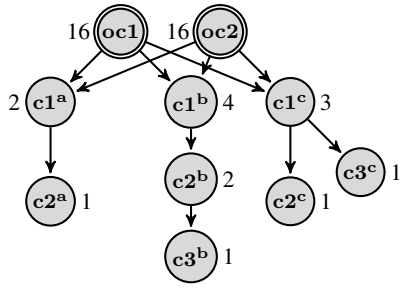


Fig. 4. Synthesised Constraint Relationship Graph with weights presented for TPD

```
float rateOfChange_a = 0.15;
float fixedChange_b = 5;
dvar int+ consStopping_b[TIMERANGE];
 [...]
subject to {
  forall (t in TIMERANGE) {
   c1_b: production["b"][t] >= 22 && production["b"][t] <= 25;
  }
   [...]
};
/* SOFT-CONSTRAINTS
oc1
oc2
oc1 >> c1_b;
oc1 >> c1_a;
oc2 >> c1_a; */
```

The last step to acquire a solvable model is then to transform the constraint relationships into weights and add decision variables and expressions to incorporate these weights as penalties to come to a scalar objective function. Essentially, all soft constraints are collected and used to index a vector of penalties and a reformulation is performed for each soft constraint $c \in \mathcal{C}_s$:

$$c' : (c \wedge p_c = 0) \vee (\neg c \wedge p_c = w(c))$$

which leads to the following changes in OPL:

```
{string} softConstraints = {"c1_b", "c1_c" ,[...]};
dvar int+ penalties[softConstraints][TIMERANGE];
dexpr float penaltySum = sum(t in TIMERANGE,
                             c in softConstraints) penalties[c][t];
 [...]
c1_b: (production["b"][t] >= 22 && production["b"][t] <= 25
       && penalties["c1_b"][t] == 0) ||
      (!( production["b"][t] >= 22 && production["b"][t] <= 25)
       && penalties["c1_b"][t] == 4);
```

We then have `penaltySum` as the expression to minimize the violation of soft constraints.

VI. Evaluation

Evaluating our approach at this stage is not straightforward as to the authors' knowledge, no public benchmark library containing unit commitment models is available. In addition, the described synthesis is primarily a methodology to incorporate heterogeneous models, not a concrete algorithm, thus performance issues are not yet in the focus of development. To

still get a quantitative (and as objective as possible) evaluation about how well individual preferences can be considered, we performed synthetic randomised experiments based on the constraints presented in the paper and power plant data[3].

The source code of both the experiments and the example discussed in Sect. V are available online[4] to facilitate replication of our experiments.

A. Experimental Design

We first generate a pool of 400 power plant models based on the types presented in Sect. V taking into account existing data of gas turbines and biomass power plants to capture plant-specific properties. We then perform a number of optimisation runs by picking a set out of these power plants randomly, synthesising them into one model and running the three-stage-optimisation with the *synthesised model*. The following parameters characterise one experiment:

- n the number of power plants
- δ_v the delta applied to the first optimum in the three-stage-optimisation as a tolerance
- δ_p the delta applied to the penalty in the third step
- d the dominance level SPD or TPD
- r the number of runs that are conducted

Several measurements are taken to evaluate properties of the synthesis and averaged over the r runs:

- v_1 – v_3 the violation (discrepancy between demand and production) for the three optimisation steps
- p_1 – p_3 the same holds for the aggregated penalties
- $\overrightarrow{v}$, $\overrightarrow{q}$ measures the relative improvement or worsening of v_1 to v_3 and p_1 to p_3 to show how the objectives change when considering penalties: $\overrightarrow{v} = \frac{v_1}{v_3}$, $\overrightarrow{p} = \frac{p_1}{p_3}$
- **#pred** is the average number of predecessors per violated constraint as a measure for how more important constraints are treated
- **#vc** sums up the total number of violated soft constraints relative to the number of all soft constraints (with $\overrightarrow{\#\text{vc}}$ denoting the relative improvement from the first optimisation step to the third)

Note that #pred and #vc are evaluated for the final result returned by the three-stage-optimisation. Some runs failed to provide a correct solution within a threshold of 3 minutes. We excluded those results from the evaluation as our focus is not yet on performance and robustness issues.

B. Experimental Results

We examine questions of interest and present the results of the experiment runs.

a) Solution quality: How does taking into account individual preferences affect the solution quality? We want to obtain an impression about how the choice of the parameters δ_v and δ_p affect the solution quality. For all choices of parameters, we found that allowing a small tolerance regarding the mismatch of demand and production allowed for comparable, substantial improvements in the satisfaction of soft constraints as Table II shows. If a violation tolerance of 10% was imposed, the solver still managed to reduce the number of violated soft constraints by half from the optimal solution to the final one while staying within a range of 1 % optimality. As expected, increasing the violation tolerance leads to better reductions in terms of penalties and soft constraints.

TABLE II
COMPARISON OF DIFFERENT DELTAS REGARDING THE SOLUTION QUALITY. MEASUREMENTS REPRESENT AVERAGES OVER 50 RUNS FOR 9 TIME STEPS WITH $n = 5, d = \text{SPD}$ AND STANDARD DEVIATIONS.

(δ_v/δ_p)	1.1/1.2	1.2/1.2	1.2/1.1	1.3/1.2
$\overrightarrow{v}$	1.0039	1.0161	1.056	1.022
	(0.0076)	(0.0286)	(0.044)	(0.034)
$\overrightarrow{p}$	0.599	0.546	0.534	0.527
	(0.116)	(0.142)	(0.142)	(0.148)
$\overrightarrow{\#\text{vc}}$	**0.555**	**0.502**	**0.486**	**0.487**
	(0.132)	(0.152)	(0.181)	(0.163)

TABLE III
COMPARISON OF DIFFERENT DOMINANCE LEVELS. MEASUREMENTS REPRESENT AVERAGES OVER 50 RUNS FOR 9 TIME STEPS WITH $\delta_v = \delta_p = 20\%$ AND STANDARD DEVIATIONS.

(n/d)	5/SPD	5/TPD	10/SPD	10/TPD
$\overrightarrow{v}$	1.022	1.0001	1.027	1.0
	(0.027)	(0.0009)	(0.038)	(0.0005)
$\overrightarrow{p}$	0.595	0.7484	0.45	0.774
	(0.169)	(0.15)	(0.085)	(0.168)
#pred	**2.905**	**2.541**	**3.23**	**2.76**
	(0.406)	(0.2569)	(0.41)	(0.281)
#vc	**0.34**	**0.468**	**0.283**	**0.462**
	(0.12)	(0.138)	(0.064)	(0.123)

b) Influence of dominance property: How does the selected dominance property affect the number of violated soft constraints? The dominance property influences how much more important a single constraint is with respect to its dominated constraints. In the case of single predecessor dominance, a constraint is only more important than one of its predecessors, not a whole set. Therefore the weights lie more closely to each other and no strong "hierarchical" difference is imposed. Choosing the property, however, is not straightforward as this substantially influences the number of soft constraints that are "dropped" by a solver in favour of more important ones. We observe this behaviour in Table III: The percentage of violated soft constraints is significantly higher when using TPD than SPD for both 5 and 10 power plants while the average number of predecessors per violated constraint (measuring its importance) is lower when using TPD. TPD semantics lead to an average dissatisfaction of about 40% of all soft constraints, whereas SPD only dissatisfies 30% albeit returning slightly worse solutions (about 2% higher demand violations when 20% higher were allowed). It is thus a relevant question for preference elicitation and requirements engineering to find out whether the system's constraints are more hierarchical or egalitarian. Constraint hierarchies correspond to TPD semantics [9] and we argue

[3]see http://www.energymap.info/ and http://www.lew-verteilnetz.de/
[4]https://github.com/Alexander-Schiendorfer/synthesisenergycoalitions

that there are certainly circumstances where having more soft constraints fulfilled is more desirable than just satisfying the most important ones and no others.

VII. Conclusion and Outlook

Although there is a vast body of literature on efficient models and algorithms for unit commitment problems (see, e.g., [20], [14], [19], [15] for a modest selection), to the best of our knowledge there is no approach that consolidates these various types of models and addresses the domain's inherent heterogeneity. We proposed an approach that attempts to leverage existing models and to simplify the engineering of optimisation problems with a well-defined modelling methodology. In addition, we showed how start-up behaviour of thermal power generators presented in [16] or [19] can be extended to account for start-up times that depend on the previous down time using transition systems. We showed a MIP formulation for this adaptive start-up problem and used heterogeneous models employing different start-up behaviour as the case study for our approach. Moreover, the modelling and synthesis strategy presented allows unit operators to express individual preferences that could not be considered before. Our first experiments indicate that this may increase the willingness of single power generators to collaborate in collective schemes where autonomy is sacrificed for potential economic benefits as we could halve the number of violated soft constraints. As the process is fully automated, we plan to combine it with models of the resource-levelling problem faced in the scheduling of (short-lived) coalitions of energy consumers to achieve better prices.

Acknowledgment

This research is partly sponsored by the German Research Foundation (DFG) in the project "OC-Trust" (FOR 1085).

References

[1] S. D. Ramchurn, P. Vytelingum, A. Rogers, and N. R. Jennings, "Putting the 'smarts' into the smart grid: a grand challenge for artificial intelligence," *Commun. ACM*, vol. 55, no. 4, pp. 86–97, Apr. 2012. doi: 10.1145/2133806.2133825. [Online]. Available: http://doi.acm.org/10.1145/2133806.2133825

[2] H. Morais, P. Kádár, P. Faria, Z. A. Vale, and H. Khodr, "Optimal scheduling of a renewable micro-grid in an isolated load area using mixed-integer linear programming," *Renewable Energy*, vol. 35, no. 1, pp. 151 – 156, 2010. doi: http://dx.doi.org/10.1016/j.renene.2009.02.031. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0960148109001001

[3] P. Arcuri, G. Florio, and P. Fragiacomo, "A mixed integer programming model for optimal design of trigeneration in a hospital complex," *Energy*, vol. 32, no. 8, pp. 1430 – 1447, 2007. doi: http://dx.doi.org/10.1016/j.energy.2006.10.023. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360544206003197

[4] L. G. Fishbone and H. Abilock, "Markal, a linear-programming model for energy systems analysis: Technical description of the bnl version," *International Journal of Energy Research*, vol. 5, no. 4, pp. 353–375, 1981. doi: 10.1002/er.4440050406. [Online]. Available: http://dx.doi.org/10.1002/er.4440050406

[5] S. Thiébaux, C. Coffrin, H. Hijazi, and J. Slaney, "Planning with MIP for supply restoration in power distribution systems," in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, ser. IJCAI'13. AAAI Press, 2013. ISBN 978-1-57735-633-2 pp. 2900–2907. [Online]. Available: http://dl.acm.org/citation.cfm?id=2540128.2540546

[6] T. Werner, "DEMS – The Decentralized Energy Management System." http://w3.siemens.com/smartgrid/global/en/smart-grid-world/experts-talk/pages/dems.aspx, 2013, [Online; accessed 07-January-2014].

[7] "PLEXOS Integrated Energy Model," http://energyexemplar.com/software/plexos-desktop-edition/, January 2014, [Online; accessed 09-January-2014].

[8] C. Tesche, "Stadtwerke München und Siemens nehmen ein virtuelles Kraftwerk in Betrieb (in German)," http://www.energie-und-technik.de/smart-grid-smart-metering/news/article/87399/0/Stadtwerke_Muenchen_und_Siemens_nehmen_ein_virtuelles_Kraftwerk_in_Betrieb/, 2012, [Online; accessed 07-January-2014].

[9] A. Schiendorfer, J.-P. Steghöfer, and W. Reif, "Synthesis and Abstraction of Constraint Models for Hierarchical Resource Allocation Problems," in *Proc. of the 6th International Conference on Agents and Artificial Intelligence (ICAART)*, vol. 2. SciTePress, March 2014. doi: 10.5220/0004757700150027. [Online]. Available: http://dx.doi.org/10.5220/0004757700150027

[10] E. Tsang, *Foundations of constraint satisfaction*. Academic press London, 1993, vol. 289.

[11] A. Schiendorfer, J.-P. Steghöfer, A. Knapp, F. Nafz, and W. Reif, "Constraint relationships for soft constraints," in *Research and Development in Intelligent Systems XXX*, M. Bramer and M. Petridis, Eds. Springer International Publishing, 2013, pp. 241–255. ISBN 978-3-319-02620-6. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-02621-3_17

[12] P. Meseguer, F. Rossi, and T. Schiex, "Soft Constraints," in *Handbook of Constraint Programming*, F. Rossi, P. van Beek, and T. Walsh, Eds. Elsevier, 2006, ch. 9.

[13] G. Anders, A. Schiendorfer, J.-P. Steghöfer, and W. Reif, "Robust Scheduling in a Self-Organizing Hierarchy of Autonomous Virtual Power Plants," in *Proc. of the 2nd International Workshop on „Self-optimisation in Organic and Autonomic Computing Systems" (SAOS14) in conjunction with ARCS 2014*, vol. 2, February 2014.

[14] L. Rani, M. Mam, and S. Kumar, "Economic load dispatch in thermal power plant taking real time efficiency as an additional constraints," *International Journal of Engineering Research & Technology (IJERT)*, vol. 2, no. 7, 2013. [Online]. Available: www.ijert.org

[15] N. Padhy, "Unit commitment-a bibliographical survey," *Power Systems, IEEE Transactions on*, vol. 19, no. 2, pp. 1196–1205, 2004. doi: 10.1109/TPWRS.2003.821611

[16] J. Arroyo and A. Conejo, "Modeling of start-up and shut-down power trajectories of thermal units," *Power Systems, IEEE Transactions on*, vol. 19, no. 3, pp. 1562–1568, 2004. doi: 10.1109/TPWRS.2004.831654

[17] L. Jarass and G. Obermair, *Welchen Netzumbau erfordert die Energiewende?: Unter Berücksichtigung des Netzentwicklungsplans Strom 2012 (in German)*, ser. MV-Wissenschaft. Monsenstein und Vannerdat, 2012. ISBN 9783869916415. [Online]. Available: http://books.google.de/books?id=jID5HQcICLkC

[18] "IBM ILOG CPLEX Optimizer," http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/, December 2013.

[19] J. Šumbera, "Modelling generator constraints for the self-scheduling problem," in *Vědecký seminář doktorandů FIS – únor 2012*, Prague, Czech Republic, Feb 2012. ISBN 978-80-245-1862-6

[20] C. Rajan, "An evolutionary programming based tabu search method for unit commitment problem with cooling-banking constraints," in *IEEE Power India Conference, 2006*, 2006. doi: 10.1109/POWERI.2006.1632557 p. 8.

[21] J. N. Mayer, N. Kreifels, and B. Burger, "Kohleverstromung zu Zeiten niedriger Börsenstrompreise," Aug. 2013, http://www.ise.fraunhofer.de/de/downloads/pdf-files/aktuelles/kohleverstromung-zu-zeiten-niedriger-boersenstrompreise.pdf/view.

[22] V. Dignum and J. Padget, "Multiagent organizations," *Multiagent Systems, G. Weiss, ed., MIT Press,*, 2013.

[23] C. Derksen, T. Linnenberg, R. Unland, and A. Fay, "Unified energy agents as a base for the systematic development of future energy grids," in *Multiagent System Technologies*, ser. Lecture Notes in Computer Science, M. Klusch, M. Thimm, and M. Paprzycki, Eds., vol. 8076. Springer Berlin Heidelberg, 2013. doi: 10.1007/978-3-642-40776-5_21. ISBN 978-3-642-40775-8 pp. 236–249. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-40776-5_21

[24] R. Marler and J. Arora, "Survey of multi-objective optimization methods for engineering," *Structural and Multidisciplinary Optimization*, vol. 26, no. 6, pp. 369–395, 2004. doi: 10.1007/s00158-003-0368-6. [Online]. Available: http://dx.doi.org/10.1007/s00158-003-0368-6

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1539–1547
DOI: 10.15439/2014F287
ACSIS, Vol. 2

A New Intrusion Prevention System for Protecting Smart Grids from ICMPv6 Vulnerabilities

Manali Chakraborty, Nabendu Chaki
University of Calcutta
Kolkata, India
Email: manali4mkolkata@gmail.com
nabendu@ieee.org

Agostino Cortesi
DAIS
Universita Ca'
Foscari Venezia
Email: cortesi@unive.it

***Abstract*— Smart Grid is an integrated power grid with a. reliable, communication network running in parallel towards providing two way communications in the grid. It's trivial to mention that a network like this would connect a huge number of IP-enabled devices. IPv6 that offers 18-bit address space becomes an obvious choice in this context. In a smart grid, functionalities like neighborhood discovery, autonomic address configuration of a node or its router identification may often be invoked whenever newer equipments are introduced for capacity enhancement at some level of hierarchy. In IPv6, these basic functionalities like neighborhood discovery, autonomic address configuration of networking require to use Internet Control Message Protocol version 6 (ICMPv6). Such usage may lead to security breaches in the grid as a result of possible abuses of ICMPv6 protocol. In this paper, some potential newer attacks on Smart Grid have been discussed. Subsequently, intrusion prevention mechanisms for these attacks are proposed to plug-in the threats.**

I. Introduction

A SMART grid is an intelligent energy network that integrates the actions of all users connected to it and makes use of advanced information, control, and communication technologies to save energy, reduce cost and increase reliability and transparency [1].

The backbone of the Smart Grid will be its communication network. This network is to connect the different components of the Smart Grid together, and provide two-way communication. IPv6 is a new technology which gained a massive attention, as a supporting layer in smart grid communication. The huge address space of IPv6 supports the network architecture of the smart grid communications. Besides, features like stateless address auto configuration (SLAAC) and IPSec support makes IPv6 more suitable for smart grid. IPv6 also supports prioritization of messages and different Quality of Service models, which complements several smart grid applications [8]. However, with these new advancements in technology, IPv6 is also exposed to various attacks, such as header modification attack, fragmentation attacks , etc. [5], [6]. In this paper, we focus on some of the possible ICMPv6 attacks that are particularly relevant in the context of building networking infrastructure between Smart Meters (SM), Data Collection Units (DCU) and Meter Data Management System (MDMS). We would demonstrate how these could affect the Smart Grid before proposing appropriate Intrusion Prevention Systems (IPS) to protect the grid from such attacks.

IPv4 networks often filter ICMP messages to avoid security concerns. However, for IPv6, this is not possible. ICMPv6 is used for basic functionalities and used by other IPv6 protocols like Neighbor Detection Protocol (NDP). Neighbor Discovery Protocol (NDP) is a protocol used with IPv6 to perform various tasks like router discovery, auto address configuration of a node, neighbor discovery, Duplicate Address Detection, determining the Link Layer addresses of other nodes, address prefix discovery, and maintaining routing information about the paths to other active neighbor nodes [4]. Thus, the implementation of IPv6 in Smart Grid needs some serious care to protect from the security vulnerabilities of the ICMPv6 protocol. NDP uses five ICMPv6 messages. These are:

- Router Solicitation (RS) message: Hosts send RS message to enquire about a legitimate router on the link.
- Router Advertisement (RA) message: Routers send RA message, either periodically or in response to RS message.
- Neighbor Solicitation (NS) message: Hosts send NS message to determine the link layer address of a specific node, and also to verify whether an address is already present on link or not.
- Neighbor Advertisement (NA) message: Hosts send NA message in response to the NS message.
- Router Redirect (RR) message: Routers send RR message to inform a host about a better router on its link.

With higher degree of autonomic control and decision making, a smart grid also becomes subject to several security concerns. Smart grid is generally considered as a heterogeneous, backward compatible, static, self adapting and self healing network, with a large number of devices, where two way communications is provided between Smart Meters and a Supervisory Control and Data Acquisition (SCADA) system. This requires special QoSs, like high restriction on delay, failure and voltage quality [3]. In smart

978-83-60810-58-3/$25.00

grid, availability and integrity are typically considered more important than confidentiality [9]. Also the risk factor is quite high in smart grid as compared to traditional networks. Thus, the existing solutions for cyber security often fall short of the typical requirements for a smart grid.

Some work has been done to secure smart meters and communication network of Smart Grid or SCADA systems [10]. An IPv6 based moving target defense system is provided in [11] to secure the communication between hosts. Most of the network attacks target some specific addresses, so, moving the target address will prevent hosts from being located for an attack. [12], [13], [14] explains different techniques for IPv6 address configuration schemes for smart grid. However, security solutions for specific IPv6 problems, like ICMPv6 attacks, for Smart Grid environment are still need to be addressed. In [16], a distributive, trust based approach to detect attacks in Duplicate Address Detection (DAD) phase was proposed. However, this concentrates only on one type of attack in DAD. In [17], the requirements and practical needs for monitoring and intrusion detection in AMI is discussed. In [18], a layered combined signature and anomaly-based IDS for HAN was proposed. This IDS was designed for a ZigBee based HAN which works at the physical and medium access control (MAC) layers. However, the work only considers the HAN part of AMI. In [19], a specification-based IDS for AMI is proposed. While the solution in [19] relies on protocol specifications, security requirements and security policies to detect security violations, it would be expensive to deploy such IDS since it uses a separate sensor network to monitor the AMI.

We have proposed a new Intrusion Prevention System for providing security against ICMPv6 attacks in smart grid networks. The structure of the paper is as follows. First, we discuss three important functionalities for a Smart Meter: Router Discovery, Duplicate Address Detection and Neighbor Discovery, using NDP and various ICMPv6 messages. Possible attacks and the effects of those attacks on smart grid are analyzed for each function. Finally, we propose an Intrusion Prevention System (IPS) to prevent the attacks in the Router Discovery phase and detect the attacks in the Duplicate Address Detection and Neighbor Discovery phase.

Notice that we do not claim that using NDP or ICMPv6 is the only option for realizing functionalities like router discovery or address configuration in a smart grid. As for example, instead of having an auto configurable addressing scheme, smart grids may also have independent Certifying Authority (CA) for providing addresses to newly installed SMs. However, the cost of installation and maintenance of such centrally controlled architecture may be avoided using auto configurable SMs. This paper aims to expose the security threats there and to propose suitable intrusion prevention mechanisms to safeguard smart grids from ICMPv6 misuses.

II. Smart Grid And ICMPv6 Abuses

Figure 1 shows the communication architecture of Smart Grid. Smart Energy Utility Network (SUN) hierarchically consists of three components: Home Area Network (HAN), Neighborhood Area Network (NAN), and Wide Area Network (WAN) [15]. The HAN provides the communication between the Smart Meters in a home and other appliances in that home. The NAN connects SMs to the Data Collection Units (DCUs), and WAN provides access between the DCUs and Meter Data Management System (MDMS). DCU collects data from hundreds of SMs and sends them to the MDMS. At the lowest level, the smart meters act as hosts in a network and DCUs are the routers of the network. We assume that

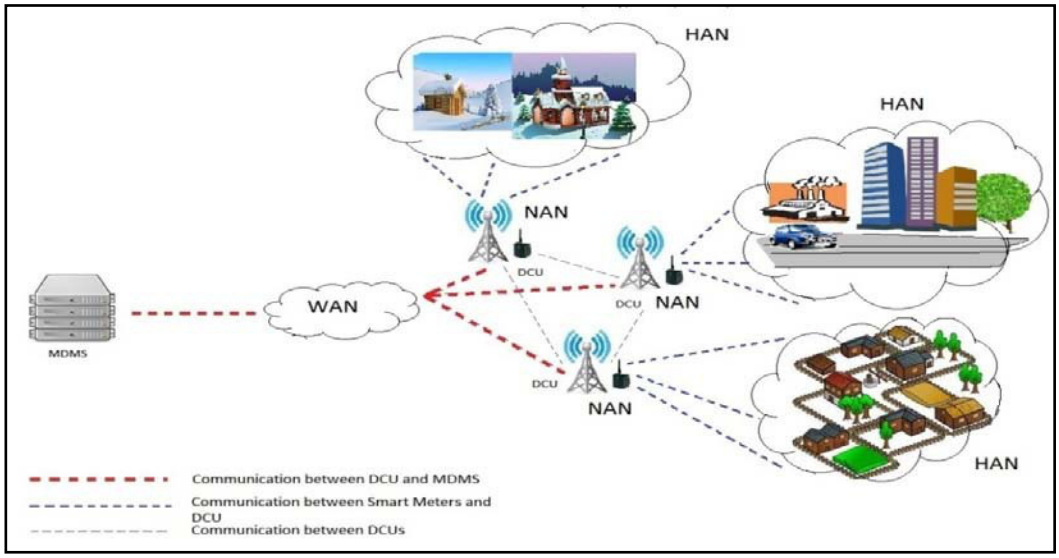


Fig. 1: Communication Architecture of Smart Grid

- Smart Meters are managed by DCUs. When a SM X is installed in a subnet, it should find a DCU, say R, to bind with. X will continue to communicate through DCU R,

until it receives any ICMPv6 Router Redirect (RR) message from R.

- Each DCU keeps a neighbor cache, storing the addresses of all DCUs in its neighborhood.
- Each subnet has a different and unique 64 bit prefix address for addressing SMs within the subnet.
- Each DCU communicates with the SMs within its subnet, and then transmits the aggregated data to DCU.
- Every SM keeps a neighbor cache to store addresses of all its one hop neighbors

A. Router Discovery

When a SM X is installed in a subnet, it should find a DCU to bind with. The Smart Meter X will continue to communicate through that DCU, until it receives any ICMPv6 Router Redirect (RR) message from the previous DCU.

Normal Procedure for Router Discovery

Normally Smart Meters discover their router or DCU through the following steps,

- First, X sends an ICMPv6 Router Solicitation (RS) message to locate a DCU in its local link.
- A legitimate DCU then responds with an ICMPv6 Router Advertisement (RA) message, with a 64 bit prefix address for its subnet.
- Then X registers that DCU as its default router in the link, and auto-configures a global unicast address based on the received prefix.

Attacks in Router Discovery phase

The most prominent attack in this phase occurs if an attacker falsely claims to be a DCU. It can spoof an RA message from a legitimate DCU and send it to the Smart Meter, with or without altering the prefix address for that subnet. In either case, the newly installed Smart Meter registers the attacker as its DCU. If the adversary alters the prefix address, then the Smart Meter will auto-configure its global address based on a wrong prefix. As a result, the Smart Meter will get blocked in the subnet and will not be able to communicate with any other Smart Meter or DCU except the attacker. The situation becomes a bit more complex when the adversary sends the RA message without changing the prefix. In this situation, the Smart Meter can communicate within its subnet. However, it becomes quite impossible for the Smart Meter to communicate beyond its subnet as the registered DCU for the Smart Meter is an attacker who is not recognized by other Smart Meters in the Neighborhood Area Network.

Once an adversary successfully convinces a newly installed Smart Meter of being its valid DCU, it can launch a myriad of conventional network attacks on the Smart Grid. It can launch a man-in-the-middle attack by intercepting packets from the Smart Meters or from the DCUs and suitable changing the Source and Destination address fields such that neither of these two entities are aware of the presence of an attacker in between. The attacker can also tweak the data contained in the intercepted packets. Another traditional network attack is the Denial-of-Service attack. The attacker can overload the network resources by generating spurious packets having the newly installed Smart Meter address as the Source Address.

B. Duplicate Address Detection

After auto configuring the address for itself, the Smart Meter X will want to know whether the address is available for use.

Normal Procedure for Duplicate Address Detection

The following steps are used for duplicated address detection.

- Smart Meter X, sends an ICMPv6 Neighbor Solicitation message for the address it wants to claim.
- If any Smart Meter on that subnet already has that address, then it sends an ICMPv6 Neighbor Advertisement message.
- If X does not receive any NA messages stating that the address has been taken, then X is able to use that address.

Attacks in Duplicate Address Detection phase

An intruder can prevent a Smart Meter from acquiring any auto-configured address, by sending an NA for the corresponding address in every NS message sent out by the Smart Meter. As a result, the Smart Meter will not be able to communicate within the network. Besides, an intruder can block a NA message from an authentic SM. This results in two or more SMs using the same address within a network. As a result of this attack, a legitimate SM can be accused of identity spoofing. Also, more than one assignment of the same address within a network can cause improper functioning during the routing phase.

In order to detect these kinds of attacks, we propose a modified version of the Duplicate Address Detection phase,

- SM X sends an ICMPv6 NS message for the address it wants to acquire.
- On receiving the NS message, every Smart Meter scans its neighbor cache information for that address. If they find the address in their cache, then they send a reply to the X.
- If any Smart Meter on that subnet already has that address, then it sends an ICMPv6 NA message.
- If the X receives neither any NA messages stating that the address has been taken nor receives any messages from its neighbors stating that the address is present in their cache, then X is able to use that address.

If X receives only the NA message from another Smart Meter but no neighborhood information about that address is received, it implies that such an address is not in existence within the subnet and some attacker is trying to prevent X from acquiring that address. If X does not receive any NA message, but its neighbors reply with their cache information stating that the address is present in their neighborhood, then the X concludes that an attacker has intercepted the NA message from the target Smart Meter and has dropped it.

Thus, X is able to use an address only when it neither receives the NA nor any neighborhood cache information from its neighbors.

If the attacker is intelligent enough, it can send both the NA message and also spoof some reply messages from other Smart Meters and change their contents. In that case, SM X will not be able to detect the attack. So, to detect this kind of attack, if a Smart Meter exists with the same address, it not only replies with an NA message but also sends its neighborhood information to X. SM X then sends unicast queries to each of the neighbors found in the reply message to verify the existence of such a Smart Meter. In this way, X can be assured whether he is being duped or whether the particular address is really being used within the subnet. However, since the reply message can also be intercepted by the attacker, it must be broadcast within the network. This will assure the delivery of the reply message to X.

C. Neighbor Discovery

Once the Smart Meter acquires a unique global address, then it can start communication through the DCU. It can also communicate with the other Smart Meters, both in its subnet and in other subnets. Smart Meters on the same subnet can communicate directly with each other without using any router or gateway when a SM has link layer addresses of other neighboring SMs. Thus it is important to store the link layer addresses of the neighboring SMs in the local cache of every SM. Neighbor Discovery facilitates the same.

Normal Procedure for Neighbor Discovery

In order to communicate with a SM B on its own subnet, a Smart Meter A has to perform the following steps,

- First, the SM A sends an ICMPv6 NS message requesting the link-layer address of B.
- If B is present in that subnet, then it replies with an ICMPv6 NA message. SM A knows the MAC address of B from this NA message.
- SM A then creates a neighbor cache entry for B that binds the MAC address of B to its IPv6 address.

Attacks in Neighbor Discovery phase

The attacks of this phase are similar to the attacks of the Duplicate Address Detection phase. Here also an intruder can try to impersonate B, and intercept all packets that are destined to B, or an intruder can block a NA reply from B so that A thinks that B is not present in the network.

III. Proposed IPS to Handle ICMPv6 Threats in Smart Grid

In section 2, we have seen three possible security breaches in Smart Grid for Router Discovery, Duplicate Address Detection and for Neighbor Discovery in sub-sections II.A, II.B and in II.C respectively. The Intrusion Prevention Systems (IPS) against each of these three attacks due to ICMPv6 vulnerabilities have been proposed in the following sub-sections.

A. Intrusion Prevention Mechanism in Router Discovery and Updation phase

In order to prevent these possible security threats, we propose a modified Router Discovery phase as follows,

- First, SM X sends an ICMPv6 RS message to locate a DCU in its local link.
- X receives an ICMPv6 RA message with a 64 bit prefix address for its subnet.
- On receiving the RA message, X extracts the DCU's address from the packet.
- X then broadcasts an ICMPv6 Echo Request message on its subnet.
- Receivers of the ICMPv6 Echo Request message will communicate with their DCU. If a new valid DCU is installed in the subnet, then the other DCUs will have information about the new DCU. If receivers of ICMPv6 Echo Request message receive Router Redirect message (RR) from their current DCU, then they reply with an ICMPv6 Echo Reply message with the address of the new DCU.
- Otherwise, Echo Reply message contains the address of the existing DCUs.
- If the DCU address in the RA message received by SM X matches with a majority of the neighbors' default routers address, then SM X concludes that the DCU is authentic. Consequently, X installs this DCU as its default router in the link, and auto-configures a global unicast address based on the received subnet prefix.
- If the received DCU's address does not match with the address of the default router of the majority of the neighbors, say C, then X concludes that it has been attacked by some adversary and C is the original DCU of that subnet.
- Subsequently, X installs C as its default router in the subnet and auto-configures a global unicast address based on the prefix of C.
- If X does not receive any Echo Reply message within a certain time, then it concludes that it has been blocked by some attacker and sends an SMS alert to the registered mobile number.

Router Updation Phase

DCUs in the Smart Grid network periodically broadcast RA messages to advertise themselves on the subnet. If a Smart Meter receives a RA from a DCU, then they change their existing DCU and register the new DCU as a router in its routing information table.

In this situation an attacker may spoof a RA message and send it to some Smart Meters. On receiving a RA message, Smart Meters then register the attacker as a router. In order to detect this kind of attacks we propose an intrusion prevention mechanism as follows,

- DCUs periodically broadcast RA message.
- On receiving a RA message with new DCU information, every Smart Meter sends a RS message to its existing DCU.

- The existing DCU, on receiving a RS message, checks whether a new DCU with higher priority is available for the subnet.
- If such a DCU exists, it sends a RR message to the SMs with the information of the new DCU. Otherwise, it advertises itself again with a RA message.
- A SM resets its DCU information if and only if it receives a RR message and the DCU information contained within the RR message matches with the previously received RA message. Otherwise, it discards the RA message.

Figure 2 shows a high level view of intrusion detection in Router Discovery and Updation phase, when an attacker spoofs a RA message from DCU and sends it to a Smart Meter X without changing the 64 bit prefix address. In the first half of the figure, an attacker spoofs a RA message and sends it to the newly installed Smart Meter X. In the second half of the figure, an attacker broadcasts a RA message to all the working Smart Meters.

The proposed IPS apparently comes with a boot-strapping limitation. It will not work properly when a new Smart Meter is installed under a new subnet. If Smart Meter X is the first meter in the subnet, then it can't consult with its neighbors to authenticate a legitimate DCU. However, in practice when a new DCU, say K, is to be introduced in a layer just on top of the SMs, some of the SMs under a neighboring DCU will be allocated under K by using RR messages from the current DCU of the respective SMs. The

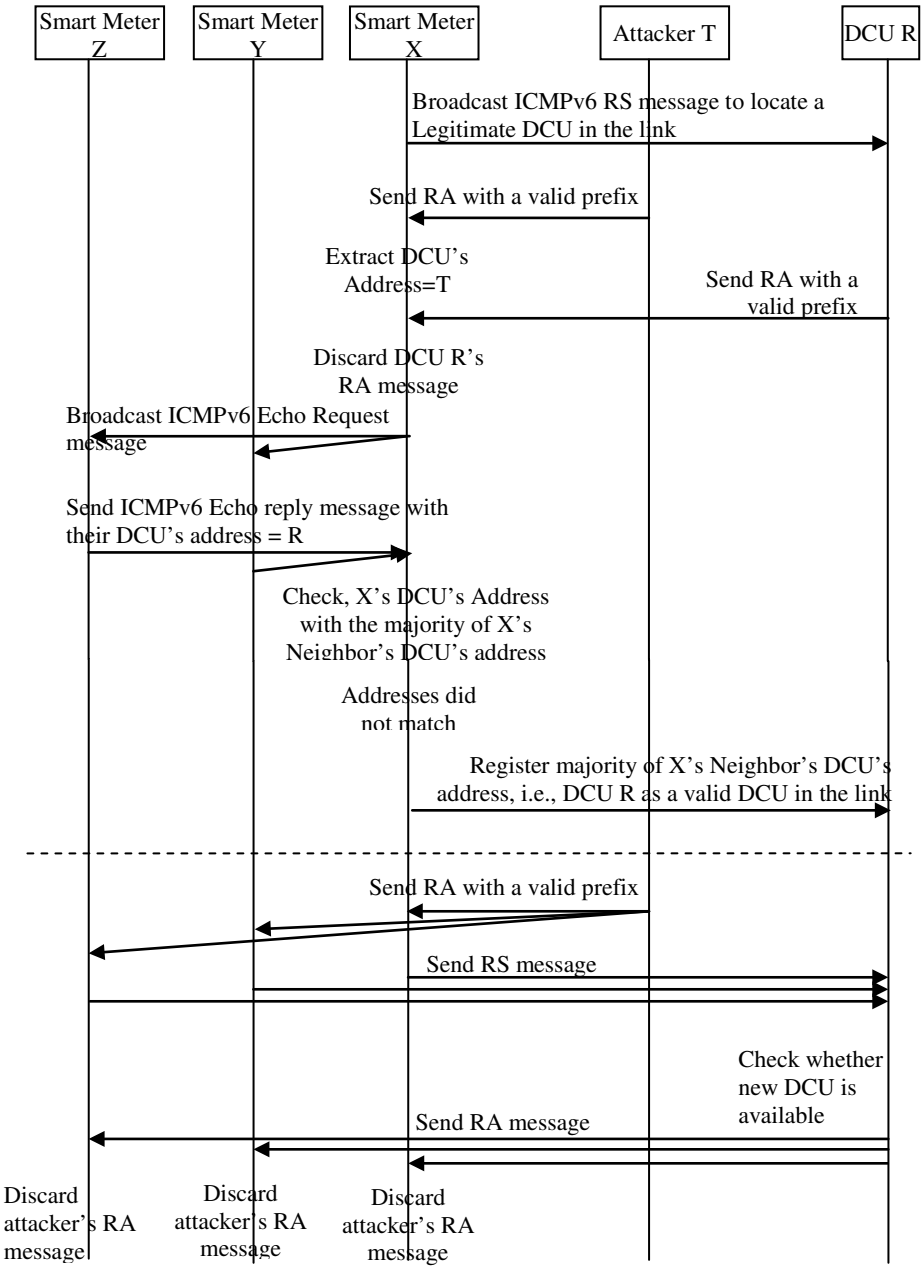


Fig. 2: High level view of Intrusion Prevention in Router Discovery and Updation phase

same is applicable for the entire Smart Grid when a new DCU is to be introduced at any higher level. Thus, the bootstrapping problem as mentioned above will not be an actual bottleneck in the context of smart grid.

B. Intrusion Prevention Mechanism in Duplicate Address Detection

In order to secure Duplicate Address Detection, the following steps are performed,

- SM X sends an ICMPv6 NS message for the address it wants to acquire, say Z.
- If Z already exists in the same subnet, then it broadcasts an ICMPv6 NA message along with the address of all neighbors in its neighbor cache.
- If Z exists, then its one-hop neighbors have Z in their neighborhood cache. These neighbors, on receiving the NS message, reply with a confirmation message.
- X builds the neighbor list of Z from the unicast confirmation messages received from Z's neighbors and verifies it with the neighborhood data sent by the node Z itself.
- If majority of the neighbors confirm the existence of Z, then X concludes that it cannot use Z. Otherwise, X sends unicast queries to those neighbors of Z from which it did not receive any confirmation message.
- Each neighbor N broadcasts Hello message to update its Neighbors. If N finds Z as a neighbor, then it sends a reply confirming existence of Z or remains silent.
- SM X continues sending these queries until either it has a majority decision or all neighbors of Z have been exhaustively queried.
- If X receives both NA message from Z and majority confirmation messages from Z's one-hop neighbors, then it repeats the process with some other auto configured address P. Otherwise, X can use the address Z.

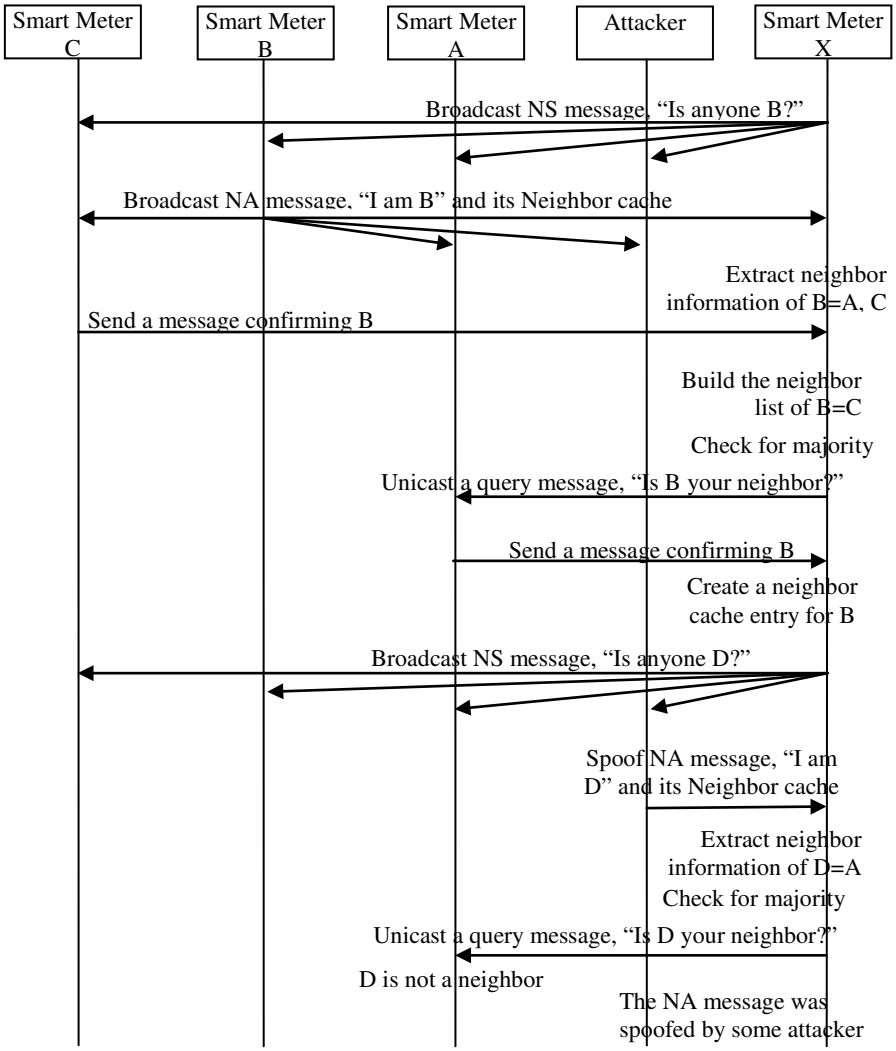


Fig. 3: High level view of Intrusion Prevention in Duplicate Address Detection phase

Figure 3 shows a high level view of intrusion detection in Duplicate Address Detection phase, when X wants to acquire address B. However, in this case, B is already present in the subnet. X verifies the presence of another SM in the subnet, with same address, i.e. B, with the help of B's neighbor list: C, A. Consequently, A wants to acquire

address D. This time an attacker falsely claims himself to be D. X successfully detects this attack.

C. Intrusion Prevention in Neighbor Discovery phase

The detection procedure is quite similar to the Duplicate Address Detection phase as discussed to section III.B. At first, SM X sends an ICMPv6 NS message requesting the link-layer address of Z. On receiving the NS message, every Smart Meter scans its neighbor cache information for that address. If they find the address of Z in their cache, then they send the address of all neighbors of its neighbor cache to X. If node Z is present in that subnet, then it replies with an ICMPv6 NA message and sends the addresses of all neighbors of its neighbor cache. From that NA message, SM X knows the MAC address of Z.

Subsequently, SM X sends unicast queries to each of the neighbors found in the reply message to verify the existence of Z. Every neighbor will broadcast their replies. If X receives a NA messages from Z and majority of reply messages from the neighbors assuring the existence of Z, then SM X creates a neighbor cache entry for Z that binds the MAC address of Z to its IPv6 address. Figure 4 shows a high level view of intrusion detection that may occur during the Neighbor Discovery phase. Here, DCU X wants to communicate with SM Z, but attacker node tries to impersonate Z.

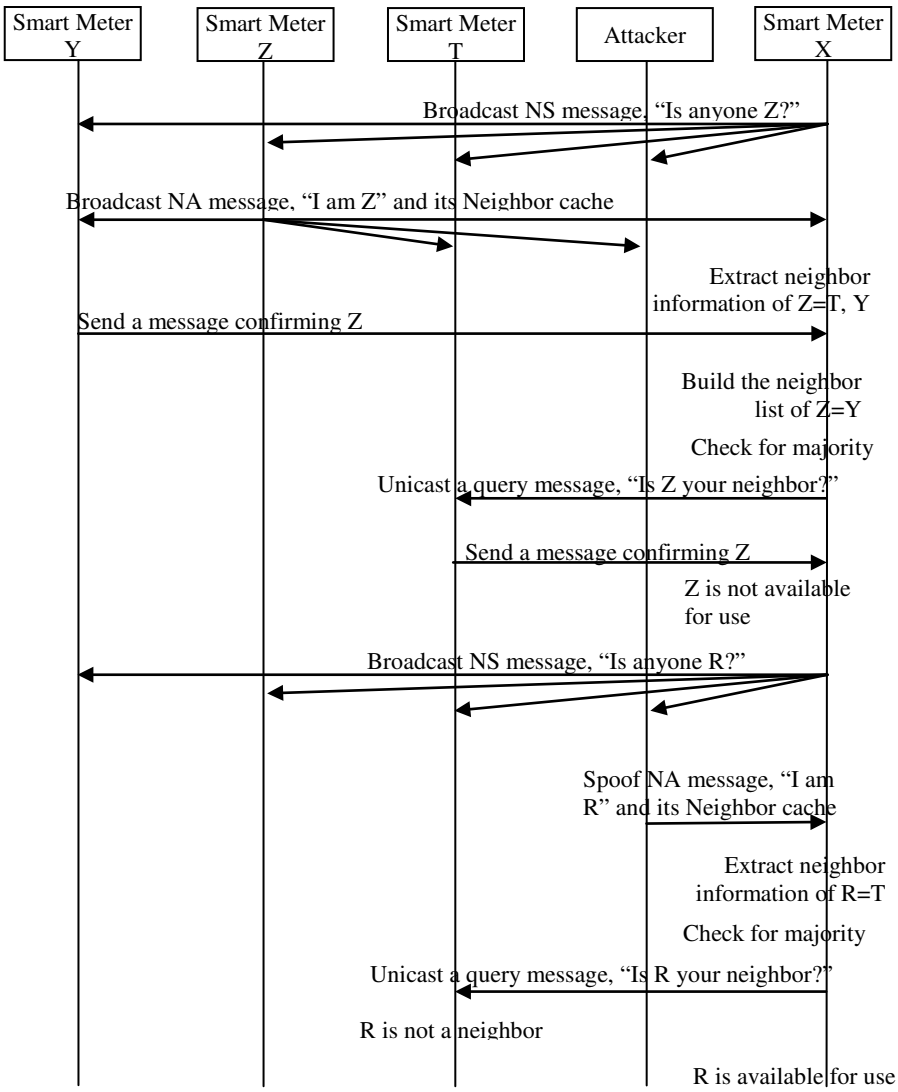


Fig. 4: High level view of Intrusion Detection in Neighbor Discovery phase

IV. Simulation Results

In order to access the performance of ICMPv6 in absence of the proposed IPS and in its presence, we have simulated an environment using Qualnet simulator software. In order to evaluate the performance of the proposed approach, two of the most important performance metrics have been considered. These are false negatives and jitter. *False negative* is measured with respect to both node density and fake router density. Jitter is compared for ICMPv6, with and without our proposed algorithm. The simulation scenario and settings are described in Table I below.

TABLE I.
SIMULATOR PARAMETER SETTINGS

Parameter	Value
Terrain area	1500X1500 m2
Simulation time	100 sec
Mac Layer protocol	DCF of IEEE 802.11b standard
Traffic Model	CBR (Constant Bit Rate)
No. of CBR applications	10 % of the number of nodes
Routing Protocol	AODV
DCU: Smart Meter	1:5

A. False Negative

False Negative occurs when a system cannot detect an attack. False negatives are often a greater threat than false positives. If there wasn't an attack and the system makes a false detection, it can affect the throughput at most. However, if there was an attack and the system is not able to detect it, then it may be disastrous. However, in our proposed IPS, there are no false positives for relatively smaller number of intruders. However, the IPS suffers from false negatives with increasing percentage of malicious nodes. Figure 5 shows that there are no false negative for 2, 4, or 6 malicious nodes out of 50 nodes. The fake router percentage represents the increasing number of fake routers or malicious nodes in a fixed number of nodes. For this experiment we take, 2, 4, 6, 8, 10, 12 fake routers in 50 nodes, with 4, 8, 12, 16, 20, 24 percentages respectively. The false negative increases with increasing number of malicious nodes. Figure 6 shows the effect on false negatives with a linear percentage of malicious nodes, i.e. a fixed percentage of fake routers or malicious nodes in an increasing number of total nodes. We carry out this experiment with 10, 20, 30, 40 and 50 nodes and 20% malicious nodes for each data set. There were no false negatives for 10, 20, 30 and 50 nodes with 20% malicious nodes. The experimental results are in line with reality where any IPS system fails when majority of nodes become compromised.

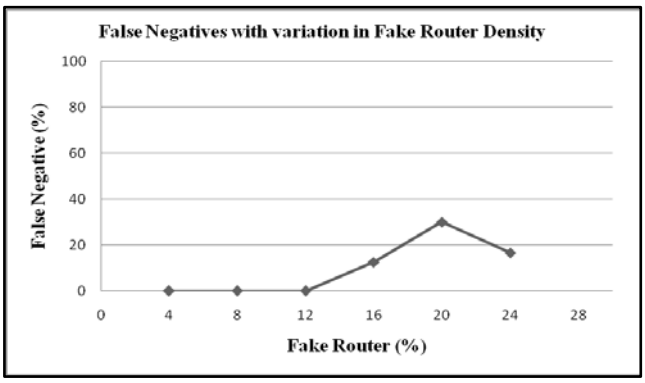


Fig. 5: False Negative vs. Number of Malicious DCUs

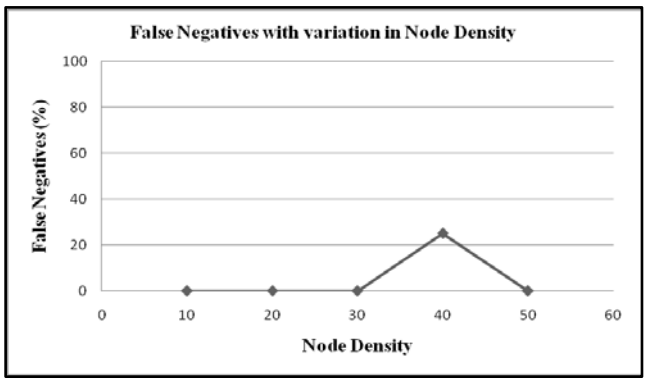


Fig. 6: False Negative vs. Node Density

B. Jitter

Jitter is expressed as an average of the deviation from the network mean latency. We measure both the Jitter for normal ICMPv6 and that with our proposed IPS for ICMPv6. Figure 7 illustrates that the proposed IPS reduces the Jitter.

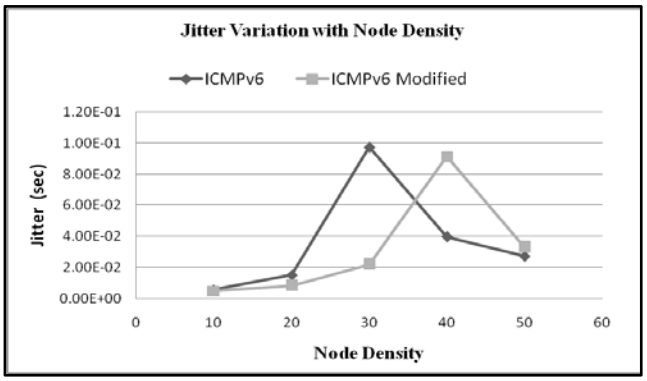


Fig 7: Jitter Vs Node Density for ICMPv6 and Modified ICMPv6

V. CONCLUSION

Integrating IPv6 with Smart Grid is quite natural, as only IPv6 could match the size of Smart Grid network. The large address space, auto configuration of addresses, QoS support technology helps Smart grid to construct a large network with a unique address specified for each and every device, efficient routing, end-to-end security. However, smart grid has very high security demand that needs to be considered before deploying IPv6 towards building Smart Grid. In this paper, the problems of using ICMPv6 in NDP and the possible effects of these problems on Smart Grid are considered. Three main functions of NDP: Router Discovery, Duplicate Address Detection and Neighbor Discovery are discussed with respect to Smart Grid environment. We first consider the normal procedure for executing each phase, and then discuss the possible attacks. Finally a prevention procedure is given to secure the system. The proposed work considers multiple security breaches on Smart Grid and provides an IPS to prevent these attacks in Router Discovery and Updation phase as well as in Neighbor Discovery and DAD phase. This, in turn, helps preventing several attacks on ICMPv6 protocol, like DoS, man-in-the-middle attack, spoofing attacks efficiently. It is also light weight and does not burden the system with unnecessary packet overhead.

A possible bootstrap problem of the proposed IPS system

has been considered and found insignificant in section III.A. The proposed methodology builds the foundation for several meaningful extensions in future. In future, we want to extend this work to detect collaborative attacks on smart grid.

ACKNOWLEDGEMENT

This work is a part of the Ph.D. work of Manali Chakraborty, a Senior Research Fellow of Council of Scientific & Industrial Research (CSIR), Government of India. We would like to acknowledge CSIR, for providing the support required for carrying out the research work. We would also like to acknowledge Jeeyan Sanyal, a student of the Masters program with Department of Computer Science and Engineering, University of Calcutta, for his contribution in the simulation process.

The work is partially supported by the PRIN "Security Horizons" project.

REFERENCES

[1] S. M. Amin, B. F. Wollenberg, " Toward a smart grid: power delivery for the 21st century." IEEE Power and Energy Mag. Vol.3. No.5. Sept.-Oct (2005) 34-41. DOI: 10.1109/MPAE.2005.1507024

[2] "Study of Security Attributes of Smart Grid Systems – Current Cyber Security Issues." Department of Energy Office of Electricity Delivery and Energy Reliability, National SCADA Test Bed, April (2009).

[3] M. R. Asghar, D. Miorandi, " A Holistic View of Security and Privacy Issues in Smart Grids." SmartGridSec, Lecture Notes in Computer Science. Vol. 7823. (2013) 58-71. DOI: 10.1007/978-3-642-38030-3_4

[4] T. Narten et al, "RFC 4861-Neighbor Discovery protocol for IPv6", september, 2007.

[5] M. A. Saad, S. Ramadass, S. Manickam, "A Study on Detecting ICMPv6 Flooding Attack based on IDS", Australian Journal of Basic and Applied Sciences, Vol.7. (2013) 175-181.

[6] S. Hogg, E. Vyncke, "IPv6 Security". 1st Edition, Cisco Press, Dec. ISBN: 978-1587055942 (2008).

[7] "The smart grid vision for India's power sector." White Paper by United States Agency for International Development, USAID India. March (2010).

[8] T. Zseby, " Is IPv6 Ready for the Smart Grid?" CYBERSECURITY '12 Proceedings of the 2012 International Conference on Cyber Security. (2012) 157-164. DOI: 10.1109/CyberSecurity.2012.27

[9] J. Liu, Y. Xiao, S. Li, W. Liang, C. L. P Chen, " Cyber Security and Privacy Issues in Smart Grids". IEEE Communications Surveys & Tutorials. Vol. 14, No. 4, 4th Quarter. (2012). DOI: 10.1109/SURV.2011.122111.00145

[10] T. Baumeister," Literature review on smart grid cyber security. "Tech. rep., Collaborative Software Development Laboratory, Department of Information and Computer Sciences, University of Hawaii, December (2010).

[11] S. Groat, M. Dunlop,W. Urbanksi,R. Marchany, J. Tront, "Using an IPv6 Moving Target Defense to Protect the Smart Grid." Innovative Smart Grid Technologies (ISGT), IEEE PES, (2012) 1-7. DOI: 10.1109/ISGT.2012.6175633

[12] C. Y. Cheng, C. C. Chuang, R. I. Chang, " Three-dimensional Location-based IPv6 Addressing for Wireless Sensor Networks in Smart Grid". 26th IEEE International Conference on Advanced Information Networking and Applications. (2012) 824-831. DOI: 10.1109/AINA.2012.42

[13] C. Y. Cheng, C. C. Chuang, R. I. Chang, " Lightweight Spatial IP address Configuration for IPv6-based Wireless Sensor Networks in Smart Grid." SENSORS 2012, IEEE. (2012) 1-4. DOI: 10.1109/ICSENS.2012.6411205

[14] A. P. Castellani, G. Ministeri, M. Rotoloni,L. Vangelista, M. Zorzi, " Interoperable and globally interconnected Smart Grid using IPv6 and 6LoWPAN." 3rd IEEE International Workshop on SmArt COmmunications in NEtwork Technologies, 10-15th June (2012) 6473-6478. DOI: 10.1109/ICC.2012.6364813

[15] M. Kim, "A Survey on Guaranteeing Availability in Smart Grid Communications". Advanced Communication Technology (ICACT). (2012) 314-317.

[16] Z. A. Baig, S. C. Adeniye, "A trust-based mechanism for protecting IPv6 networks against stateless address auto-configuration attacks". 17th IEEE International Conference on Networks. Singapore (2011). 171-176. DOI: 10.1109/ICON.2011.6168470

[17] R. Berthier,W. H. Sanders, H. Khurana, "Intrusion detection for advanced metering infrastructures: Requirements and architectural directions" In First IEEE International Conference on Smart Grid Communications (SmartGridComm), Oct. (2010). 350–355. DOI: 10.1109/SMARTGRID.2010.5622068

[18] P. Jokar, H. Nicanfar, V. Leung, "Specification-based intrusion detection for home area networks in smart grids". In IEEE International Conference on Smart Grid Communications (SmartGridComm), Oct. (2011).208–213. DOI: 10.1109/SmartGridComm.2011.6102320

[19] R. Berthier, W. H. Sanders, "Specification-based intrusion detection for advanced metering infrastructures." In IEEE 17th Pacific Rim International Symposium on Dependable Computing (PRDC), Dec. (2011). 184–193. DOI: 10.1109/PRDC.2011.30

Software Systems Development & Applications

SSD&A is a FedCSIS conference area aiming at integrating and creating synergy between FedCSIS events that thematically subscribe to the discipline of software engineering. The SSD&A area emphasizes the issues relevant to developing and maintaining software systems that behave reliably, efficiently and effectively. This area investigates both established traditional approaches and modern emerging approaches to large software production and evolution.

Events that constitute SSD&A are:

- ATSE'14 - 5th International Workshop Automating Test Case Design, Selection and Evaluation
- MDASD'14 - 3rd Workshop on Model Driven Approaches in System Development

5th International Workshop Automating Test Case Design, Selection and Evaluation

TRENDS such as globalisation, standardisation and shorter lifecycles place great demands on the flexibility of the software industry. In order to compete and cooperate on an international scale, a constantly decreasing time to market and an increasing level of quality are essential. Software and systems testing is at the moment the most important and mostly used quality assurance technique applied in industry. However, the complexity of software systems and hence of their development is increasing. Systems get bigger, connect large amounts of components that interact in many different ways on the Future Internet, and have constantly changing and different types of requirements (functionality, dependability, real-time, etc.). Consequently, the development of cost-effective and high-quality systems opens new challenges that cannot be faced only with traditional testing approaches. New techniques for systematization and automation of testing are required.

Even though many test automation tools are currently available to aid test planning and control as well as test case execution and monitoring, all these tools share a similar passive philosophy towards test case design, selection of test data and test evaluation. They leave these crucial, time-consuming and demanding activities to the human tester. This is not without reason; test case design and test evaluation are difficult to automate with the techniques available in current industrial practice. The domain of possible inputs (potential test cases), even for a trivial program, is typically too large to be exhaustively explored. Consequently, one of the major challenges associated with test case design is the selection of test cases that are effective at finding flaws without requiring an excessive number of tests to be carried out. This is the problem which this workshop wants to attack.

This workshop will provide researchers and practitioners a forum for exchanging ideas, experiences, understanding of the problems, visions for the future, and promising solutions to the problems in automated test case generation, selection and evaluation. The workshop will also provide a platform for researchers and developers of testing tools to work together to identify the problems in the theory and practice of software test automation and to set an agenda and lay the foundation for future development.

Topics

Topics include (but are not limited to):

- techniques and tools for automating test case design:
  - model-based,
  - combinatorial.based,
  - optimization-based,
  - etc.
- Evaluation of testing techniques and tools on real systems, not only toy problems.
- Benchmarks for evaluating software testing techniques

Event Chairs

Eldh, Sigrid, Ericsson & Karlstad University, Sweden
Prasetya, Wishnu, University of Utrecht, Netherlands
Vos, Tanja, Universidad Politecnica de Valencia, Spain

Program Committee

Afzal, Wasif, Mälardalens Högskola
Aho, Pekka, VTT
Bagnato, Alessandra, Softeam, France
Condori, Nelly, Universidad Politecnica de Valencia, Spain
Datar, Advaita
Escalona, Maria Jose, Universidad de Sevilla, Spain
Harman, Mark
Jia, Yue
Marchetto, Alessandro, Centro Ricerche Fiat - CRF, Italy
Marin, Beatriz, Universidad Diego Portales, Chile
Memon, Atif, University of Maryland, United States
Noack, Thomas
Polo, Macario, Uiversidad de Castilla la Macha, Spain
Roper, Marc, University of Strathclyde, United Kingdom
Shehory, Onn, IBM, Israel
Sundmark, Daniel, Swedish Insititute of Computer Science
Tonella, Paolo, Fondazione Bruno Kessler,, Italy
Tuya, Javier, Universidad de Oviedo, Spain

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1553–1558
DOI: 10.15439/2014F93
ACSIS, Vol. 2

Tool for Automatic Testing of Web Services

Ilona Bluemke, Michał Kurek, Małgorzata Purwin
Institute of Computer Science Warsaw University of Technology
Nowowiejska 15/19, 00-665 Warsaw, Poland
Email: I.Bluemke@ii.pw.edu.pl.

***Abstract*— The Web Services technology has received a significant amount of attention in recent years because it allows to easily utilize and integrate existing software applications to create new business services. With the increase of interest and popularity of Web Services, web applications are developed. This way of software development causes new issues for Web Service testing to ensure the quality of service that are published. This paper describes the tool (named WSDLTest) for automatic testing of web services. The tool can be used for testing of web services for which WSDL 1.1 or WSDL 2.0 document are available. WSDLTest parses the WSDL document, and based on, it tests the web service by sending automatically generated messages. Some examples of the usage of our tool are given.**

I. Introduction

Web services are very important in providing software in the World Wide Web. They have emerged as the next generation of Web-based technology for exchanging information. They are modular, self - descriptive, self - contained applications that are based on open standard. A Web Service is a kind of Internet application based on SOAP [1] and XML [2] technology. The W3C [3] (World Wide Web Consortium) defines web service as "a software system designed to support interoperable machine-to-machine interaction over a network". Web Service relies on a family of protocols to describe, deliver, and interact with each other, such as the Web Service Description Language (WSDL) [4], the Universal Description, Discovery and Integration (UDDI) [5] protocol or the Web Service Inspection Language (WSIL) [6], and the Simple Object Access Protocol (SOAP) [1]. WSDL, UDDI, WSIL, and SOAP are all based on XML [2]. Due to the nature of standards-based architecture and XML-based messaging, Web Service is vendor-neutral, language-agnostic, and platform-independent. Therefore, developers of a Web Service are not able to assume which type of clients will use the Web Service and the developers at the client side, are not aware of which language and platform are used at the server side. A Web Service and its clients can be developed in totally different programming languages and on different platforms.

Web service can be described by a Web Service Description Language [4] document. WSDL was established as a standard by the W3C [3]. The WSDL document describes web service in terms of its interfaces with operations, types of data the web service takes or returns. Other systems interact with the Web service in a manner prescribed by its description using SOAP [1] messages.

Testing effort is often a major cost factor during software development, which sometimes consumes more than 50 % of the overall development effort [7] . Consequently, one major goal is often to reduce the testing effort e.g. by providing a tool facilitating the testing. With services the difficulty of testing increases because of the changes that this architectural style induces on both the system and the software business/organization models.

The purpose of our research was to determine the possibility and methods of testing web service based on its WSDL document. An application, named WSDLTest, has been designed and implemented at the Institute of Computer Science Warsaw University of Technology. This application can test web services for which WSDL 1.1 or WSDL 2.0 document are available. WSDLTest parses the WSDL document, and based on it, tests the web service by sending automatically generated messages.

The paper is organized as follows. Section II identifies key features of web services testing, also WSDL is briefly presented. In Section III some testing tools, using WSDL documents as a basis for testing, are discussed. Section IV focuses on the architecture of our testing tool (WSDLTest). Section V presents some results of testing a simple web service with WSDLTest. Finally, Section VI concludes the paper, highlighting some issues and future research directions.

II. Web services Testing

Testing services and service-centric systems poses new challenges to traditional testing approaches. Testing challenges derive primarily from the dynamic nature of web services and the clear separation of roles between the users, the providers, the owners, and the developers of a service. Automated service discovery and binding mean that the

This work was not supported by any organization

complete configuration of a system is known only at execution time, and this cause integration testing to be difficult. Canfora and Di Penta in [8] indicated unique features of services that add complexity to the testing.

Service testing can be performed at different levels namely:

- unit testing of atomic services and service compositions,
- integration/interoperability testing,
- regression testing,
- testing of non-functional properties.

Services testing is a recent area of investigation, numerous contributions have been presented in the literature, primarily in the areas of unit testing of services and orchestrations, integration testing, regression testing, and testing of non-functional properties. The literature also reports several investigations to improve the testability of services and service-centric systems. Canfora and Di Penta made SOA testing survey [8]. They also pointed out several problems which remain open and need additional research work e.g.: improving testability, combining testing and run-time verification, validating fully decentralized systems.

Comprehensive and excellent surveys on Web services testing can also be found in [9-12]. Several Web Services testing approaches were developed to address these new challenges, the survey can be found in [9]. Based on the surveyed Web Services testing approaches found in the literature, the existing approaches were divided by Ladam [9] into four classes:

1. WSDL-based test case generation,
2. mutation-based test case generation (e.g. [13, 14]),
3. test modeling (e.g. [15]),
4. XML-based approaches (e.g. [16]).

Hanna and Munro [17] proposed a framework that can be used to test the robustness quality attribute of a Web Service. This framework is based on analyzing the Web Service Description Language (WSDL) document of Web Services to identify what faults could affect the robustness attribute and then test cases were designed to detect those faults.

Bai and Dong [18] also are using the WSDL document to generate tests in a testing framework that includes test case generation, test controller, test agents and test evaluator.

Often the service is not able to satisfy the needs of a user but it is possible to combine existing services together in order to fulfill this need. The process of combining these services is called Web Service Composition (WSC). However much researches have been focused on the discovery, selection and composition of Web services, the testing of Web service composition is still immature [19]. Bucchiarone et al. [20] and Zakaria et al. [21] provide surveys focusing in on Web service composition. Rusli et al. provided [12] an interesting overview and evaluation of current approaches to WSC testing.

We concentrated on the WSDL-based approach to unit testing because it enables the automatic generation of tests.

A. WSDL

WSDL [4] - Web Services Description Language, is an abstract, XML-based language, which specifies location and functionalities offered by a web service. It contains specific information about the web service, for example needed parameters, returned data. XML schema is used for the presentation of WSDL description containing the messages send and received from the service. To communicate with the web service SOAP messages are exchanged and they are described by WSDL as operations. WSDL describes the format for interfaces, it can also describe interactivity of given service. WSDL description only shows the possible but not required interactions. The current, recommended by W3C [3], version of WSDL is 2.0. The older version - WSDL 1.1 is still quite common and in some software only this version is supported. Sections `<messages>` and `<portType>` were combined to create new section `<interface>` (in version 2.0)

WSDL 2.0 contains following sections:

- "*description*" - a container, inside which the remaining sections are located,
- "*types*" describes the data types send and received by a web service,
- "*interface*" describes the abstract functionality the web service provides (what messages it sends and receives, possible fault messages),
- "*binding*" provides information how to access the service,
- "*service*" provides information where to access the service.

There are also two optional sections in WSDL document:

- "*documentation*" provides textual description of a web service,
- "*import*" is used to import other XML schemas.

B. Usage of WSDL in testing

WSDL document is a description of functionalities provided by a web service. Inside this document, all "methods" and data types can be found. It is divided into several sections directly connected with each other. The section "types" contain data types, which are input and output parameters for operations. Operations are described in "interface" section. For each interface, there is a binding with the message protocol. The "binding" section of WSDL document provides knowledge about protocols bonded to given interface. "Service" sections provide information where to find endpoint for each binding. All those sections allow to determine the input and the output, as well as how and where to access the web service. Nonetheless, there is no information about what exactly the web service produces as output from input. To use effectively WSDL in testing software, following elements are needed:

- WSDL parser tool, which reads the WSDL document and extracts information.
- Oracle mechanism determining if the results of the test are correct.
- SOAP messages generator.

The process of testing based on WSDL document includes several steps. Firstly the WSDL document is parsed into

addressable memory objects. Information from parsing can be used to generate SOAP requests. Such message may contain either randomly generated parameters based on data type taken from WSDL document, or can be filled by a tester. Such approach to web service testing is applicable only when communication is bidirectional – request and response from the web service. When SOAP request is send, after a while the web service should send back response with results of its operation. These results may be later compared with oracle to determine if the web service is working properly.

III. TESTING TOOLS

There are several testing tools available on the market, which can be used for testing web services. The most recognizable tools are:

1. HP QuickTest Professional,
2. Parasoft SOAtest,
3. SOAPSonar,
4. SoapUI.

Some of these tools offer the validation of WSDL document provided by WS-I [22].

HP QuickTest Professional [23] is an automatic testing suite. It accepts both 1.1 and 2.0 versions of WSDL. The application performs automated tests on a variety of software and in many environments. It has graphical user interface, which allows performing regression and functional testing. The main testing algorithm identifies objects to be tested and performs testing of interface operations. Important function of HP QuickTest Professional is the ability to perform the validation of WSDL using the WS-I [22] tool. In this application the modification, insertion, and removal of test parts are easy. HP QuickTest Professional has a Web Service Testing Wizard Click that can be used for the definition of different options e.g. a checkpoint can be created or some actions may be selected. Everything defined in the wizard is later converted and inserted into the test. This tool has functions allowing the customization of reports with user-defined images and screenshots. Client performance-related errors (e.g. memory leakage) can be included in the standard reports with a direct link from the report to the test script. HP QuickTest Professional automates the design of tests and test cases.

Parasoft SOAtest [24] is used for testing applications build in Service-oriented Architecture. It can automate application testing, message/protocol testing, cloud testing and security testing. Its graphical user interface is Eclipse based (uses the same GUI library). It allows creating tests, defining the behavior of the tests and configuring specific tests (e.g. queue managers or database connections). SOAtest primary function is to test Web services. It automatically generates tests from key market platforms such as WSDL [4], UDDI [6], WSIL [7], XML [2], BPEL [25], HTTP traffic and other. SOAtest can validate WSDL documents and emulate the client or the server. Tests can be grouped to be executed in a sequence. Failed tests are highlighted. Parameters for requests can be entered by the user or can be read from a file. SOAtest can trace and visualize how messages and events flow through ESBs (Enterprise service bus) [26], message brokers, applications and databases, while tests are executed. Such tracing allows the interpretation of problems.

SOAPSonar [27] is testing and analysis tool specifically designed for Web Services. It can perform functional regression tests, performance tests; generate compliance reports, vulnerability checks and identity tests. It has clear graphical user interface. Tests are created via drag and drop selection. Application has also WSDL Region WS-I validation [22]. It accepts WSDL 1.1 and WSDL 2.0 documents. The professional edition of SOAPSonar has test flow management options and can create WSDL requests - response chain or data driven test for the exchange of messages. Functional tests use load testing for the performance monitoring. Security penetration tests are performed at the message layer. All versions of SOAPSonar use as test inputs and response analysis data from e.g. database tables or Microsoft®Excel files. Application can automatically change variables in message headers, message body, tasks and can change variables. SOAPSonar has also vulnerability mode, which allows associating each test request with a set of attack. It can parse the WSDL documents and generate a list of the operations available on interface described by it. It can be used to send SOAP request messages to the target Web service and capture the response.

SoapUI [28] is an open source web service testing application for service-oriented architectures. SoapUI has clear and easy to use graphical user interface, which allows for quick creation of any test scenarios. It requires WSDL files to generate tests, messages, validations and MockServices. SoapUI supports only 1.1 version of WSDL and it can easily create functional, regression, compliance, and load tests. SoapUI provides code free test environment, all tests are created by drag and drop actions. SoapUI Pro Test Debugging is an option, which allows following the flow of test, variables, properties requests and context. Data for data-driven tests can be provided from external editors such as Microsoft ® Excel. SoapUI MockServices is a feature, which allows mimicking Web Services before they are implemented. SoapUI uses WS-I profile [22] for 1.1 version of WSDL.

IV. WSDLTEST

WSDLTest is a simple application supporting unit testing of web services which are described in WSDL. In the Fig. 1 the high-level structure of WSDLTest is presented. External connections and internal interfaces are also shown. The frame at the top of figure, which consists of WSDL document and web service itself, represents remote location

that is accessed in different manner by WSDLTest application.

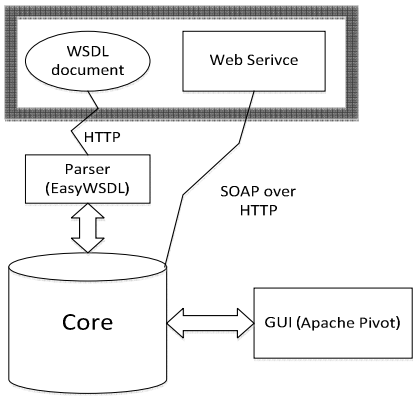


Fig. 1 Structure of WSDLTest

WSDL document is accessed via parser using HTTP connection, whereas web service is communicated thru SOAP messages transported over the HTTP connection. Core module consists of classes that support main functionality of application. WSDLTest is divided logically into Core, Parser and GUI. As a parser we used EasyWSDL Toolbox [29]. It can be used to parse WSDL 1.1 and WSDL 2.0 descriptions and transforms them in a unified object model (based on the WSDL 2.0 entities). Library is written in Java language and therefore can be used within java compatible application only. Moreover, WSDL parser can additionally export information from application object to an editable WSDL document. Other implementation details can be found in [30].

In the Fig. 2 main functions of core module are presented. Based on information from the parsed WSDL document interface bonded to SOAP [1] connection is found. From this interface all operations are extracted to determine what kind of parameters operations require and how many of them. GUI creates panel for input values of parameters or input of regular expression. WSDLTest generates all required parameters and performs tests.

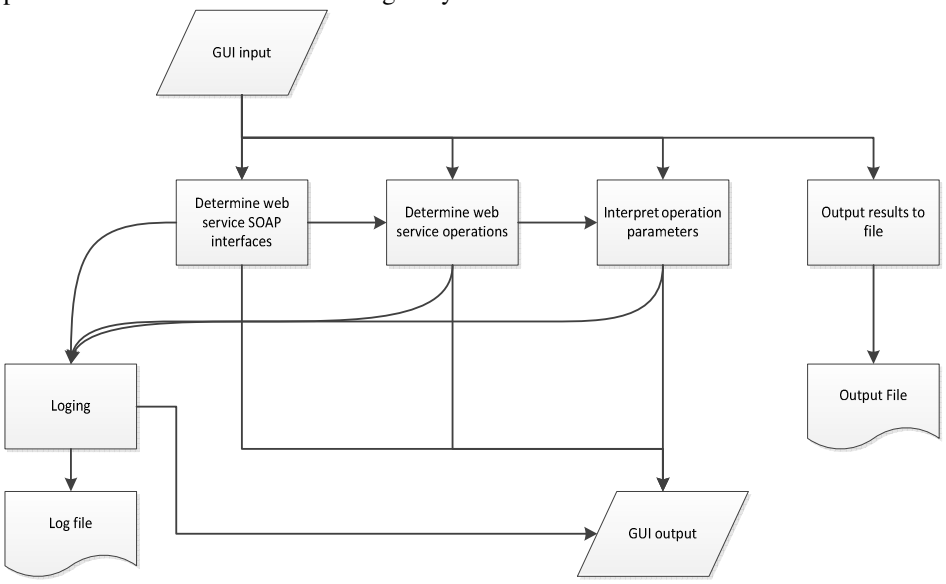


Fig. 2 Activities in Core module of WSDLTest

```
<s:element name="ValidateCardNumber">
        <s:complexType>
               <s:sequence>
                      <s:element minOccurs="0" maxOccurs="1"
                             name="cardType"
                             type="s:string"/>
                      <s:element minOccurs="0" maxOccurs="1"
                             name="cardNumber"
                             type="s:string"/>
               </s:sequence>
        </s:complexType>
</s:element>
<s:element name="ValidateCardNumberResponse">
        <s:complexType>
               <s:sequence>
```

```
14                          <s:element minOccurs="0" maxOccurs="1"
15                                 name="ValidateCardNumberResult"
16                                 type="s:string"/>
17                    </s:sequence>
18             </s:complexType>
19    </s:element>
20    ...
21    <wsdl:portType name="CCCheckerSoap">
22             <wsdl:operation name="ValidateCardNumber">
23                    <wsdl:documentation xmlns:wsdl=
24                           "http://schemas.xmlsoap.org/wsdl/">
25                           Please enter card type as VISA or
26                                 MASTERCARD or DINERS or AMEX
27                    </wsdl:documentation>
28                    <wsdl:input message=
29                           "tns:ValidateCardNumberSoapIn"/>
30             </wsdl:operation>
31    </wsdl:portType>
```

Fig. 4 WSDL document for web service ValidateCardNumbert

A. Testing modes

WSDLTest has three testing modes available for the user (shown in Fig. 3):

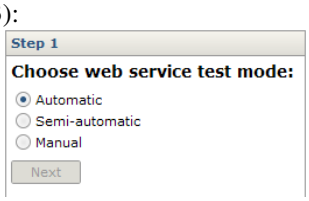


Fig. 3 Modes of operation of WSDLTest

1- *Automatic* – all tests are performed automatically. User only specifies the number of tests to be performed. All possible operations (functions of web service) defined in WSDL document are tested. For each operation random parameters are generated based on theirs defined in XML types.

2- *Semi-automatic* – parameters of operations are provided as regular expressions. For these expressions values of parameters, used in tests, are generated. User can specify number of tests and operations to be tested.

3- *Manual* – the simplest of modes, parameters of operations are given explicitly by the user.

In semi-automatic and automatic mode the oracle, containing expected test results, can be defined.

V. Example

Web service Validate Card Number [31] is a simple web service with only one method named `ValidateCardNumber`, used for credit card number validation. In Fig 4 the definition of operation `ValidateCardNumber` [lines 22-30] is given, input parameter of operation `ValidateCardNumber` [lines 1-10] and output of operation parameter `ValidateCardNumberResponse` [lines 11-19].

This service was tested using semiautomatic mode. In first test the parameter `cardNumber` was inputted as regular expression which matches any random sequence of digits of the length equal 16 (typical length of credit card number). Input window is presented in Fig.5.

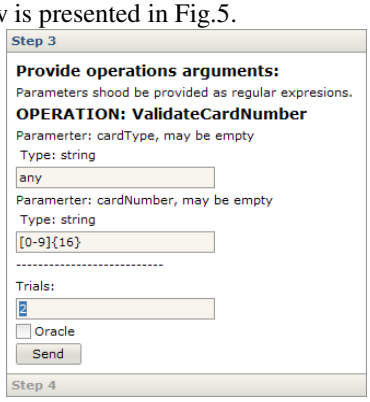


Fig. 5 Input window of WSDLTest – setting parameter values

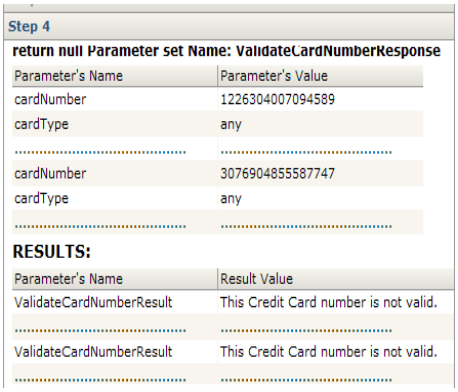


Fig. 6 Result of tests for parameters given as digits

Results, presented in Fig. 6, indicate whether the card number generated by WSDLTest is valid or not. The parameter `cardNumber` is defined as string (line 7 in Fig. 4.), so random string consisting of letters was used in next test. Credit card number can consist only of numbers, so error should be reported. WSDLTest was able to find the errort inside returned SOAP message. Results are shown in Fig.7.

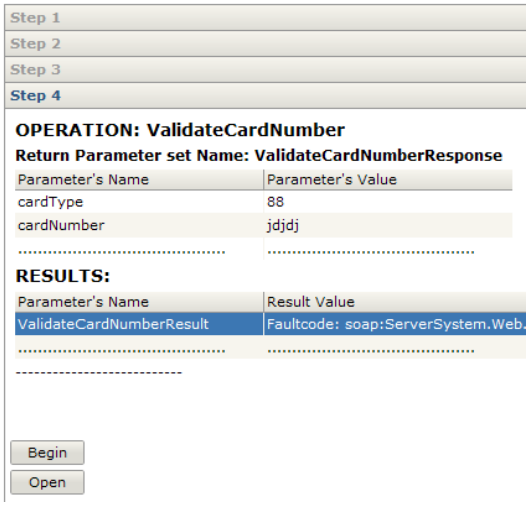


Fig. 7 Result of tests for parameters given as random string

VI. Conclusion

Web service can be specified in a WSDL [4] document. This document may be used as a basis in the automatic testing of the web service. We developed the application, called WSDLTest, which uses the WSDL document to prepare unit tests of web service. WSDLTest allows user, with basic technical knowledge, to test a web service. Testing can be performed in fully automatic mode for theoretically infinite number of trials with different parameters. The automated process of testing includes the generation of parameter values, the comparison of results with oracle and a basic validity check of the WSDL document. Tester can also provide parameter values and oracle in form of regular expression. Such form is more general than the explicit one. Regular expressions used in oracle are used to generate expected values of results, which are compared with actual results of tests.

WSDLTest can test a variety of web services described by WSDL. In this paper we presented simple example (section V). Even on this simple example it can be easily noticed that WSDLTest can be very useful in testing web services. Our tool does not have all functionalities available in other tools presented in section III, e.g. is not able to do load testing. In current version, WSDLTest is not able to read parameters from external files or mimicking web service, such functions are usually available in commercial tools. It might have some problems in testing web services with very complex data. These drawbacks can be eliminated in future versions.

WSDLTest has also an advantage over other similar tools. It can generate values of parameters from regular expressions; such functionality is not common in other tools.

References

All URL in references were valid in May 2014.

[1] SOAP: Simple Object Access Protocol, http://www.w3.org/TR/soap/.
[2] XML – Extensible Markup Language, http://www.w3.org/XML/.
[3] W3C Official Website:, http://www.w3.org/.
[4] WSDL: http://www.w3.org/TR/wsdl.
[5] UDDI: http://uddi.xml.org/.
[6] WSIL www.ibm.com/developerworks/webservices/library/ws-wsilspec.html
[7] F. Elberzhager, A. Rosbach, R. Eschbach, J. Münch, "Reducing Test Effort: A Systematic Mapping Study on Existing Approaches", *Information and Software Technology*, vol. 54, no. 10, October 2012, pp. 1092-1106, DOI=10.1016/j.infsof.2012.04.007.
[8] G. Canfora, M. Di Penta, M. 2009." Service-Oriented Architectures Testing: A Survey". In *Software Engineering*, De Lucia, A. and Ferrucci, F., Ed. Springer, 78-105. DOI= 10.1007/978-3-540-95888-8_4.
[9] M. I. Ladan, "Web Services Testing Approaches: A Survey and a Classification", in F. Zavoral et al. (Eds.): NDT 2010, Part II, CCIS 88, pp. 70–79, 2010. Springer-Verlag Berlin Heidelberg 2010
[10] M. Bozkurt, M. Harman, Y.Hassoun, Y. *Testing Web Services: A Survey*. Technical Report. King's College London, 2010.
[11] A. Metzger, S. Benbernou, M. Carro, M. Driss, G. Kecskemeti, R. Kazhamiakin, K. Krytikos, A. Mocci, A., E. Di Nitto, B. Wetzstein, F. Silvestri, 2010. "Analytical Quality Assurance". In *Service Research Challenges and Solutions for the Future Internet*, Papazoglou, M., Pohl, K., Parkin, M. and Metzger, A., Ed. Springer, 209-270. DOI= 10.1007/978-3-642-17599-2_7
[12] H. M. Rusli, M. Puteg, S. Ibrahim, and S. G. H. Tabatabaei, "A comparative evaluation of state-of-the-art web service composition testing approaches," in *International Workshop on Automation of Software Test*, 2011, pp. 29-35
[13] Reda, S., Nashat, M.: Testing Web Services. IEEE Software (2005)
[14] Andre, L., Silvia Regina, V.: Mutation Based Testing of Web Services. IEEE Software (2009)
[15] Feudjio, A.-G.V., Schieferdecker, I.: Availability Testing for Web Services, ISSN 0085- 7130 Telektronikk 1 (2009)
[16] Tsai, W.T., Paul, R., Song, W., Cao, Z.: An XML-Based Framework for Web Services Testing. IEEE Software (2002)
[17] S. Hanna, M. Munro: "An Approach for WSDL-Based Automated Robustness Testing of Web Services". *Information Systems Development Challenges in Practice, Theory, and Education* 2 (2008)
[18] X. Bai, W. Dong: 'WSDL-Based Automatic Test Case Generation for Web Services Testing". *IEEE Software* (2005)
[19] G. Canfora, M. Di Penta: "Testing services and service-centric systems: challenges and opportunities". *IT Professional*. 8, 2 (March-April 2006), 10-17. DOI= 10.1109/MITP.2006.51.
[20] A. Bucchiarone, H. Melgratti, F. Severoni, F. 2007. "Testing service composition". In *Proceedings of the 8th Argentine Symposium on Software Engineering* (Mar del Plata, Argentina, August 29-31, 2007). ASSE 2007.
[21] Zakaria, Z., Atan, R., Ghani, A. A. A., Sani, N.F.M. 2009. „Unit testing approaches for BPEL: A systematic review" In *Proceedings of the 2009 Asia-Pacific Software Engineering Conference* (Penang, Malaysia, December 1-3, 2009). APSEC 2009. IEEE, 316-322. DOI= 10.1109/APSEC.2009.7
[22] WSI: http://www.oasis-ws-i.org
[23] HP QuickTest Professional: http://www8.hp.com/us/en/software-solutions/unified-functional-testing-automation/.
[24] Parasoft: http://www.parasoft.com/soatest .
[25] BPEL: http://docs.oasis-open.org/wsbpel/2.0/wsbpel-v2.0.pdf
[26] ESB: http://en.wikipedia.org/wiki/Enterprise_service_bus.
[27] SOAPSonar: http://www.crosschecknet.com/products/soapsonar.php.
[28] SoapUI: http://www.soapui.org/.
[29] Easy WSDL Toolbox, http://easywsdl.ow2.org/.
[30] M. Kurek, M. Purwin: *Automatic Testing of Web Services Based on WSDL Document*, Bachelor Thesis, Institute of Computer Science, Warsaw University of Technology, 2014.
[31] WSDL, "Credit Card", http://www.webservicex.net/CreditCard.asmx?WSDL.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1559–1568
DOI: 10.15439/2014F73
ACSIS, Vol. 2

Handling Conflicts to Test Transport Protocol's Parallel Routing on a Vehicle Gateway System

Hassan Mohammad
MBtech Group GmbH & Co. KGaA
Sindelfingen 71063, Germany
Email:hassan.mohammad@mbtech-group.com

Muhammad Shamoon Saleem
Ingolstadt University of Applied Sciences
Ingolstadt 85049, Germany
Email:ia2632@thi.de

***Abstract*—This paper addresses the issue of verifying transport protocol's parallel routing functionality on a vehicle gateway system. The focus of the paper is to construct a conflict-free input parameter model for testing this functionality. The input parameter model shall support the reduction of combinations to be tested and serves as a basis for automatic test case generation from a large space of input parameters. In the proposed approach, defined similarity criteria are used to cluster system input parameters represented as transport protocol routing instances into groups which stimulate similar behavior in the gateway when transport protocol routing is established. Subsequently, the two conflict-handling methods *sub-models* and *avoid* are utilized to prohibit invalid combinations of transport protocol routing instances. The proposed approach is applied on a complex example of real gateway with five buses, 390 transport protocol routing instances and diverse conflicts to illustrate its applicability.**

***Index Terms*—Transport Protocol's Parallel Routing, Input Parameter Model (IPM), Conflict Handling**

I. Introduction

TODAY'S vehicles Electric/Electronic (E/E) systems are designed as distributed systems in order to overcome the increasing complexity and to meet the diversity of requirements such as performance, comfort, safety and costs. In a vehicle distributed system, gateways are indispensable. They enable Electric Control Units (ECUs) within connected networks to interchange information necessary for accomplishing specified functionalities. During information interchange, the gateway routes data between its connected networks although they work on different communication protocols.

In a vehicle gateway system, routing data packets which are larger than a single frame of the corresponding network communication protocol is carried out with the help of transport protocol implementations and such type of routing is called TP routing. Since TP data packets are mostly large in size (flash data for example) and routing of such large data packets requires longer duration, modern gateways support TP parallel routing where multiple TP routing instances are established in parallel over the gateway in order to save time and resources, e.g., flashing multiple ECUs in parallel.

Verifying TP parallel routing of a gateway system is not a trivial problem, since a large number of possible combinations of communicating ECUs can be built for test case selection (the combinatorial explosion problem). Furthermore, in case of established TP parallel routing, different types of conflicts between ECUs need to be handled.

Combination strategies [1] are test case selection methods that focus on solving the combinatorial explosion problem raised while testing the interactions between system input parameter values by defining coverage criteria to satisfy. However, the problem in the case of verifying TP parallel routing on a vehicle gateway system is different than described problem of combination test strategies (see II-B). Hence, new techniques need to be defined.

In combination strategies, Input Parameter Models (IPMs) [2] [3] [4] [5] are essential. They represent the System Under Test (SUT) on an abstract level. Mostly, IPMs contain conflicts which must be resolved. A conflict in an IPM is due to an invalid combination of input parameter values and hence this combination must be omitted or avoided while generating test cases. Diverse conflict handling strategies such as *sub-models* and *avoid* have been suggested in literature [6] to overcome conflicts in IPMs.

This paper suggests an approach to build a conflict-free IPM used to test TP parallel routing on a vehicle gateway system. The IPM is utilized in a recursive way for test case selection, generation and execution in order to overcome the combinatorial explosion problem.

The remainder of this paper is organized as follows. Section II gives background information on terminology and the combinatorial explosion problem of testing TP parallel routing. An IPM along with conflict handling mechanisms are presented in details in section III. In section IV, the suggested IPM along with described conflict handling mechanisms are utilized in a method to reduce test suite size. The complete methodology is applied on a complex example of real gateway in section V. Section VI discusses the approach and section VII concludes the paper.

II. Background

A. Vehicle Bus Communication Systems

Generally, different bus communication systems are utilized in E/E system. In this subsection, two common types of automotive buses are briefly explained.

This work was supported by MBtech Group GmbH & Co. KGaA

978-83-60810-58-3/$25.00

1) CAN Communication Bus: Controller Area Network (CAN) bus system was originally invented to reduce the wiring harness in automotive E/E systems by developing a serial communication protocol. CAN has the capability to connect multiple ECUs directly to one medium, which leads to better management of the system complexity and the reduction of manufacturing costs. CAN protocol is an asynchronous serial protocol that enables real time communication. In CAN, the medium access is based on the concept of Carrier Sense Multiple Access/Collision Detection (CSMA/CD).

2) FlexRay Communication Bus: FlexRay was developed to deliver deterministic, fault-tolerant and high-speed communication bus required for x-by-wire applications such as steer-by-wire and break-by-wire. These applications demand more safety, performance and reliability than provided by CAN. FlexRay unifies time- and dynamic event-triggering mechanisms in one protocol. The communication in FlexRay is based on the Time Devision Multiple Access (TDMA) and Flexible Time Division Multiple Access (FTDMA) schemes of networking.

B. Terminology

The vehicle gateway is part of a distributed network system. It is a special ECU that has multiple communication channels used to communicate with other ECUs in the network in order to route data between them. Communication channels of the gateway are mostly heterogeneous in respect of characteristics and behavior. Generally, a number of ECUs (u) can exchange data over the gateway in predefined fashions. Each fashion is characterized through a set of gateway configuration parameters. These are required by the gateway to establish routing between communicating ECUs connected to different communication channels. For TP routing over the gateway, following definitions are considered:

A *TP_Routing_Fashion* describes a possible routing behavior of TP data and is characterized through a particular set of gateway configuration parameters. An example of a $TP\_Routing\_Fashion_F$ with P parameters shall be described (1) (see [7] for configuration parameters of CAN TP).

$$TP\_Routing\_Fashion_F = \{P_{F_1}, P_{F_2}, ..., P_{F_P}\} \quad (1)$$

A *TP_Connection_Channel* is an instance of a *TP_Routing_Fashion*. It has the same set of gateway configuration parameters and is utilized to route TP data between respective ECUs. A $TP\_Connection\_Channel_x$ in the $TP\_Routing\_Fashion_F$ shall be described (2).

$$TP\_Connection\_Channel_x = \{P_{F_1x}, P_{F_2x}, ..., P_{F_Px}\} \quad (2)$$

The gateway has a number of configured *TP_Connection_Channel*s for connected ECUs utilized to establish TP routing in different scenarios. Examples of TP scenarios are flashing and Onboard Diagnostic (OBD). *TP_Routing_Scenario*s shall be described as a group of s scenarios (3).

$$TP\_Routing\_Scenarios = \{S_1, S_2, ..., S_s\} \quad (3)$$

A *TP_Routing_Instance* is a relationship between a specific *TP_Connection_Channel* and a possible *TP_Routing_Scenario*. An example of a $TP\_Routing\_Instance_x$ shall be described (4).

$$TP\_Routing\_Instance_x = \{P_{F_{1x}}, P_{F_{2x}}, ..., P_{F_{Px}}, S_x\} \quad (4)$$

The gateway can be configured to serve y *TP_Routing_Instance*s in parallel. The number of parallel *TP_Routing_Instance*s *"y"* is a configuration parameter which needs to be verified. In the case of errors, the next determined *"y"* should be verified.

C. The Combinatorial Explosion Problem of Testing TP Parallel Routing

The combinatorial explosion problem mentioned in literature [8] [9] [10] shall be explained as in the following example:

Assume a distributed system consisting of a central unit interacting over communication channels with u units of the network U_1, U_2, ..., U_u. Each unit U_i uses a defined parameter p_i for communication. The parameter p_i shall have v_i possible configuration values. By assuming that configuration values of parameters are independent from each other, the number of possibilities in which the system can be configured would be $v_1 * v_2 * ... * v_u$. If each possible configuration requires c test cases to verify it, the number of test cases for exhaustive test would be $c * v_1 * v_2 * ... * v_u$. In a nontrivial software system, the values of u and v_i are large which leads to a huge number of possible combinations of parameter values.

Related to testing TP parallel routing, the goal of test is to measure the performance of the gateway to handle parallel *TP_Routing_Instance*s. The problem is more complex because:

- System input parameters are *TP_Routing_Instance*s where each consists of a set of configuration parameters.
- The number of system input parameters is not fixed. It can be different for every new release of the system.
- System input parameters include also timing parameters where the interactions are difficult to resolve.
- The number of parallel *TP_Routing_Instance*s *"y"*, which is also a configuration parameter, is used to build possible combinations to be tested. Combinations are any y elements from the system input parameter set. In case of errors, one of the goals is to determine the next *"y"* and verify it (performance measurement).
- In TP parallel routing, each additional instance will consume resources of the system and may lead to errors. Hence, it is not only a specific combination of *TP_Routing_Instance*s which can affect the behavior and may reveal errors, but also the number of included *TP_Routing_Instance*s and their values.

To verify *"y"* parallel routing of *TP_Routing_Instance*s and determine the next *"y"* in case of errors, all possible combinations from 1 to y of *TP_Routing_Instance*s should be included at least once in test cases. This results in a number

of combinations to be tested which can be calculated using equation(5).

$$X = \frac{u!}{y!(u-y)!}.y^s + \frac{u!}{(y-1)!(u-(y-1))!}.(y-1)^s + ... + \frac{u!}{1!(u-1)!}.1 \quad (5)$$

The equation (5) calculates the sum of all possible combinations for a number of ECUs (u) communicating in *TP_Routing_Scenario*s (s), where the selected number of ECUs in each term varies from y to 1.

Even for a simple system, the duration of testing for all these combinations is too high and therefore not acceptable.

III. Input Parameter Model and Conflict Handling

A combination test strategy consists generally of two steps:

1) Building a suitable IPM
2) Selecting combinations of parameter values from IPM to satisfy defined coverage criteria.

In this paper, the proposed IPM shall be used to test TP parallel routing in a recursive approach, i.e., one combination is constructed at a time. Subsequently, a test case is generated for that combination and executed. This procedure is then repeated until satisfying the defined coverage criterion. The decision on using a recursive selection and execution approach has two reasons:

1) Information from executed test cases is collected and can be used to reduce the number of combinations for the successive test cases.
2) Gained information from executed test cases is utilized to determine the worst case combinations.

In order to build a suitable conflict-free IPM, conflicts need to be defined. In a TP parallel routing, three types of conflicts are stated:

1) **Type$_A$-Conflicts**: Conflicts between *TP_Routing_Scenario*s. These are constraints describing *TP_Routing_Scenario*s not practiced in parallel.
2) **Type$_B$-Conflicts**: Conflicts between *TP_Connection_Channel*s. These are constraints describing *TP_Connection_Channel*s not allowed to be combined in parallel.
3) **Type$_C$-Conflicts**: Conflicts because of configuration parameters. These are constraints describing known or desired capabilities of the SUT.

The following example explains raised conflicts in testing TP parallel routing on a vehicle gateway system:

Assume a gateway having six *TP_Routing_Instance*s in *TP_Routing_Fashion*s (F_x and F_y) used to route TP data between connected ECUs as following:

$$TP\_Routing\_Instance_1 = (P_{F_{x11}}, P_{F_{x21}}, P_{F_{x31}}, P_{F_{x41}}, S_1)$$
$$TP\_Routing\_Instance_2 = (P_{F_{x12}}, P_{F_{x22}}, P_{F_{x32}}, P_{F_{x42}}, S_2)$$
$$TP\_Routing\_Instance_3 = (P_{F_{x13}}, P_{F_{x23}}, P_{F_{x33}}, P_{F_{x43}}, S_1)$$
$$TP\_Routing\_Instance_4 = (P_{F_{x14}}, P_{F_{x24}}, P_{F_{x34}}, P_{F_{x44}}, S_2)$$
$$TP\_Routing\_Instance_5 = (P_{F_{y15}}, P_{F_{y25}}, P_{F_{y35}}, S_1)$$
$$TP\_Routing\_Instance_6 = (P_{F_{y16}}, P_{F_{y26}}, P_{F_{y36}}, S_2)$$

Assume also that the gateway supports two *TP_Routing_Instance*s in parallel. Then, following examples explain the three types of conflicts:

- **Type$_A$-Conflicts**: *TP_Routing_Scenario* S_1 and *TP_Routing_Scenario* S_2 are non-combinable. That is, a combination of *TP_Routing_Instance*$_1$ and *TP_Routing_Instance*$_2$ is for example an invalid combination.
- **Type$_B$-Conflicts**: *TP_Connection_Channel*$_2$ and *TP_Connection_Channel*$_4$ are non-combinable, i.e., a combination of *TP_Routing_Instance*$_2$ and *TP_Routing_Instance*$_4$ is an invalid combination.
- **Type$_C$-Conflicts**: Maximum of two *TP_Routing_Instance*s are combinable, i.e., all combinations of more than two *TP_Routing_Instance*s are invalid.

As discussed in the previous section, a *TP_Routing_Instance* of gateway is a relationship between *TP_Connection_Channel* and possible *TP_Routing_Scenario*. *TP_Routing_Instance*s are input parameter values required to build a conflict-free IPM for testing. To achieve building a conflict-free IPM which supports the reduction of combinations in a recursive approach, following steps are required:

1) Collecting and extending *TP_Routing_Instance*s based on similarity criteria.
2) Handling conflicts.

A. Collecting and Extending TP_Routing_Instances based on Similarity Criteria

In collecting *TP_Routing_Instance*s, instances that stimulate similar behavior in the gateway are clustered into groups. Two *TP_Routing_Instance*s are said to be similar if and only if they have the same values for all related parameters such as routing parameters, network relationship parameters and mapped *TP_Routing_Instance*. Creating groups of similar *TP_Routing_Instance*s helps in reducing the number of combinations required for testing. The resulting combination from formulating groups are defined as *"reduced combinations"*.

Following example explains reduction achieved after groups are constructed.

Assume that the SUT has 4 *TP_Routing_Instance*s A, B, C and D (see Fig. 1) and it is configured to support 2 *TP_Routing_Instance*s in parallel. The number of possible combinations of 2 instances out of 4 would be 6 (the order has no effect). If *TP_Routing_Instance*s A,B and C,D are similar to each other, then groups can be constructed based on similarity criteria such that *Group*$_1$ consists of instances A and B, whereas *Group*$_2$ consists of instances C and D. After grouping, the number of combinations could be rather reduced from 6 to 3, because all other possible combinations would resemble a similar behavior, i.e., combinations of instances

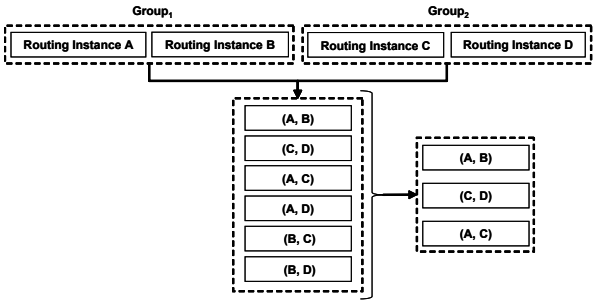


Fig. 1. Advantages of Building Similar Groups

(A, C), (A, D), (B, C) and (B, D) are all similar and can be replaced by only one substitute (A, C). Therefore (A,B), (C,D) and (A,C) are the *"reduced combinations"*.

In extending *TP_Routing_Instance*s, *Similarity Number*s and *Stress Factor*s are assigned to constructed groups. The same *Similarity Number* will be assigned to groups with *TP_Routing_Instance*s having the same routing parameters, the same *TP_Routing_Scenario* and the same characteristics for network relationship. Concerned network characteristics are the protocol type and channel bandwidth. The *Stress Factor* is calculated based on aspects such as processing time, memory usage, channel bandwidth and other channel specific aspects. *Stress Factor*s shall be assigned during test case execution. *Similarity Number*s and *Stress Factor*s are required for reducing combinations during testing. This will be discussed later in details. The basic idea to resolve the combinatorial explosion problem is by using *"reduced combinations"* formulated from constructed groups based on similarity criteria.

B. Conflict Handling

In [2], four different methods for handling conflicts in IPMs were investigated. The result of the study was that the *avoid* method is best suited with respect to size of test suite if it can be utilized. To handle conflicts with the *avoid* method; a procedure is integrated in the test case selection algorithm to prohibit choosing of conflicting combinations. Another method mentioned in the study is the *sub-models* method, in which conflicts are removed by splitting the IPM into multiple smaller conflict-free IPMs used separately to generate test cases. In this paper, a combination of these two conflict handling methods is utilized to handle the defined conflicts for testing TP parallel routing.

1) Type$_A$-Conflict Handling: Since constructed groups possess *TP_Routing_Instance*s similar to each other in all assigned parameters, *TP_Routing_Instance*s of each group will have the same attached *TP_Routing_Scenario*. To handle conflicts existing between *TP_Routing_Scenario*s, the *sub-models* method is utilized. The *TP_Routing_Scenario* is used to split IPM into sub-IPMs which are Type$_A$-Conflict-free. The resulted sub-IPMs shall have no combinable *TP_Routing_Instance*s among each other. This step can be completely achieved before executing any test case, i.e., it is not part of the recursive mechanism. Subsequently, resulted sub-IPMs are processed successively in a recursive methodology.

2) Type$_B$-Conflict Handling: As described before, a *TP_Routing_Instance* is a relationship between a *TP_Connection_Channel* and a *TP_Routing_Scenario*. In order to handle conflicts between *TP_Connection_Channel*s, the *avoid* method is utilized. To implement the *avoid* method, two reserved parameters, i.e., *"Token"* and *"Include Times"*, are attached to individual *TP_Routing_Instance*s. The value of the first parameter *"Token"* is used to decide whether a *TP_Routing_Instance* is allowed to be included in the next combination or not. The *Token* parameter can have one of the following values:

- **0:** Related *TP_Routing_Instance* having no Type$_B$-conflicts and is allowed to be included in a combination.
- **1:** Related *TP_Routing_Instance* having Type$_B$-conflicts and is allowed to be included in the next combination.
- **2:** Related *TP_Routing_Instance* having Type$_B$-conflicts and is not allowed to be included in the next combination.

The second parameter *"Include Times"* is used to hold the number of times a *TP_Routing_Instance* is included in constructed combinations. It is utilized to choose *TP_Routing_Instance*s from the same group that have not yet been included or less included than other instances. Every *TP_Routing_Instance* having Type$_B$-conflict is assigned an *"Include Times"* value "0" that gets incremeted whenever that particular *TP_Routing_Instance* becomes part of a generated combination.

In order to avoid combinations having Type$_B$-conflict, the *Rotate ()* function is called. Each time the function is called, it assigns a suitable value for parameter *Token* based on Type$_B$-conflicts and the value of parameter *Include Times*. After calling the *Rotate ()* function, Type$_B$-conflicting *TP_Routing_Instance*s with the minimum value of parameter *Include Times* will be assigned the value 1 for *Token* and all other Type$_B$-conflicting *TP_Routing_Instance*s will get the value 2. Type$_B$-conflict-free *TP_Routing_Instance*s will be assigned the value 0. The algorithm for generating the final reduced combinations will avoid *TP_Routing_Instance*s with the value 2 for *Token* parameter.

3) Type$_C$-Conflict Handling: Conflicts based on configuration parameters are constraints considered in the test case selection and generation procedure. These constraints shall be corrected if error arises, e.g., correcting the maximum number of parallel routing instances of CAN transport protocol.

IV. Reduction of Test Suite

In order to reduce the size of test suite, a new recursive test mechanism consisting of the following two test phases is proposed:

1) Testing TP parallel routing of Single Network Relationships (SNRs). An individual SNR comprises groups of *TP_Routing_Instance*s responsible for routing TP data between two specific networks of the gateway.

2) Testing TP parallel routing of Mixed Network Relationships (MNRs). An individual MNR comprises groups of *TP_Routing_Instance*s from all SNRs belonging to the same sub-IPM.

In the first test phase, testing of TP parallel routing for individual pairs of connected networks is carried out by means of groups belonging to individual SNRs. The goal of this test is to cover simple interactions (routing) between network pairs. After completing the first test phase, the second test phase is conducted. The goal of the second test phase is to cover complex interactions between multiple interacting networks by means of groups belonging to individual MNRs. Testing SNRs and MNRs separately is very useful for managing the test complexity and reducing the number of combinations. Furthermore, it provides information about possible reason for the occurrence of errors and their relationship with certain pairs of networks or parameter.

A. Coverage Criterion

Coverage criterion is an essential element of combination test strategies. It affects the complexity and the thoroughness of the test. In the proposed approach for testing TP parallel routing, coverage criteria is determined with respect to combinations of *TP_Routing_Instance*s from constructed groups. For example, *1-wise* coverage requires that at least one possible combination with maximum allowed *TP_Routing_Instance*s from every input group is included at least once in a test case. Since all *TP_Routing_Instance*s of a group stimulate a similar behavior in the gateway, any one of such a combination from each group is sufficient enough to satisfy *1-wise* coverage criterion. *N-wise* coverage criterion (exhaustive testing) requires that all possible combinations of *TP_Routing_Instance*s from *N* input groups are included in some test cases, where *N* is the number of input groups.

For an arbitrary number of input groups, the algorithm generates all possible combinations satisfying the *N-wise* interactions between the groups. The maximum number of resulting combinations is calculated as in (6):

$$X = \frac{(n+y-1)!}{y!(n-1)!} \tag{6}$$

Where *n* is the number of groups and *y* is the maximum number of allowed parallel *TP_Routing_Instance*s. Table I explains the algorithm for generating combinations with the help of an example of 4 input groups (G_1,G_2,G_3,G_4) and a maximum of 3 allowed parallel *TP_Routing_Instance*s. Resulting rows in Table. I represent the combinations and the numbers in columns represent the number of *TP_Routing_Instance*s selected from each corresponding group for test case generation. The maximum number of resulting combinations (rows) in this example can be calculated according to (6) and is equal to 20.

While generating test cases, following aspects are considered:

- To guarantee that each *TP_Routing_Instance* is included at least once in some combinations, the *Rotate ()* function is used. Upon calling this function, it generates new *Token* values for *TP_Routing_Instance*s based on the last values of *Include Times* variables and $Type_B$-Conflicts.
- The sum of column elements must be greater than or equal to the number of elements in the corresponding group in order to guarantee that each *TP_Routing_Instance* has been included at least once. If this is not the case, the remaining *TP_Routing_Instance*s from that group shall be tested individually.
- The number in each column must be lesser than or equal to the total number of *TP_Routing_Instance*s in the corresponding group. Otherwise, the combination in the related row is invalid and must be omitted.

TABLE I
COMBINATIONS FROM 4 GROUPS WITH MAXIMUM OF 3 *TP_Routing_Instance*S

G_1	G_2	G_3	G_4
3	0	0	0
2	1	0	0
2	0	1	0
2	0	0	1
1	2	0	0
1	1	1	0
1	1	0	1
1	0	2	0
1	0	1	1
1	0	0	2
0	3	0	0
0	2	1	0
0	2	0	1
0	1	2	0
0	1	1	1
0	1	0	2
0	0	3	0
0	0	2	1
0	0	1	2
0	0	0	3

B. TP Parallel Routing of SNRs

The procedure for testing TP parallel routing of SNRs is depicted in Fig. 2. The procedure accepts constructed $Type_A$-Conflict-free sub-IPMs as input and processes them successively. For each $Type_A$-Conflict-free sub-IPM, the *Selector ()* function extracts groups for the first SNR. In the next step, a combination is built for the current processed SNR with the help of the mechanism explained previously satisfying *N-wise* coverage criterion. After that, a test case will be generated and executed for the build combination. During test case generation, $Type_B$- and $Type_C$-conflicts are handled as explained previously. Result of the test case is analyzed in the following step in which *Stress Factor* is determined and configuration parameters are corrected if a variance has been observed. This procedure is repeated for all combinations until the *N-wise* coverage criteria is satisfied for each SNR. This

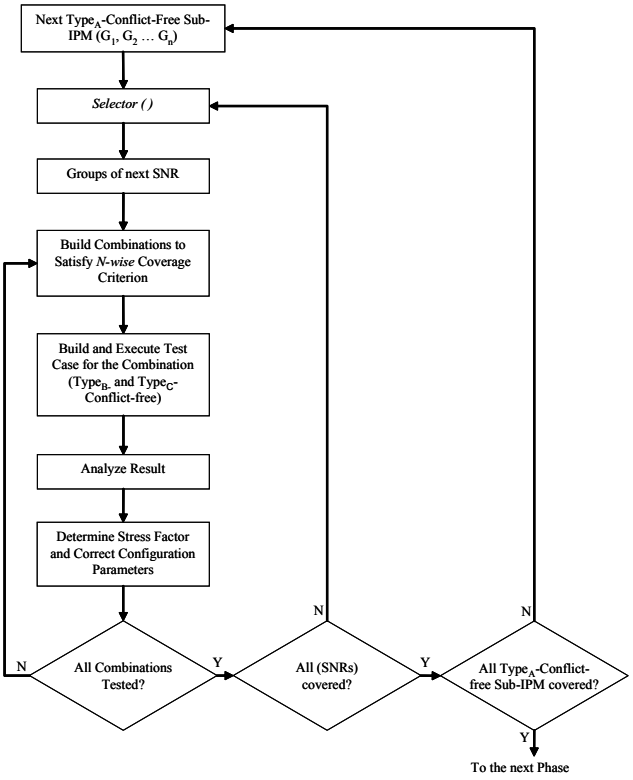


Fig. 2. TP Parallel Routing for SNRs

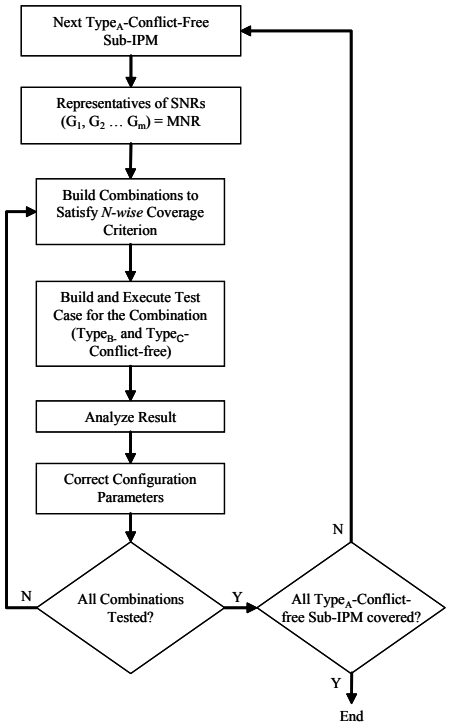


Fig. 3. TP Parallel Routing for MNRs

process is then repeated until all Type$_A$-conflict-free sub-IPM are covered. Processing combinations successively has the advantage to reduce further invalid combinations depending on the corrected configuration parameters. After finishing this phase, each group will be assigned a calculated *Stress Factor*, which will be used by the next phase.

C. TP Parallel Routing of MNRs

The procedure for testing TP parallel routing of MNRs is depicted in Fig. 3. In this procedure, Type$_A$-Conflict-free sub-IPMs are processed successively. Each Type$_A$-conflict-free sub-IPM consists of an arbitrary number of SNRs. A representative group with the best calculated *Stress Factor* for each of these SNR is selected. These representative groups are called MNR and they are the base for testing in this phase. That is, for each processed Type$_A$-Conflict-free sub-IPM, a MNR with selected groups is constructed. After selecting groups of a MNR, groups having the same *Similarity Number*s will be omitted in oder to reduce repititions in combinations. Then, the same procedure as that in previous phase is utilized to build combinations and execute test cases until *N-wise* coverage criterion is satisfied. The procedure will stop once all Type$_A$-Conflict-free sub-IPMs are processed. During recursive testing, constraints are corrected if an error is observed. This helps in further reduction of remaining combinations required to satisfy defined coverage criteria.

V. EXPERIMENT

The central gateway used in this experiment is a special electronic control unit which connects five different networks representing five functional domains of a modern vehicle. Bus systems of the five networks are listed as follows:

- A CAN network for diagnostic functional domain with 500 kilo baud, denoted as bus 1.
- A CAN network for periphery and power train functional domain with 500 kilo baud, denoted as bus 2.
- A CAN network for body functional domain with 250 kilo baud, denoted as bus 3.
- A CAN network for telematics functional domain with 500 kilo baud, denoted as bus 4.
- A FlexRay network for chassis functional domain with 10 MB, denoted as bus 5.

The gateway has 390 *TP_Connection_Channel*s which are separated in:

- 190 *TP_Connection_Channel*s used to transfer data from external device to available ECUs over the gateway.
- 190 *TP_Connection_Channel*s used to transfer data from available ECUs to external device over the gateway.
- 4 *TP_Connection_Channel*s used to transfer data between 4 couples of ECUs in the one direction.
- 4 *TP_Connection_Channel*s used to transfer data between 4 couples of ECUs in the another direction.
- 2 functional *TP_Connection_Channel*s.

Similarity criteria used to form groups are:

- Network relationship of *TP_Connection_Channel*s (source and destination networks).
- Addressing format (normal or mixed).
- Block Size (BS) and separation minimum time (STmin) if they are available.
- Cycle Repetition if it is available.
- ID Type (normal or extended).
- Address Type (physical or functional or extended).
- *TP_Routing_Scenario*

*TP_Routing_Scenario*s determined are:

- Flashing (S_1).
- Uploading (S_2).
- OBD in the one direction (S_3).
- OBD in the other direction (S_4).
- Functional routing (S_5).

Table II, table III, table IV and table V represent the resulting sub-IPMs with their corresponding groups.

Determined conflicts are listed below:

- $Type_A$-Conflicts
  - Flashing and Upload are non-combinable
  - Flashing and OBD in both directions are non-combinable
  - Upload and OBD in both directions are non-combinable
  - OBD in both directions are non-combinable
- $Type_B$-Conflicts
  - *TP_Connection_Channel*s with the same ID but different value for Node Address for Diagnose (NAD) are non-combinable
  - *TP_Connection_Channel*s with the same Slot-ID, Base-Cycle and Cycle-Repetition but different source address are non-combinable
- $Type_C$-Conflicts
  - The maximum configured parallel routing instances for CAN-CAN routing <= 30
  - The maximum configured parallel routing instances for CAN-FlexRay routing <= 8

Transport protocols implemented are based on the ISO standards, ISO 10681 for FlexRay TP [11], and ISO 15765 for CAN TP [7]. Handling $Type_B$- and $Type_C$-Conflicts during test lead to following maximum number of combinations in the first test phase:

- First sub-IPM
  - First SNR: 10 combinations
  - Second SNR: 152 combinations
  - Third SNR: 1 combination
  - Fourth SNR: 31 combinations
- Second sub-IPM
  - First SNR: 10 combinations
  - Second SNR: 78 combinations
  - Third SNR: 1 combination
  - Fourth SNR: 31 combinations
- Third sub-IPM
  - First SNR: 1 combination
  - Second SNR: 1 combination
  - Third SNR: 1 combination
  - Fourth SNR: 1 combination
- Fourth sub-IPM
  - First SNR: 1 combination
  - Second SNR: 1 combination
  - Third SNR: 1 combination
  - Fourth SNR: 1 combination

Table VI represents an example of a resulting combination from the first SNR of the first Sub-IPM. The first SNR of the first Sub-IPM consists of the groups G_1, G_2, G_3, G_4, G_{18} and G_{19} from table II. Following numbers of *TP_Routing_Instance*s have been selected to generate the first combination:

- 5 *TP_Routing_Instance*s from group G_1
- 8 *TP_Routing_Instance*s from group G_2
- 14 *TP_Routing_Instance*s from group G_3
- 3 *TP_Routing_Instance*s from group G_4

VI. Discussion

Depending on the grade of diversity in parameters for *TP_Connection_Channel*s and in connected networks; the number of formulated groups can increase. The idea is to cover the interactions between formulated groups instead of interactions between *TP_Routing_Instance*s. This helps in avoiding similar combinations and contribute to reduce the size of the test suit. If the number *N* of groups is high, either a *2-Wise* or *3-wise* coverage criteria can be used. A drawback of proposed approach is the need of system functionality expertise to define similarity criterion and calculate the Stress Factors of the groups. However this needs to be performed only once. Later on, for each new release of the system, combinations can be generated automatically.

VII. Conclusion and Future Work

In this paper, an approach for building a conflict-free IPM to test TP parallel routing on a vehicle gateway system has been proposed. Firstly, the approach utilizes defined similarity criteria to cluster system input parameters represented as *TP_Routing_Instance*s into groups stimulating similar behavior. Secondly, the two conflict handling methods *sub-models* and *avoid* are used to prohibit invalid combinations of *TP_Routing_Instance*s. In order to reduce the size of the test suite, two phases of testing have been suggested, testing TP parallel routing for SNRs in order to cover simple interactions between network pairs, and testing TP parallel routing for MNRs to cover complex interactions between multiple interacting networks. The proposed approach has been applied on a complex example of a gateway with five buses, 390 *TP_Routing_Instance*s and diverse conflicts. Generating test cases for resulting combinations from the first test phase and conducting the second test phase will be achieved in the future. Furthermore, a restbus simulation is under development to execute the generated test cases for build combinations.

TABLE II
FIRST SUB-IPM'S GROUPS

Groups	Network Relationship	Addressing Format	BS and STmin	Cycle Repetition	ID Type	Address Type	TP Routing Scenario
G_1 (19 elements)	1,2	Mixed	(32,0) (32,0)	-	Normal	Physical	S_1
G_2 (8 elements)	1,2	Normal	(32,0) (8,10)	-	Norma	Physicall	S_1
G_3 (14 elements)	1,2	Normal	(32,0) (32,0)	-	Normal	Physical	S_1
G_4 (8 elements)	1,2	Normal	(8,10) (8,10)	-	Extended	Physical	S_1
G_5 (58 elements)	1,3	Mixed	(32,0) (32,0)	-	Normal	Physical	S_1
G_6 (29 elements)	1,3	Normal	(32,0) (8,10)	-	Normal	Physical	S_1
G_7 (5 elements)	1,3	Mixed	(32,0) (32,20)	-	Normal	Physical	S_1
G_8 (3 elements)	1,3	Normal	(32,0) (32,0)	-	Normal	Physical	S_1
G_9 (8 elements)	1,4	Normal	(32,0) (32,0)	-	Normal	Physical	S_1
G_{10} (2 elements)	1,4	Normal	(4,10) (4,10)	-	Normal	Physical	S_1
G_{11} (13 elements)	1,4	Normal	(32,0) (8,10)	-	Normal	Physical	S_1
G_{12} (5 elements)	1,4	Mixed	(32,0) (32,0)	-	Normal	Physical	S_1
G_{13} (8 elements)	1,5	Normal	(32,0) (-,-)	2	Normal	Physical	S_1
G_{14} (1 elements)	1,5	Normal	(32,0) (-,-)	1	Normal	Physical	S_1
G_{15} (1 elements)	1,5	Mixed	(32,0) (-,-)	4	Normal	Physical	S_1
G_{16} (1 elements)	1,5	Normal	(8,10) (-,-)	1	Extended	Physical	S_1
G_{17} (7 elements)	1,5	Normal	(32,0) (-,-)	4	Normal	Physical	S_1
G_{18} (1 elements)	1,(2,3,4,5)	Normal	(-,-) (-,-)	-	Normal	Functional	S_5
G_{19} (1 elements)	1,(2,5)	Normal	(-,-) (-,-)	-	Extended	Functional	S_5

TABLE III
SECOND SUB-IPM'S GROUPS

Groups	Network Relationship	Addressing Format	BS and STmin	Cycle Repetition	ID Type	Address Type	TP Routing Scenario
G_1 (15 elements)	2,1	Normal	(8,0) (8,0)	-	Normal	Physical	S_2
G_2 (7 elements)	2,1	Normal	(8,10) (8,0)	-	Normal	Physical	S_2
G_3 (19 elements)	2,1	Mixed	(8,0) (8,0)	-	Normal	Physical	S_2
G_4 (8 elements)	2,1	Normal	(8,0) (8,0)	-	Extended	Physical	S_2
G_5 (63 elements)	3,1	Mixed	(8,0) (8,0)	-	Normal	Physical	S_2
G_6 (29 elements)	3,1	Normal	(8,10) (8,0)	-	Normal	Physical	S_2
G_7 (3 elements)	3,1	Normal	(8,0) (8,0)	-	Normal	Physical	S_2
G_8 (8 elements)	4,1	Normal	(8,0) (8,0)	-	Normal	Physical	S_2
G_9 (13 elements)	4,1	Normal	(8,10) (8,0)	-	Normal	Physical	S_2
G_{10} (5 elements)	4,1	Mixed	(8,0) (8,0)	-	Normal	Physical	S_2
G_{11} (2 elements)	4,1	Normal	(4,10) (4,10)	-	Normal	Physical	S_2
G_{12} (8 elements)	5,1	Normal	(-,-) (8,0)	2	Normal	Physical	S_2
G_{13} (1 elements)	5,1	Normal	(-,-) (8,0)	1	Normal	Physical	S_2
G_{14} (7 elements)	5,1	Normal	(-,-) (8,0)	4	Normal	Physical	S_2
G_{15} (1 elements)	5,1	Mixed	(-,-) (8,0)	4	Normal	Physical	S_2
G_{16} (1 elements)	5,1	Normal	(-,-) (8,0)	1	Extended	Physical	S_2
G_{17} (1 elements)	1,(2,3,4,5)	Normal	(-,-) (-,-)	-	Normal	Functional	S_5
G_{18} (1 elements)	1,(2,5)	Normal	(-,-) (-,-)	-	Extended	Functional	S_5

TABLE IV
THIRD SUB-IPM'S GROUPS

Groups	Network Relationship	Addressing Format	BS and STmin	Cycle Repetition	ID Type	Address Type	TP Routing Scenario
G_1 (3 elements)	3,4	Normal	(4,10) (4,10)	-	Normal	Physical	S_3
G_2 (1 elements)	4,5	Normal	(4,20) (-,-)	16	Normal	Physical	S_3
G_3 (1 elements)	1,(2,3,4,5)	Normal	(-,-) (-,-)	-	Normal	Functional	S_5
G_4 (1 elements)	1,(2,5)	Normal	(-,-) (-,-)	-	Extended	Functional	S_5

TABLE V
FORTH SUB-IPM'S GROUPS

Groups	Network Relationship	Addressing Format	BS and STmin	Cycle Repetition	ID Type	Address Type	TP Routing Scenario
G_1 (3 elements)	4,3	Normal	(4,10) (4,10)	-	Normal	Physical	S_4
G_2 (1 elements)	5,4	Normal	(-,-) (8,20)	16	Normal	Physical	S_4
G_3 (1 elements)	1,(2,3,4,5)	Normal	(-,-) (-,-)	-	Normal	Functional	S_5
G_4 (1 elements)	1,(2,5)	Normal	(-,-) (-,-)	-	Extended	Functional	S_5

TABLE VI
AN EXAMPLE OF A RESULTING COMBINATION

Request ID	Response ID	Network Relationship	BS and STmin	Message DLC	NAD	Addressing Format	ID Type	TP Routing Scenario
0x4e9	0x499	1,2	(32,0) (32,0)	8	14	Mixed	Normal	S_1
0x4e8	0x498	1,2	(32,0) (32,0)	8	32	Mixed	Normal	S_1
0x4c4	0x494	1,2	(32,0) (32,0)	8	5	Mixed	Normal	S_1
0x4d0	0x490	1,2	(32,0) (32,0)	8	8	Mixed	Normal	S_1
0x4e7	0x497	1,2	(32,0) (32,0)	8	13	Mixed	Normal	S_1
0x450	0x5d9	1,2	(32,0) (8,10)	8	-	Normal	Normal	S_1
0x456	0x5d5	1,2	(32,0) (8,10)	8	-	Normal	Normal	S_1
0x7e4	0x7ec	1,2	(32,0) (8,10)	8	-	Normal	Normal	S_1
0x625	0x5a5	1,2	(32,0) (8,10)	8	-	Normal	Normal	S_1
0x654	0x5d4	1,2	(32,0) (8,10)	8	-	Normal	Normal	S_1
0x652	0x5d2	1,2	(32,0) (8,10)	8	-	Normal	Normal	S_1
0x624	0x5a4	1,2	(32,0) (8,10)	8	-	Normal	Normal	S_1
0x653	0x5d3	1,2	(32,0) (8,10)	8	-	Normal	Normal	S_1
0x64e	0x5ce	1,2	(32,0) (32,0)	8	-	Normal	Normal	S_1
0x7e5	0x7ed	1,2	(32,0) (32,0)	8	-	Normal	Normal	S_1
0x7e1	0x7ea	1,2	(32,0) (32,0)	8	-	Normal	Normal	S_1
0x665	0x5e5	1,2	(32,0) (32,0)	8	-	Normal	Normal	S_1
0x778	0x788	1,2	(32,0) (32,0)	8	-	Normal	Normal	S_1
0x64d	0x5cd	1,2	(32,0) (32,0)	8	-	Normal	Normal	S_1
0x656	0x5d6	1,2	(32,0) (32,0)	8	-	Normal	Normal	S_1
0x650	0x5d0	1,2	(32,0) (32,0)	8	-	Normal	Normal	S_1
0x7e2	0x7e8	1,2	(32,0) (32,0)	8	-	Normal	Normal	S_1
0x7e6	0x7ee	1,2	(32,0) (32,0)	8	-	Normal	Normal	S_1
0x7d9	0x7e7	1,2	(32,0) (32,0)	8	-	Normal	Normal	S_1
0x7e3	0x7eb	1,2	(32,0) (32,0)	8	-	Normal	Normal	S_1
0x662	0x5e2	1,2	(32,0) (32,0)	8	-	Normal	Normal	S_1
0x65b	0x6db	1,2	(32,0) (32,0)	8	-	Normal	Normal	S_1
0x18da69f1	0x18daf169	1,2	(8,10) (8,10)	8	-	Normal	Extended	S_1
0x18da66f1	0x18daf166	1,2	(8,10) (8,10)	8	-	Normal	Extended	S_1
0x18da43f1	0x18daf143	1,2	(8,10) (8,10)	8	-	Normal	Extended	S_1

REFERENCES

[1] M. Grindal, J. Offutt, and S. F. Andler, "Combination Testing Strategies: A survey," GMU Technical Report ISE-TR-04-05, July 2004.

[2] M.N. Borazjany, L. S. Ghandehari, Y. Lei, R.N. Kacker, and D.R. Kuhn, "An Input Space Modeling Methodology for Combinatorial Testing," Software Testing, Verification and Validation Workshops (ICSTW), 2013 IEEE Sixth International Conference, pp. 372-381, Luxembourg 2013.

[3] M. Grindal and J. Offutt, "Input Parameter Modeling for Combination Strategies," In Proceeding SE'07 Proceedings of the 25th conference on IASTED International Multi-Conference: Software Engineering, pp. 255-260, USA 2007.

[4] M. Grindal, J. Offutt, and J. Mellin, "Handling Constraints in the Input Space when Using Combination Strategies for Software Testing," Technical Report HS-IKI-TR-06-001. School of Humanities and Informatics, University of Skövde 2006.

[5] S. A. Vilkomir, W. T. Swain, and J. H. Poore, "Software Input Space Modeling with Constraints among Parameters," Computer Software and Applications Conference, COMPSAC '09. 33rd Annual IEEE International, pp. 136-141, Seattle. WA. 2009.

[6] M. Grindal, J. Offutt, and J. Mellin, "Managing Conflicts when Using Combination Strategies to Test Software," Software Engineering Conference, ASWEC 2007. 18th Australian, pp. 255-264. Melbourne, Vic. 2007.

[7] "Road vehicles-Diagnostics on Controller Area Networks (CAN)-," ISO 15765:2004(E), Switzerland : ISO.

[8] D. M. Cohen, S. R. Dalal, M. L. Fredman, and G. C. Patton, "The AETG System: An Approach to Testing Based on Combinatorial Design," IEEE Transaction on Software Engineering, vol. 23 ,pp. 437-444, July 1997.

[9] R. Mandl, "Orthogonal Latin Squares: An application of experiment design to compiler testing," Communications of the ACM, 28(10):1054 -1058, October 1985.

[10] C. J. Colbourn, M. B. Cohen, and R. C. Turban, "A Deterministic Density Algorithm for Pairwise Interaction Coverage," In: Proc. of the IASTED Intl. Conference on Software Engineering, pp. 345-352, Austria 2004.

[11] "Road vehicles-Communication on FlexRay-," ISO 10681:2010(E), Switzerland : ISO.

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 1569–1574

DOI: 10.15439/2014F342
ACSIS, Vol. 2

Automating Acceptance Testing with tool support

Tomasz Straszak, Michał Śmiałek
Warsaw University of Technology
Warsaw, Poland
Email: {straszat, smialek}@iem.pw.edu.pl

***Abstract*—During acceptance testing different areas of delivered software system are reviewed. Usually these are functionality, business domain logic, non-functional characteristics, user interface. Although they are related to the same particular functional area, they are verified separately. This paper presents the concept and the Requirements Driven Software Testing (ReDSeT) tool, which allows for automatic integrated test generation based on different types of requirements. Tests are expressed in newly introduced Test Specification Language (TSL). The basis for functional test generation are detailed use case models. Furthermore, by combining different types of requirements, relations between tests are created. The constructed tool acknowledges validity of the presented concept.**

I. Introduction

SOFTWARE testing is one of the main steps of each development process. In this step the compliance of delivered software with the requirements is being verified. Verification procedures of comparing the system under development to its requirements and needs of its users are encapsulated in the form of acceptance tests [1]. These requirements should be understandable for the stakeholders and at the same time precise enough for the developers to produce efficient software.

To describe the expected functionality of the software system, use cases are commonly used [2]. Use cases describe interactions between external actors and the system, which lead to specific goals according to the given scenarios. Such requirements can be claimed as satisfactory to define the tests, that will be performed during acceptance testing.

A number of automatic test generation mechanisms based on use cases were proposed. Examples of such approaches can be found in work by El-Attar and Miller [3], Gutiérrez et al. [4], and Nebut et al. [5]. Beside use cases, requirement specifications contain other types of requirements, that describe different aspects of the desired software. These requirements should also be verified by executing corresponding tests. Some work has been done on the generation of tests based on business rules (see Junior et al. [6]), GUI requirements (see Bertolini and Mota [7]), and even on non-functional requirements (see Dyrkom and Wathne [8]).

All these mechanisms use model transformation, forming the area of Model-based testing (MBT), which is an evolving technique for generating suites of test cases from requirements [9]. Although different types of tests are generated from requirements models describing the same software system, usually they are not related, because they verify different aspects of the system.

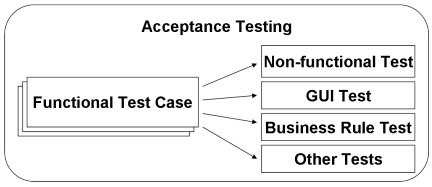


Fig. 1. Acceptance test suite based on functional test cases

This article extends the basic description of the idea presented in our previous work [10]. It focuses on automatic generation of different types of tests, integrated in functional test cases with test scenarios executed during acceptance testing. These tests are generated on the basis of interrelated requirements describing many aspects of the developed software system, making this idea MBT-compliant. The element that joins the different types of testing is the functional test case related to the use case scenario as shown in Figure 1.

This concept is based on the test metamodel defined as the Test Specification Language (TSL) and implemented within the ReDSeT tool (Requirements Driven Software Testing). The tests are generated automatically based on the requirement specification created with the Requirements Specification Language (RSL) [11]. As RSL provides notation for precise use case scenarios, generation of test cases verifying the system behaviour is significantly facilitated. Additional information contained in scenario sentences (notions from the domain vocabulary) and other related requirements allow for generating tests of different types. All the tests generated on the basis of RSL-based requirements form a complete test suite for acceptance testing.

II. Detailed requirements expressed in RSL

As in other test generation solutions, the basis for automatic generation of tests is the precise specification of requirements. As mentioned above, the described solution is based on requirements specifications created with RSL. The main features of this language are: clear separation of descriptions of the system's behaviour and descriptions of the system's domain. Functional requirements can be presented in three equivalent forms: structured text with hyperlinks to domain elements, an activity diagram or a sequence diagram. The elements describing the system's domain are depicted as notions on so-called notion diagrams. Each notion has operations that can be performed in regard to the particular notion. RSL allows for precise specification of requirements, which is understandable even for ordinary people who do not have technical expertise.

978-83-60810-58-3/$25.00

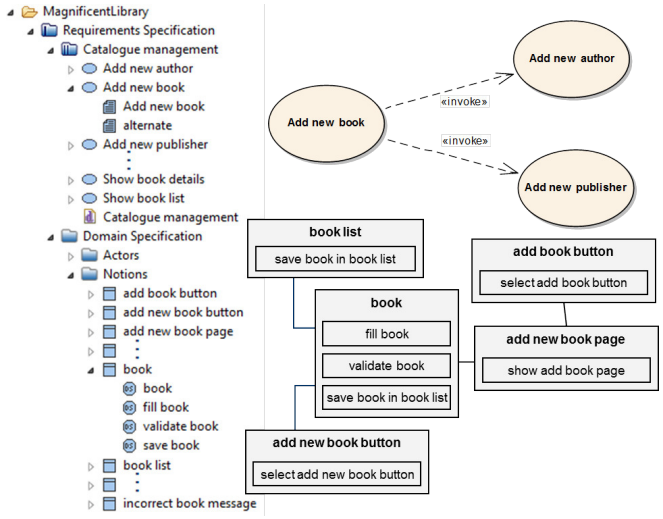


Fig. 2. Example of requirement specification structure, use cases and notion diagrams expressed with RSL

Main Scenario

precondition: book list is ready to add new book
1. User *selects* add new book button
2. System *shows* add new book page
=>invoke/INSERT Add new author
=>invoke/INSERT Add new publisher
3. User *fills* book
4. User *selects* add book button
5. System *validates* book
=>cond: book valid
6. System *saves* book in book list
final: success
postcondition: new book is added to book list

Alternate scenario

precondition: book list is ready to add new book
1. User *selects* add new book button
2. System *shows* add new book page
=>invoke/INSERT
=>invoke/INSERT
3. User *fills* book
4. User *selects* add book button
5. System *validates* book
=>cond: book invalid
5.1.1 System *shows* incorrect book me
final: failure
postcondition: book list has not changed

Fig. 3. Use case scenarios - textual representation

The language has a precise specification of its syntax and semantics [11] with methods of its use explained e.g. by Nowakowski et al. [12]. Figures 2, 3 and 4 shows an example requirements specification, created in RSL.

All the elements of a requirement specification are grouped in packages in a tree structure. Simple requirements described with free text can be used to define business rules or non-functional aspects of the system. Use cases defining the functionality of the system are described with structured scenarios (see Figure 3). Scenarios consist of numbered sentences in a simple grammar called SVO-O (Subject Verb Object - indirect Object). These sentences are constructed with notions stored in the domain vocabulary. This is illustrated with two scenarios (main and alternative) of the *Add new book* use case. The same information is presented in Figure 4 in the form of an activity diagram that is generated automatically from the scenarios.

The notions are referred-to in scenario sentences through hyperlinks (*book*, *book list*, *edit book button*, *edit book page*) and are presented on a notion diagram, that is similar to a class diagram. The relationships between notions, and notion operations are defined automatically according to the scenario sentences where these notions appear or are defined manually by the requirements engineer. The notions and their operations used in use case scenarios, describe the business logic and the user interface elements.

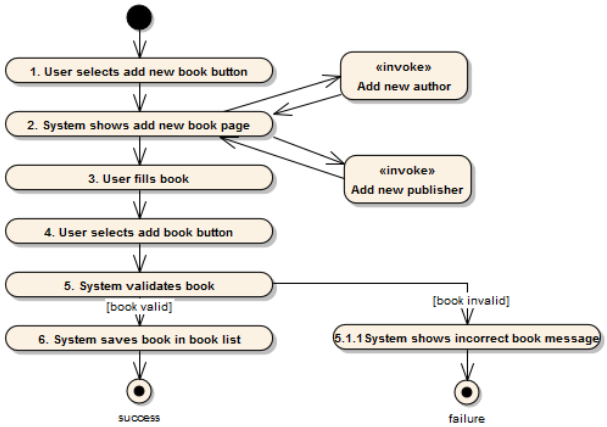


Fig. 4. Use case scenarios - activity representation

All the requirements can be related. To depict relations between use cases, a special invoke relationship is used (see Figures 2 and 4). It allows to determine under what conditions and in which step of a use case scenario another use case is to be called (see Nowakowski et al. [12] for more details). What is important, RSL is based on a formal metamodel. This allows for automatic processing of information contained in the requirement specification. This characteristics of RSL will be used for generating test cases.

III. Automating Test Generation

To define acceptance test suite and to ensure accurate and automatic transition from RSL-based requirements to tests, Test Specification Language (TSL) was developed. This language is based on a metamodel defined in the Eclipse Modeling Framework (EMF) [13] and is out of scope of this paper.

The main idea of TSL is to provide notation for reusable tests, that are understandable for non technical people and precise enough for detailed verification of the software system. All tests are grouped in a tree structure, called the Test Specification (see Figure 5), that groups tests assigned to a specific release of software. Each test contained in a test specification represents a procedure for verifying software in the context of a single requirement. Such a verification is made by examining all the check points defined inside the test.

The basic structure of a TSL test specification consists of two packages: Abstract Tests and Concrete Tests. The first of these includes tests generated directly from the requirement specification: mostly use case tests, notions, as well as tests of other types. A use case test corresponds to a use case, and includes test scenarios, as shown in Figure 5. Tests of other types, in addition to use case tests (verifying the behaviour of the system), can verify the business logic, user interface, non-functional aspects (performance, usability, etc.) or any other aspect of the system that is described through requirements.

A use case test scenario includes the initial condition (a precondition sentence) that must be met before the execution

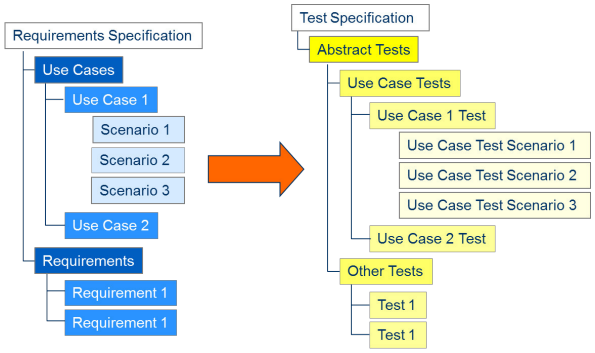


Fig. 5. Test generation based on the requirement specification

of actions described in this scenario and the final condition (a postcondition sentence) that describes the desired state of the system after the scenario is executed.

Every use case test scenario, generated from an RSL use case scenario, is a sequence of actions forming a dialogue between the primary actor and the system. Every such action is expressed by a single sentence in the SVO grammar (see Graham [14] for an original idea). These sentences describing single actions can have check points assigned. In addition to action sentences, two additional sentence types were introduced: condition and control sentences. They are used in a scenario to express the flow of control between alternative scenarios of the same use case as well as between scenarios of different use cases (see work by Śmiałek et al. [15]).

An important feature of requirement specified with RSL, is the possibility to create relationships. Due to generation of test specifications on the basis of these requirements, also relationships between tests can be created. The invocation relations between use cases are translated to become relations between use case tests. This provides information on the steps of the use case test scenario and on the conditions under which scenarios of other use case tests should be called. Relationships to other requirements are translated to relationships from use case tests to tests of other types.

Having two languages (RSL and TSL), which have definitions that are based on metamodels, automatic transformation from requirements to tests becomes possible also allowing for further acceptance test composition. There is a couple of common rules applied in the transformation:

- The structure of the packages containing use cases and notions is reflected in the structure of the packages in the Test Specification.
- Each element of a Test Specification that reflects an element of a requirement specification holds the identifier and the tree path of that element.

The transformation is performed in several steps according to the transformation procedure, that is presented in Figure 6. At the beginning of the transformation a new Test Specification structure is created (step 1). The basic Test Specification structure consists of a root node named using the pattern of "Software Case Name - date" and two child Test Packages named "Abstract Tests" and "Concrete Tests".

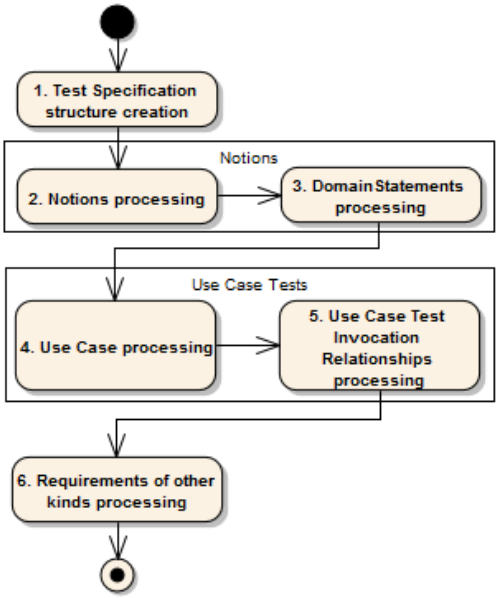


Fig. 6. RSL to TSL transformation procedure

All SVO sentences within the use case test scenarios are built of the notion's domain statements. For this reason, RSL notions should be transformed first (step 2). For each RSL notion, a TSL notion is created. The name, description and attached notion attributes are transferred.

For all TSL notions, domain statements contained in corresponding RSL notions are created (step 3). For each RSL notion, its domain statements are transferred into a TSL domain statement. The phrases contained in the RSL domain statement notions, used as direct and indirect objects, are pointed to by the *directNotion* and the *indirectNotion* attributes.

Having the notions with the domain statements transferred, the use case tests can be processed (step 4). For each RSL use case, a use case test is created and placed in a proper test package within the use case test structure. The name and the description are transferred directly. All the scenarios contained by the RSL use case are transferred into a use case test scenario. On the basis of the RSL scenario pre- and postcondition, adequate pre- and postonditions are created and attached to the use case test scenario. The sentences of the RSL scenarios are transferred into the correct specialisation of the use case test scenario sentence. The ordering number and the sentence text are set. For every SVO sentence, a proper domain statement is found and a relation to the corresponding domain statement is created. For every control sentence, a test invocation relationship is created with an empty use case test as its target.

The target use case tests of the test invocations are set after all the use cases are transferred into the use case tests (step 5). For each test invocation relationship contained in the control sentences, a correct use case test is found and set.

At the end of the transformation (step 6), tests of other kinds are created. Each RSL requirement that is not a use

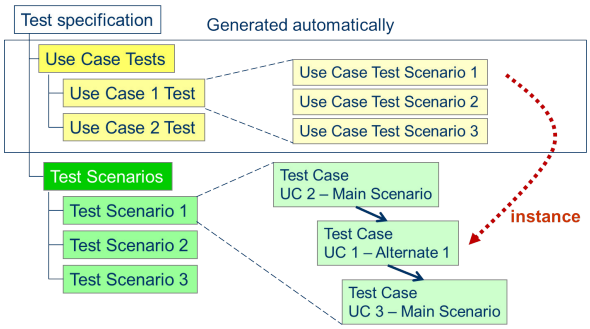


Fig. 7. Test scenarios composed of test cases

case and is classified as a requirement of specific type (e.g. business logic requirement, user interface requirement, non-functional requirement) is the basis for generating a test, which supplements use case tests, scenarios, sentences or notions. RSL requirements relations between use cases, notions and requirements of specific types are transformed into test relationships. As RSL currently does not support basic requirements of specific types, only non-functional requirements are automatically transformed into non-functional tests.

IV. Instantiating Tests

A scenario of a use case test determines the conditions, steps and check points that will be subject to verification for the use case implementation. These elements will be used in acceptance testing after placing them in test scenarios and assigning specific test data values. Test scenarios are grouped within second-level packages in the basic structure of a test specification named Concrete Tests, as shown in Figure 7. They are defined by a test engineer as a set of ordered instances of use case test scenarios, that we call test cases. A test case describes a procedure for verification of a system's functionality and is composed of ordered steps in the form of SVO sentences. Each step can contain check points with assigned test data values and can be related with instances of other type tests. These other test instances are automatically created during instantiation. They are related to a test case, just as abstract tests of other types are related to use case test scenarios and particular scenario sentences.

A test scenario constructed with test cases also builds the context for the test data. The initial test data values are set by the test engineer as the precondition values of the test scenario. Test data values describe basic business objects as well as GUI elements. The test data in the scope of one test scenario is passed between test cases as their pre- and postcondition values. The test data values are changing according to the functionality and the business logic that is under tests. It can be noted that although the test cases cannot be formally related to each other, within the manually created test scenarios, they indirectly refer to business processes that are implemented within the system being tested.

The instantiation procedure of a use case test scenario is performed in at least as much steps as the test scenario is supposed to have. Figure 8 presents the procedure for

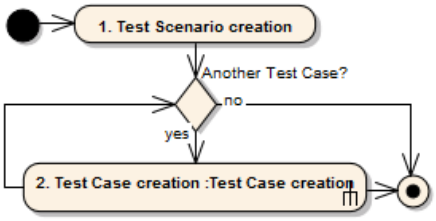


Fig. 8. Test Scenario creation procedure

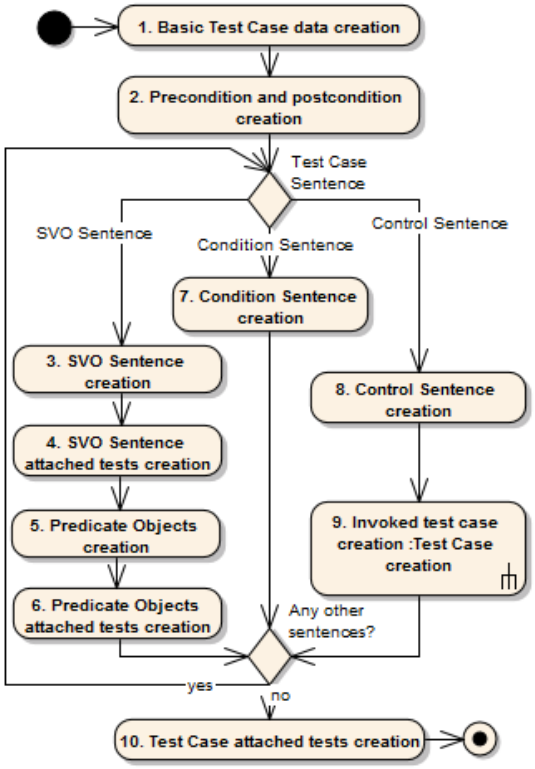


Fig. 9. Test Case instantiation procedure

test scenario creation. The test scenarios are inserted in the instance tests package. During creation, the name and the description of the test scenario should be given. All the steps of a test scenario are created as test cases. The number of test cases depends on the test engineer, who defines the steps of the test scenario. Each time a new test case is created, the instantiation procedure is performed. The procedure is presented in Figure 9.

At the beginning of the procedure, the test case order number is set. For each nested test case, the order number is segmented, e.g.: *2.3.1*. The name of the chosen use case test scenario and the description of the corresponding use case test are transferred into the test case (step 1), the same as for the use case test scenario pre- and postcondition (step 2).

Having the test case created, SVO, condition and invoke sentences are being created. SVO sentences (step 3) are transferred with their sentence order number and sentence text. Direct and indirect objects of the sentence predicates are

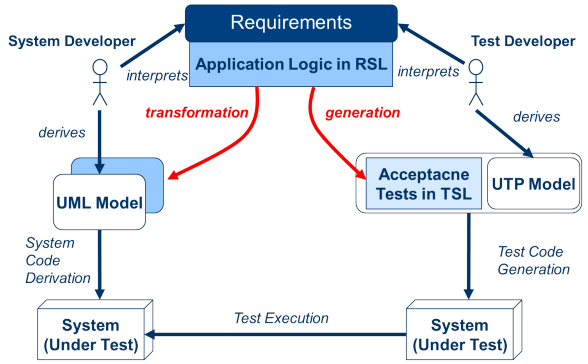


Fig. 10. TSL as UML Testing Profile complement

created on the basis of the direct and the indirect notions of the domain statement (step 5). All abstract tests of other types, related to the domain statement pointed to by the predicate relation and by the domain statements' direct notion and indirect notion relations, are transferred into adequate instance test specialisations (step 4 and 6). These instance tests of other types are contained in the SVO sentences of the test case and in the direct or the indirect object. The condition sentences are transferred with the sentence order number and the sentence text (step 7).

Depending on the test engineer's decision, the nested test case depicted by the control sentence can be created on the basis of a use case test related through a test invocation relationship to the currently processed use case test (step 8 and 9). If the use case test invocation is used, one of the invoked use case test scenarios in instantiated recursively. At the end all other abstract tests related to a use case test scenario are transferred into adequate instance test specialisations (step 10). These tests are contained within the test case.

V. TSL as an Extension of the UML Testing Profile

TSL can be seen as a stand-alone language but it can be easily interfaced with other languages for model-based testing. Prominently, it can be used in conjunction with the UML Testing Profile [16].

Figure 10 presents an appropriate usage scenario. Starting from requirements, a system developer delivers UML models, which are the basis for developing the system. The same requirements are used for manual creation of UTP models. On the basis of these test models, detailed unit and integration tests can be executed. However, this does not include high-level acceptance tests. Here, TSL offers an extension allowing to derive such tests directly and automatically from functional requirements.

The usage of RSL for defining application logic allows for automatic transformation of requirements into code and into acceptance tests in TSL. During these transformations, requirements-to-UML models and requirements-to-TSL-test traces are being created. These traces facilitate linking of UTP models with corresponding acceptance tests in TSL through UML models.

VI. Tool Support

The tool supporting the described idea of automatic test generation based on requirements is called ReDSeT (Requirements Driven Software Testing). It is based on the Eclipse Rich Client Platform. This enables integration with the ReDSeeDS tool (www.redseeds.eu, [17]) which provides advanced editors for requirements (for use cases, notions and other requirements) described with RSL. The generated test specification can be included in the same Eclipse project as the requirements specification and code. This allows for integration of activities at different stages of the software development project. As the repository of the test specification is based on the EMF technology [13], the TSL meta-model can be easily extended in order to handle other types of tests that are adapted to different types of requirements associated with use cases.

To start working with the ReDSeT tool, the requirement specification should be transformed into the test specification. When the automatic transformation is complete, the test engineer is able to manage the test specification organised in a tree structure using the Test Specification Browser and other dedicated editors enclosed in the ReDSeT perspective. Use case tests and use case test scenarios, along with test scenarios and test cases are presented in the Test Editor area. The Detailed Test View is dedicated for viewing check points and editing test values contained in all the types of tests. To create test scenarios and to instantiate use case test scenarios as test cases, dedicated wizards are available.

In order to perform acceptance tests according to test scenarios defined in the ReDSeT tool, the test execution scripts need to be generated. The test scripts, that are composed of test cases, contain detailed steps for the testers in the form of structured text. Each line represents one test with its name, description, input data values and expected state of the system for the specified elements.

The example of test execution script in the form of a CSV text file is presented in figure 11. The rows describing steps of a test scenario have solid background and are shown as numbered test cases. Each sentence of the test case is numbered with the test case number and the SVO sentence number. For each direct and indirect object of the SVO sentence, an additional row for data input or output appears. For example, in the step *2.1 SVO 3*, there are presented the input test data and the values for the author attribute (row number *2.1.SVO 3 DirObj*). Additional tests of other types, related to a sentence or to a direct or indirect object appear as separate rows. For example, a GUI test is presented in the step *2 SVO 2 DirObjGui Test*. This test is attached to the direct object of the step *2 SVO 2*.

On the basis of such a test execution script, the testers can verify the delivered software system. The results of each test step can be noted in an additional column. In case the test fails, the corresponding requirement can be found by tracing to the appropriate requirements element. The implementation units can be precisely located by examining traces to code, as described by Śmiałek et al. [18].

Test Type	Step	Test Name	Description	Test Data	Input Value	Characteristic to test	Expected result
Test Case	1	Show book list	---	---	---	---	---
SVO Sentence	1 SVO 1	User selects show book list button	---	---	---	---	---
Domain Object	1 SVO 1 DirObj	show book list button	---	---	---	---	---
SVO Sentence	1 SVO 2	System fetches book list	---	---	---	---	---
Domain Object	1 SVO 2 DirObj	book list	---	---	---	---	---
SVO Sentence	1 SVO 3	System shows book list page	---	---	---	---	---
Domain Object	1 SVO 3 DirObj	book list page	---	---	---	---	---
Test Case	2	Add new book	---	---	---	---	---
SVO Sentence	2 SVO 1	User selects add new book button	---	---	---	---	---
Domain Object	2 SVO 1 DirObj	add new book button	---	---	---	---	---
SVO Sentence	2 SVO 2	System shows add new book page	---	---	---	---	---
Domain Object	2 SVO 2 DirObj	add new book page	---	---	---	---	---
GUI Test	2 SVO 2 DirObjGUI Test	Asterisk marking mandatory fields	Empty mandatory fields should be marked with asterisk	---	---	Asterisk marking mandatory fields, when they are set	Empt mandatory fields marked with asterisk
Control Sentence	2 Control	INSERT Add new author	---	---	---	---	---
Test Case	2.1	Add new author	---	---	---	---	---
SVO Sentence	2.1 SVO 1	User selects add new author button	---	---	---	---	---
Domain Object	2.1 SVO 1 DirObj	add new author button	---	---	---	---	---
SVO Sentence	2.1 SVO 2	System shows add new author page	---	---	---	---	---
Domain Object	2.1 SVO 2 DirObj	add new author page	---	---	---	---	---
SVO Sentence	2.1 SVO 3	User fills author	---	---	---	---	---
Domain Object	2.1 SVO 3 DirObj	author	---	Author's name and surname	Mark Twain	---	---
SVO Sentence	2.1 SVO 4	User selects add author	---	---	---	---	---

Fig. 11. Test execution script - attached test and SVO sentence object test

VII. Conclusion

The proposed idea and the ReDSeT tool bring a complete solution for creating acceptance tests for the systems that are focused on user-system interaction. The preparation of test specifications can begin during the requirements formulation stage. Consecutive generation of tests allows to reach test complexity that corresponds to the level of detail of the final requirements. The basis for creation of a set of test scenarios are detailed use cases. Requirements defined in RSL significantly facilitate automatic test generation, and TSL allows for expressing interrelated tests in a way that is comprehensible to the audience responsible for acceptance testing. It can be noted that the proposed method is based on black box testing and is independent of the implementation technology of the system under test. On the other hand, since RSL and TSL are based on metamodels, the whole idea is close to Model Based Testing which goes into the details of system design.

In terms of future work, traces from requirements to test cases are planned to be used for generating requirements with test coverage reports. These traces will be subject to further research on regression test selection. Another area that is planned to be a subject of further research is using RSL, TSL and model transformations [19] as implementation of Test Driven Development (TDD) [20] and Behaviour Driven Development (BDD) [21]. These ideas assume that test effort is already incorporated at an earlier point of software development process. The proposed solution seems to have potential for automating the process of generating unit and acceptance tests and executing them on the basis of RSL requirements. It is also planned to develop the mechanism for extracting of test scripts as input for tools that automate test execution (e.g. IBM Rational Functional Tester, Selenium). It would bring a complete solution for detailed use case based testing.

References

[1] G. J. Myers, C. Sandler, and T. Badgett, *The Art of Software Testing*, 3rd ed. Wiley Publishing, 2011.

[2] A. Cockburn, *Writing Effective Use Cases*. Addison-Wesley, 2000.

[3] M. El-Attar and J. Miller, "Developing comprehensive acceptance tests from use cases and robustness diagrams," *Requir. Eng.*, vol. 15, no. 3, pp. 285–306, Sep. 2010. [Online]. Available: http://dx.doi.org/10.1007/s00766-009-0088-6

[4] J. J. Gutiérrez, M. J. Escalona, M. Mejías, and J. Torres, "An approach to generate test cases from use cases," in *Proceedings of the 6th international conference on Web engineering*, ser. ICWE '06. New York, NY, USA: ACM, 2006, pp. 113–114. [Online]. Available: http://doi.acm.org/10.1145/1145581.1145606

[5] C. Nebut, F. Fleurey, Y. L. Traon, and J. marc Jézéquel, "Automatic test generation: A use case driven approach," *IEEE Transactions on Software Engineering*, vol. 32, pp. 140–155, 2006. [Online]. Available: http://dx.doi.org/10.1109/TSE.2006.22

[6] E. Mendes Bizerra Junior, D. Silva Silveira, M. Lencastre Pinheiro Menezes Cruz, and F. Araujo Wanderley, "A method for generation of tests instances of models from business rules expressed in ocl," *Latin America Transactions, IEEE (Revista IEEE America Latina)*, vol. 10, no. 5, pp. 2105–2111, 2012. [Online]. Available: http://dx.doi.org//10.1109/TLA.2012.6362355

[7] B. C. and M. A., "A framework for gui testing based on use case design," in *Proceedings of the 2010 Third International Conference on Software Testing, Verification, and Validation Workshops*, ser. ICSTW '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 252–259. [Online]. Available: http://dx.doi.org/10.1109/ICSTW.2010.37

[8] K. Dyrkorn and F. Wathne, "Automated testing of non-functional requirements," in *Companion to the 23rd ACM SIGPLAN conference on Object-oriented programming systems languages and applications*, ser. OOPSLA Companion '08. New York, NY, USA: ACM, 2008, pp. 719–720. [Online]. Available: http://doi.acm.org/10.1145/1449814.1449828

[9] S. R. Dalal and et al., "Model-based testing in practice," in *Proceedings of the 21st international conference on Software engineering*, ser. ICSE '99. New York, NY, USA: ACM, 1999, pp. 285–294. [Online]. Available: http://doi.acm.org/10.1145/302405.302640

[10] T. Straszak and M. Śmiałek, "Acceptance test generation based on detailed use case models," in *Advances in Software Development*, J. Swacha, Ed. PIPS, 2013, pp. 116–126.

[11] H. Kaindl, M. Śmiałek, P. Wagner, and et al., "Requirements specification language definition," ReDSeeDS Project, Project Deliverable D2.4.2, 2009, www.redseeds.eu.

[12] W. Nowakowski and et al., "Requirements-level language and tools for capturing software system essence," *Computer Science and Information Systems*, vol. 10, no. 4, pp. 1499–1524, 2013. [Online]. Available: http://dx.doi.org/10.2298/CSIS121210062N

[13] D. Steinberg, F. Budinsky, M. Paternostro, and E. Merks, *EMF: Eclipse Modeling Framework 2.0*, 2nd ed. Addison-Wesley Professional, 2009.

[14] I. M. Graham, "Task scripts, use cases and scenarios in object-oriented analysis," *Object-Oriented Systems*, vol. 3, no. 3, pp. 123–142, 1996.

[15] M. Śmiałek and et al., "Complementary use case scenario representations based on domain vocabularies," *Lecture Notes in Computer Science*, vol. 4735, pp. 544–558, 2007, mODELS'07. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-75209-7_37

[16] "UML Testing Profile (UTP) Version 1.2," Object Management Group, Tech. Rep. formal/2013-04-03, Sep. 2012. [Online]. Available: http://www.omg.org/spec/UTP/1.2/

[17] M. Smialek and T. Straszak, "Facilitating transition from requirements to code with the ReDSeeDS tool," in *Requirements Engineering Conference (RE), 2012 20th IEEE International*. IEEE, 2012, pp. 321–322. [Online]. Available: http://dx.doi.org/10.1109/RE.2012.6345825

[18] M. Smialek and et al., "Translation of use case scenarios to Java code," *Computer Science*, vol. 13, no. 4, pp. 35–52, 2012. [Online]. Available: http://dx.doi.org/10.7494/csci.2012.13.4.35

[19] M. Smialek, W. Nowakowski, N. Jarzebowski, and A. Ambroziewicz, "From use cases and their relationships to code," in *MoDRE*. IEEE, 2012, pp. 9–18. [Online]. Available: http://dx.doi.org/10.1109/MoDRE.2012.6360084

[20] Beck, *Test Driven Development: By Example*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2002.

[21] D. North. (2006, Mar.) Introducing BDD. [Online]. Available: http://dannorth.net/introducing-bdd/

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1575–1584
DOI: 10.15439/2014F316
ACSIS, Vol. 2

Abstractions on Test Design Techniques

Marc-Florian Wendland
Systems Quality Center
Fraunhofer Institute FOKUS
Berlin, Germany
marc-florian.wendland@fokus.fraunhofer.de

***Abstract*—Automated test design is an approach to test design in which automata are utilized for generating test artifacts such as test cases and test data from a formal test basis, most often called test model. A test generator operates on such a test model to meet a certain test coverage goal. In the plethora of the approaches, tools and standards for model-based test design, the test design techniques to be applied and test coverage goals to be met are not part of the test model, which may easily lead to difficulties regarding comprehensibility and repeatability of the test design process. This paper analyzes current approaches to and languages for automated model-based test design and shows that they are lacking important information about the applied test design techniques. Based on this analysis, we propose to introduce another layer of abstraction for expressing test design techniques in a tool-independent, yet generic way.**

Keywords- Model-based testing (MBT), test generation, automated test design, test design techniques, UML Testing Profile (UTP)

I. INTRODUCTION

THE degree of automation in industrial software testing was consequently raised within the last two decades. In the 1990s, efforts had been undertaken to increase the degree of automation for test execution, resulting in today's accepted technologies like keyword-driven testing [3]. Standards have been built upon this principle like TTCN-3[1] or the newly developed ISO 29119 [9] standards family. With the wide-spread acceptance of UML in the late 1990s and the advent of UML 2 early 2000s, the idea to automate also parts of the test design activities was pursued in research and industry. The outcome of these efforts is what is known today as model-based testing (MBT) and test generation. Both, automation of test execution and automation of test design rely on abstraction of irrelevant details. Of course, when it comes down to actually test execution, the abstracted details need to be provided, but this is commonly accepted to be pertinent and indispensable. In keyword-driven testing approaches, the so called adaptation layer is in charge of making logical test cases executable [21].

[1] http://www.ttcn-3.org

The UML Testing Profile (UTP) [12] is a modeling language for MBT approaches based on the UML. It is the first industry-driven, standardized modeling language for MBT. It was adopted by the Object Management Group (OMG) as far back as 2003 and is currently under major revision. In addition, the European Telecommunications Standardizations Institute (ETSI) has funded efforts to develop its own modeling language for MBT, called Test Description Language (TDL). Thus, two important technical standardization bodies offer languages to build MBT methodologies upon.

Interestingly, none of the above mentioned standards provides concepts to specify the test design techniques that shall be applied for test generation. This seems inconsistent, since one of the most communicated benefits of MBT is automated generation of test artifacts and the increased systematics, comprehensibility and repeatability of the test design process [20]. Until today, there is no generally accepted approach found in the literature how test design techniques for model-based test generation shall be specified the best. In fact, almost every test generator provides its own proprietary configuration for specifying the test coverage goal. This lead to several issues regarding comprehensibility and repeatability the automated test design activities. Moreover, the exchangeability of test generators on models, even of the same modeling language, becomes risky since it bears a great potential for loss of relevant knowledge.

This paper addresses the abstraction from technical, tool-dependent representations of test design techniques by providing an extensible language framework for specifying tool-independent test design techniques that can be shared across multiple test generators. This step is a consequent evolution of the automation through abstraction principle already applied in keyword-driven testing or test generation. The contributions of the work are:

- A thorough analysis of current approaches to model-based test generation.
- The development of a conceptual model of test design based on the ISO 29110 standard. The conceptual model builds the foundation on which the abstractions of test design techniques rely on.

978-83-60810-58-3/$25.00

- The refinement of the conceptual model with an approach to test generation that is motivated by means of directives and strategies.
- Provision of an extensible, yet flexible UML profile-based implementation of the refined conceptual model as an extension to the UTP

The remainder of this paper is structured as follows:

Section 2 describes the problems of today's approaches to automated test design from the viewpoint of comprehensibility and repeatability. Section 3 elaborates the conceptual model of test design and the refinement towards test design directives and test design strategies. Section 4 discusses the extension of the UTP with test design directives and test design strategies. Section 5 demonstrates the feasibility of our approach by applying it to two non-commercial test generators, i.e., the Spec Explorer and Graphwalker. Section 6 presents the work related to ours. Finally, section 7 concludes our work and highlights future work in that context.

II. Problem Statement

Most of the today's model-based test generators are able to work on UML or derivatives. In this paper, we employ the SpecExplorer[2] and the Graphwalker[3] generation engine that do not operate directly on UML, but on closely related concepts. The SpecExplorer input is actually based on a textual representation of an Abstract State Machine (ASM) [7] which is called Spec#. On contrast, the input for Graphwalker is GraphML, a XML format for describing graph structures. Both test generators can operate on graph structures, although the input format is different. These input formats can be derived from UML behaviors, though. In the last years, we have in particular integrated these two test generators with UTP, which allows us to generate the required input format from the very same UTP model for both generators ([23], [22]). Our overall vision is to integrate a wide variety of test generators with UTP to counteract the broadening of proprietary, yet technically incompatible modeling languages.

Hence, the following problem statement was identified in the context of MBT with UTP, so is the technical solution presented in this paper. The conceptual solution, however, is not bound to any particular modeling language.

A. Test Design Techniques in MBT

If we consider the commonly understood advantages of MBT – such as efficient solutions for test design, increasing the degree of automation, prevention of loss of knowledge by using (semi-)formal models, more systematic and, even most important, repeatability of test case derivation and self-explanatory of test specifications ([20],[6]) – MBT comes along with an indispensable change of paradigms for testing activities. The most central artifact in an MBT approach should be the model itself, so test processes have to move from a document-centric to a model-centric paradigm. A model that describes test-relevant information from a tester's point of view is called *test model*. A test model is a "… model that specifies various testing aspects, such as test objectives, test plans, test architecture, test cases, test data etc." [12]. The UTP, in combination with UML, provides a test engineer with numerous possibilities for building test models, since it offers the expressiveness of UML and amends it with test-specific artifacts. Thus, UTP is deemed suitable to support the change to the model-centric paradigm where test models are single source of truth. Even though not as a dedicated concept, UTP allows for modeling the inputs for test generation just by relying on the underlying UML concepts. Inputs to test generation are referred to as test model in the ISO 29119 terminology as well. To avoid confusion with the much broader understanding of a test model given by the UTP specification, we henceforth refer to inputs to test generation as test design models. ISO 29119 defines test design as all activities in a test process that actually derive test cases, test data and test configuration from test conditions. This derivation may be carried out automatically or manually. The term automated test design is commonly known as test generation.

Even though UTP is an expressive language, it does not offer concepts to specify the test design techniques that shall be applied for deriving test artifact. If we consider the before mentioned benefits of MBT, first and foremost test generation, it is most surprising that one of the most important information for automating the test design activities is missing in the test model: The information about which test design techniques shall be carried out on the test design model by the test generator. In other words, the information why a certain test artifact has been generated is not part of the test model. As a matter of fact, today's test generators define their own proprietary presentation of test design techniques that resides within the tool. It complicates, however, the application of an entire model-centric approach to testing, for it does not allow integrating all the required information into the test model. This can have severe implications, since it may easily happen that the applied test generator shall be replaced, for whatever reason, while the defined test design techniques shall be retained. If this happened and access to the previous test generator is not given any longer, the information where a certain test artifact originated from in the first place is lost.

Figure 1 (a) illustrates this problem in a three-layered-approach. The domain layer encodes the test model (more specific, the test design model), which it is completely decoupled from a certain test generator and simply focus on the specification of a system under test. The engine layer, in

[2] http://research.microsoft.com/en-us/projects/specexplorer/
[3] http://www.graphwalker.org

TABLE I.
INVESTIGATION OF TEST GENERATORS

Generator	Input	Output	Configuration	Paradigm	License
Spec Explorer	Spec# (C#)	NUnit Test Cases	Coord Language	1-way (not centric)	Free
MBTsuite	UML Activity / StateMachines	Proprietary	Settings in the tool	1-way (not centric)	Commercial
Conformiq	UML-approximated StateMachine	Proprietary	Settings in the tool	1-way (not centric)	Commercial
Graphwalker	GraphML	Sequence of Strings	Command line parameter	1-way (not centric)	Open Source
Tedeso	UML approximated-Activity	Proprietary	Settings in the tool	1-way (not centric)	Commercial
CertifyIT	UML and OCL	Proprietary	Settings in the tool	1-way (not centric)	Commercial

contrast, is the most fundamental component of a test generator. It is, conversely, completely decoupled from the domain layer and simply operates on its inputs, without taking into account where that input comes from. Both capabilities, complexity and underlying principle of the engine layer vary among test generators from powerful symbolic execution (like the SpecExplorer) to a simple traversal engine that simply operates on already explored graph structures such as finite state machines (like the Graphwalker). Regardless how powerful or sophisticated the generation engine actually is, it is necessary to transform the information encoded in the domain layer into a format understood by the generation layer. A mediator, an adaptation layer, is required. An adaptation layer is a tool-specific component that serves two purposes. First, it transforms (***γ(i)***) the input ***i*** (i.e., a test design model contained in the test model) into a format understood by the test generation engine. Secondly, it offers some kind of interface to the test analyst in order to configure the generation engine. We call this the configuration (***c***) of a test generator. The configuration contains the information of which test design technique shall be applied to the input i. For example, the SpecExplorer is configured with a proprietary language called coord. If the user wants to ensure a certain traversal order of events, he/she has to specify a regular expression over events. This is called a scenario in coord terminology. The semantics of the coord scenario matches with the standardized specification-based test design technique scenario testing in ISO 29119. This information ought to be part of the test model for it contains important test-relevant information to comprehend the automated test design activities. This holds true for other test design techniques and further test generators as well. TABLE I lists a few of the commercially relevant or prominent open source test generators that fit into our view of MBT. Interestingly, all of the investigated generators do offer proprietary means to configure the generation engine. Furthermore, none of these generators really follow the model-centric paradigm, since they simply employ the test design models for the purpose of test generation. There is commonly no feedback of the generated test cases into the test model in order to abide by the single source of truth principle as proposed and described by Wendland ([23], [24]). This is a situation we strive to improve with our work with our work.

B. *Abstractions of Test Design Techniques*

We propose to abstract from a tool-specific representation to tool-independent representation of test design techniques. Fig. 1 b) illustrates this abstraction. The configuration (shaded grey) are extracted from the tool-specific layer and abstracted (***α(i)***) to test design techniques that are part of the test model itself. The adaptation layer is still required to transform the input ***i*** (i.e., the test design model plus tool-independent test design techniques) into the input format ***γ(i)*** for the generation engine. As such, it is possible to share test design techniques across multiple test generators that provide an adapter for the utilized test design technique. With test design techniques removed from the realm of a specific test case generator and becoming part of the test model a more holistic approach to test design is being provided and the process gains transparency, repeatability and comprehensibility. Such an approach is in-line with the idea of abstraction for test generation as it is done for MBT [18], but for the specification of test design techniques instead of the test design model. This abstraction of test design techniques has not yet been discussed in the literature.

III. A CONCEPTUAL MODEL OF TEST DESIGN

The conceptual field of test design techniques is actually well known. Several academic and industrial literature deals with the application and formalization of test design techniques for different test design models. A good overview is given by Utting [21] and ISO 29119-4 [9]. Based on the concepts and terminology provided in ISO 29119, a conceptual model of test design can be deduced (see Fig. 2) which will be explained in great detail in the subsequent sections.

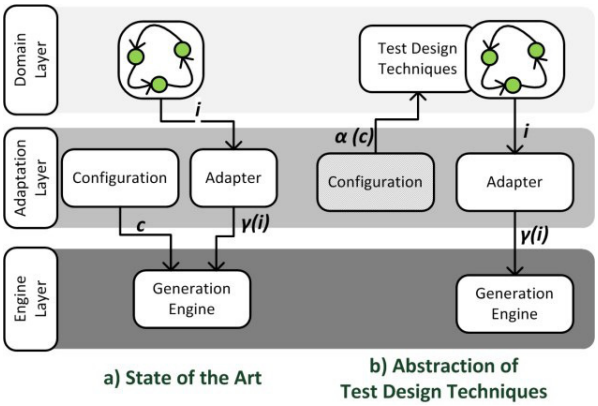


Fig. 1 Abstraction of test design techniques

A. The Principles of Test Design

The derivation of test artifacts is usually done by applying a test design technique, or ad-hoc if no systematic approach is applied. A test design technique is a method or a process, often supported by dedicated tooling that derives a set of test coverage items from an appropriate test design model. The test design model is obtained from the identified test conditions. According to ISO 29119, a test condition is a "testable aspect of a component or system, such as a function, transaction, feature, quality attribute, or structural element identified as a basis for testing." A test analyst utilizes the information accompanied with the test conditions to construct the test design model in whatever representation. This gave rise to Robert Binder's famous quote that testing is always model-based [1]. A test design model refers to a specification of the expected behavior of the system under test that is represented either as mental model, informal model, semi-formal model, or formal model.

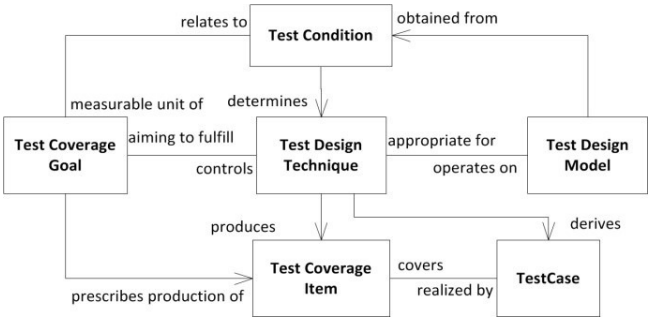


Fig. 2 A conceptual model of test design

As always with models [19] the test design model must be appropriate for the test design technique to be applied. An inappropriate model might not be able to produce an optimal result. The details of a test design model can usually be derived from the test conditions of the test basis.

There is a correlation between the test design technique and the test design model, however, since both are determined or influenced by the test conditions. For example, if the test condition indicates that the system under test might assume different states while operating, the test design model may result in a finite state machine (FSM) or similar. Consequently, a test design technique (like state-based test design) is most likely to be applied on this test design model.

A test design technique tries to fulfill a certain test coverage goal (the term used by ISO 29119 is *Suspension Criteria*, which is actually not that commonly understood). A test coverage goal determines the number and kind of test coverage items that have to be derived from a test design model and represented as test cases. The actual derivation activity might be carried out manually or in an automated manner.

A test coverage item is an "attribute or combination of attributes to be exercised by a test case that is derived from one or more test conditions by using a test design technique" [9]. The term test coverage item has been newly introduced by ISO 29119, thus, it is expected not to be fully understood at first sight. A test coverage item is usually been obtained from the test condition, and made been explicit (in a sense of that it can be used for coverage analysis etc.) through a test design technique. The following example discusses the subtle differences between test condition, test design model and test coverage item:

Let us assume there is a functional requirement that says the following: "If the On-Button is pushed and the **system** is *off*, the **system** shall be *energized*."

The bold words indicate the system under test, the italic words potential states the system under test shall assume und the underlined word an action that triggers a state change. According to ISO 29119, all the identifiable states (and the transitions and the events) encoded in the functional requirement represent the test conditions for that system under test. A state machine according to the test conditions would represent the test design model. As test design technique could be decided to be structural coverage criterion like transition coverage or similar. The test coverage goal would represent a measurable statement about what shall be covered after the test design technique has operated on the test design model. This might be one of Chow's N-Switch-Coverage 0 like full 1-Switch-Coverage (or transition-pair coverage). The test coverage items would eventually be represented by all transition pairs that have been derived by the test generator, and which are finally covered by test cases.

However, there are certain inaccuracies in the ISO 29119's test design concepts which are subsequently classified into three issues.

1) Test Coverage Calculation

At first, the term *test coverage*, defined by ISO 29119 as the "degree, expressed as a percentage, to which specified coverage items have been exercised by a test case or test cases", does not take the actual number of potentially

available test coverage items into account. According to the given definition of test coverage, the coverage would always be 100% since it is calculated on the actual derived test coverage items. What is missing is a calculation of all test coverage items that actually should be derived. Otherwise, it would be possible to state that 100% test coverage has been achieved, even though just 10% of all to be covered test conditions were actually covered. This is in particular relevant for model-based approaches to test design, for the test coverage items are usually not explicitly stored for further test case derivation, but rather automatically transformed into test cases by the test generator on the fly. This means that in today's model-based test generators the test cases always cover 100% of the derived test coverage items. This is just consequent, since the test coverage items were derived according to a specific test coverage goal, thus, the test design technique only selected those test coverage items (out of all potentially reachable test coverage items) that are required to fulfill the test coverage goal. Ending up in a situation where the eventually derived test cases would not cover 100% of the produced test coverage items would violate the whole idea of specifying test coverage goals.

2) Output of Test Design Techniques

Test design techniques do not only derive test cases, but also test data or test configurations. The test design process deals with the derivation of all aspects that are relevant for finally executing test cases. The test configuration or test interface (i.e., the identification of the system under tests, its interfaces and the communication channels among the test environment and the system under test) is a crucial part of each test case, when it comes down to execution. Same, of course, holds true for test data. In this paper we deal not with the generation of test configuration, however, test data generation is covered.

3) Test Design Techniques Are Not Monolithic

The concept of a test design technique, as defined and described by ISO 29119, needs to be further differentiated. In relevant, yet established standards for industrial software testing (such as ISO 29119, IEEE:829 and ISTQB) a test design technique is described as a monolithic concept. This is not the case, because the actual test derivation process consists of a number of techniques that represent distinguished course of actions to achieve test coverage. These courses of actions operate in combination with each other to derive the desired test coverage items. Those techniques contribute their semantics to the overall test design activity for a given test design model. Examples for well-known strategies are structural coverage criteria or equivalence partitioning, but also less obvious and rather implicit strategies like the naming of test cases or the physical structure or representation format of test cases. For example, the so called State-Transition test design technique might be based on an extended Finite State Machine (EFSM). Hence, the sole application of structural coverage criteria (like all-transition-coverage etc.) might not suffice to produce executable test cases, for EFSM may also deal with data inputs, guard conditions etc. By adding also data-related techniques (such as equivalence partitioning) to structural coverage criteria, it is possible to explore and unfold the EFSM into an FSM that ultimately represents the available test coverage items for finally deriving test cases. So, the discussion gives rise to the fact that the conceptual model of ISO 29119 regarding test design techniques shall be further differentiated to allow combining several test design techniques with each other in a systematic manner.

B. Towards Strategies and Directives

When further differentiating the concept of test design technique, a wider search beyond the field of testing of software systems seems to be appropriate. Based on what was discussed earlier, test design techniques are required to be grouped by different test design process, thus, they are reusable. Test design techniques are consequently decomposed into a thing that groups different techniques and the techniques themselves. This conceptual structure is similar to the Business Motivation Model (BMM) [14] concepts for directives and strategies (see Fig. 3). We adapt the terms directives and strategies for the scope test design.

The BMM provides a fine-grained conceptual model to analyze the visions, reasons and influencers of a business (or endeavor) in order to deduce its overall motivation. The BMM is enunciated in the Semantics of Business Vocabulary and Business Rules (SBVR) [15], a standard which is by the OMG, a standard which is adopted by the OMG to formalize a vocabulary for semantically documentation of an organization's business facts, plans and rules.

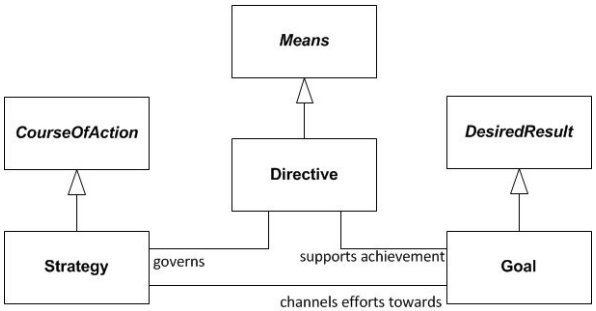


Fig. 3 Relations of Strategy, Directive and Goal

Notwithstanding the motivation for BMM to apply directives is outside of MBT in the first place, the BMM contains concepts and notations that can be beneficial to the realization of test design directives and test design strategies. According to BMM a *directive* is a means to achieve a certain goal. A *goal* is a statement about a state or condition of the endeavor to be brought about or sustained through appropriate *means*. Therefore, a directive (as specialized

means) utilizes (a set of) strategies that are governed by the directive to achieve the goal. A *strategy* channels efforts towards the achievement of that goal. This means that the same strategy can be utilized by different directives in order to achieve different goals, hence, strategies are reusable across directives. The notions of strategy, directive and goal can be mapped to the *business* test design. The BMM goal would map almost inherently to the ISO 29119 concept test coverage goal. As with a goal, a test coverage goal imposes a condition on the test design activity that need to be achieved in order to deem the test design activity completed. Since BMM strategies are the actual actions that need to be carried out in a controlled manner, the notation of BMM strategy stands for a single test design technique, such as equivalence partitioning, all-transition-coverage or similar. The directive, however, does not have a direct counterpart in the ISO 29119 conceptual model on test design. From a logical point of view, it is part of the test design technique concept even though not explicitly enunciated. We are going to leverage the notion of strategies and directives for the area of model-based test generation in order to refine the ISO 29119 conceptual model with test design strategies and directives that replaces the monolithic test design technique.

C. Refined Conceptual Model of Test Design

This section mitigates the conceptual imprecisions of the ISO 29119 conceptual model of test design by further differentiating the test design technique into test design directives and test design strategies. These notions are adopted from the BMM. Fig. 4 shows the redefined test design conceptual model in which the monolithic test design technique is split up into test design strategy and test design directive.

A *test design strategy* describes a single, yet combinable (thus, not isolated) technique to derive test coverage items from a certain test design model either in an automated manner (i.e., by using a test generator) or manually (i.e., performed by a test analyst). A test design strategy represents the logic of a certain test design technique (such as structural coverage criteria or equivalence partitioning) in a tool- and methodology-independent way and is understood as logical instructions for the entity that finally carries out the test derivation activity. Test design strategies are decoupled from the test design model, since the semantics of a test design strategy can be applied to various test design models. This gives rise to the fact that test design strategies can be reused across different test design models. This fits with the more general notation of a strategy that can be utilized by several means. The intrinsic semantics of a test design strategy, however, needs always be interpreted and applied within the context of a test design model. According to and slightly adapted from the BMM, this context is identified by a test design directive.

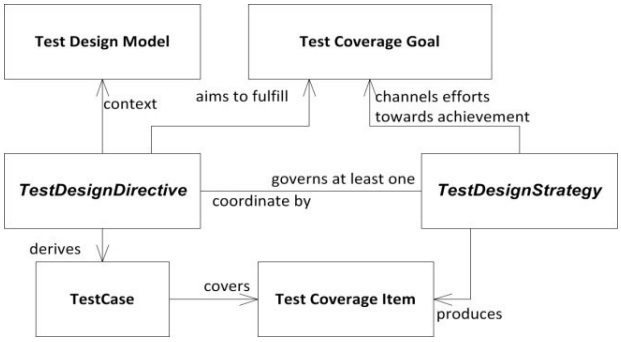


Fig. 4 Redefined conceptual model on test design

A *test design directive* governs an arbitrary number of test design strategies that a certain test derivation entity has to obey to within the context of a test design model. A test design directive is in charge of achieving the test coverage goal. Therefore, it assembles appropriately deemed test design strategies to eventual fulfill the test coverage goal. The assembled test design strategies, however, channel the efforts of their intrinsic semantics towards the achievement of the test coverage goal. The test coverage items that are produced by test design strategies are always fully covered, thus, they are reduced to a pure transient concept. According to Fig. 1 b), the test design directive will be passed to the tool-specific adaptation layer, since it is the test design directive that has access to all required information. At first, it specifies the test design models out of which test artifacts shall be generated. Next, it governs the test design strategies that shall operate on the test design models. In the next sections, we show an implementation of the conceptual model as UML profile.

IV. A Language Framework for Test Design

The implementation of the refined ISO 29119 conceptual model on test design was from the very first idea incepted as an extension of the UTP. This mitigates one of the most obvious deficiencies of the UTP and allows the creation of fully comprehensible test models. The extension is kept most flexible, so that new test design directives and test design strategies can be easily incorporated.

A. Realization as UML Profile

Since the test design framework has to be kept minimalistic and left open for multiple modeling and testing methodologies, it is important to find a means to not being too intrusive while defining the framework. Fortunately, a UML profile grants all capabilities of MOF-based metamodels, so it is possible to utilize the concepts of derived unions and subsetting properties. Fig. 5 shows the elements of the language framework.

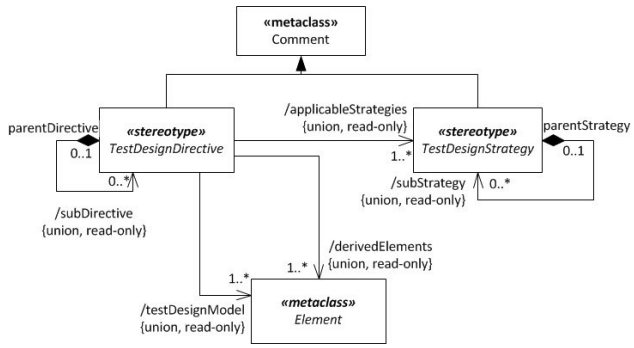


Fig. 5 A UML profile of test design

Both stereotypes test design strategy and test design directive extend the UML metaclass *Comment.* Comment is a semantic-free element of UML, usually used to add notes or documentations to other elements. Applying a test design strategy stereotype onto a Comment turns the Comment into a test design strategy. Same, of course, holds true for test design directive. According to the BMM and the refined ISO 29119 conceptual model of test design, a test design directives establishes associations to one or multiple test design strategies. This means that a test design directive might be composed of several test design strategies. Both test design strategy and test design directive are abstract concepts which means they need to be further specialized in order to be applicable. This is, of course, on purpose, since it is infeasible to foresee each and every test design directive or test design strategy of every existing or new methodology in the future. Thus, the language framework must under all circumstances not restrict a test analyst if he/she wants to define new test design strategies or directives.

A test design directive establishes two further associations to the most fundamental metaclass in the UML metamodel, i.e., *Element*. One association captures the semantics of identifying the allowed test design models for a specialized test design directive, the other represents a trace link to the actually derived elements. The language is again left open as much as possible, because there is no reason why a certain elements of UML should be spared out as test design model or generated element. As with the associations to test design strategies, the specializations of test design directives are responsible to fill these derived unions with reasonable information.

The main benefit and most powerful characteristics of the language framework is the fact that an appropriate tooling can access all available information of any existing specialized test design strategy or test design directive by using the MOF reflection capability at runtime (i.e., modeling time). This enables tool vendors to build a complete and sophisticated tooling on basis of this minimalistic framework without taking care of future extensions. As soon as new specializations of test design strategies and directives are made accessible to the modeling tool, they can be utilized at once.

B. Libraries of Test Design Strategies

As already stated in section *Towards Strategies and Directives*, test design strategies bear the potential to be leveraged by different test design directives. It is not possible to navigate from a test design strategy to a test design library explicitly (it has to be said that the MOF capabilities allows the navigation of so called non-navigable association ends – this is an essential precondition for the language frameworks adaptability). A test design strategy actually does not have to be aware with which test design directives it is associated with, since the evaluation and realization of the semantics of a test design strategy in the context of a test design model is done by the adaptation layer of a test generator. On domain level, it is sufficient to limit the combination of test design strategies in the context of test design directives. Since test design strategies are more or less autarkic concepts, it is possible to develop libraries of well-known and accepted test design strategies and to make them accessible to the developer of a new test design directive.

Predefined libraries for test design strategies make in particular sense when it is prescribed (by a standard or company-wide test strategy or policy) which test design strategies are permitted within the test process. By building libraries (and appropriate tool support), test managers or test analyst are able to define a canonical list of test design strategies that shall be exclusively used. This counteracts, for example, violations of such test strategies and fosters, at the same time, consistency among different test design activities of the same test process.

C. Integration of Test Generator Capabilities

In order to integrate on the language framework, it is required that each integrated test generator need to propagate its meta-information into the language framework. The most important meta-information [5], of course, is the information about which test design strategies the test generator can realize via its adaptation layer. The language framework provides all required information to enunciate these meta-information. Fig. 6 illustrates the integration of the SpecExplorer and Graphwalker with the language framework.

D. Provision of Test Generation Services

As a proof-of-concept, we have combined the language framework with the existing UTP and integrated it into our academic test modeling environment Fokus!MBT [24]. As test generators we employed the previously mentioned Graphwalker and SpecExplorer. As part of the tool integration, we created two test design directives (one for each test generator) with appropriate test design strategies. The result can be seen in Fig. 6.

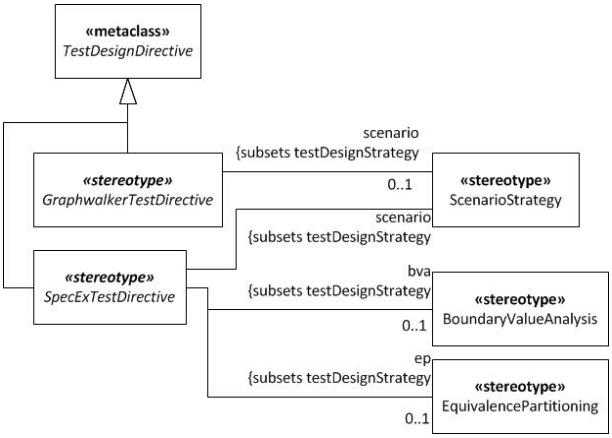


Fig. 6 Definition of test design directives

In Fokus!MBT, a test design directive is considered to identify an accessible test generation service. A test generation service is in its simplest essence a single test generator that is integrated with the language framework. However, a test generation service might also consist of a set of test generators realizing the associated test design strategies. The term test generation service abstracts from the physical representation of the test generation approach.

A test design directive is tightly bound to a test generation service, which in turn comes along with an appropriate adaptation layer. Since Fokus!MBT is based on Eclipse, we utilize the extension point mechanism of Eclipse to identify a test generation service by means of the test design directive. The input validation process is crucial, since only test design models the adapter can map into the input format of its respective engine are allowed to be passed to the engine. In Fokus!MBT, a specialized test design directive can be associated with a set of validation constraints. What is not shown in this picture is the case, if the input validation detects constraint violations. If this happens, the tester is notified about the details of the input validation process.

V. EVALUATION

Let us assume we have an EFSM as test design model that consists of five states and five transitions and a global state variable *x* of type Integer. The according state machine is defined by the following state-transition-table (TABLE II):

TABLE II.
EXAMPLE EFSM

Source	Input (i)	Action	Output (o)	Target
initial	-	x := 0	-	*s1*
s1	i ≥ 0 and i ≤ 5	x:= x+i	x	*s2*
s1	i = 1000	x:= x+i	x	*s3*
s2	i = 3	x:= x-i	x	*accepting*
s3	-	x:= x-5	x	*accepting*

This implies that there are two paths available, which are *sc1* = {*initial→s1, s1→s2, s2→accepting*} and *sc2* = {*initial→s1, s1→s3, s3→accepting*}, where → denote the transition. There are no further guard conditions defined on the transitions for the sake of simplicity. Let us further assume that we want to apply the scenario test design strategy with both the Graphwalker and SpecExplorer engine, where the desired scenario is *sc1*. In addition, we want to apply boundary value analysis, if data generation is possible. So, we define the following two test design directives:

```
SpecExTestDirective 'specex' :=
      testDesignModel = EFSM
      testDesignStrategies :={
       scenario {events = sc1} and
       boundary value analysis {values = 1}
      }

GraphwalkerTestDirective 'gw' :=
      testDesignModel := EFSM
      testDesignStrategies :={
       scenario{events=sc1}
      }
```

The SpecExplorer is capable of generating inputs based on given constraints. The transition *S1→S2* has a constraint defined that the input value must be in a range [0..5]. In Fokus!MBT, range constraints are modeled as UML::Interval. As said before, the configuration of the SpecExplorer engine I s done via *coord*. The respective adaptation layer transforms the test design directive and the referenced test design model into the coord representation. The scenario test design directive is transformed into a SpecExplorer scenario, which is then further used for the exploration. The transformation of the EFSM into the ASM is not part of the paper. Please refer to the DOME case study of the REMICS case study.

In contrast, the *GraphwalkerTestDirective* does not have a test design strategy for boundary value analysis defined, since the generation engine is not capable of generating input values.

After executing bot test generation services, the following test coverage items have been derived:

- SpecExplorer – 2 Test Cases

  #1: *initial→s1(-/-),s1→s2(0/0), s2→s3(3/-2)*
  #2: *initial→s1(-/-),s1→s2(5/5), s2→s3(3/2)*
- Graphwalker – 1 Test Case

  #1:initial→s1,s1→s2, s2→s3

This result is not surprising since the Graphwalker is not able to generate any data at all. Depending on the applied methodology to test generation, this might be not a problem. As a matter of fact, the resulting test cases are mostly of abstract nature, i.e., they are lacking concrete data. These

data information need to be applied later on in order to make the test cases executable.

VI. Related Work

The term *test directive* was firstly introduced and defined by [2] in her PhD as "… a collection of test-specific information which, when combined with the system model, derives a test model." The PhD of Dai, however, dealt solely with the derivation of test configurations from existing system models. Therefore, she treated test directives as specifications of single model-to-model transformation rules in a platform-independent manner. The set of combined test directives yielded a complete transformation which generated a test configuration. Our approach, in contrast, deals with the generation of test cases and test data in addition to test configuration.

Friske [5] presented an early idea to specify test coverage goals in generator-independent manner. He leveraged OCL in the context of UML State Machines to define structural coverage criteria like transition coverage or 1-Switch coverage 0. They specified two requirements for the integration of arbitrary test generators into their OCL framework. Firstly, the test generator to be integrated had to follow a two-step generation process. The first step has to generate the test goals (or test coverage items as defined by ISO 29119, the second step to derive test cases from these test goals. The second requirement requests that each generator publishes its metadata regarding its generation capabilities. The idea is quite similar to what we proposed in our work, however, there is no proof-of-concept described that actually realizes the pure theoretically OCL-based approach. Furthermore, it is not clear how strategies for automated test design can be integrated that are not pertinent to the generation of test coverage items (or test goals) but to the derivation of test cases from these test coverage items. For example, the strategies how generated test cases shall be named or finally realized in a compositional sense (decompose test case behavior into several separated pieces of behavior or build monolithic test case behaviors) are essential decision in each test design activity. In our approach, such adjacent strategies can be integrated in the same way as any other strategy. Finally, our experiences with the application of model-based testing techniques in the industry have shown that most testers are not familiar with formal languages like OCL.

Fourneret et al. [4] have presented an approach to model-based security verification and testing of smart-cards based on a dedicated language for security properties. The language is integrated with the UMLsec approach [10] and allows test engineers specifying what a security test generation shall generate. Thus, it enables domain experts to express the strategies for the test generation approach in a non-technical way. A proof-of-concept was done with the commercial CertifyIt [4] test generation tool. The specific test design strategies of that language could be integrated with the test design directives framework in order to make them applicable outside of UMLsec.

Wendland has used of test directives in the context of MBT based on a proprietary metamodel for testing purposes [22]. This work can be seen as the very first approach for an abstracted view on test design techniques based on test directives. Back in those days, the notion of test strategies was not used, though. In contrast, to the work presented in this paper, the former approach tried to develop a modeling framework that was intended to condense the indivisible parts of test design techniques in order to construct a concrete test design technique dynamically. This was much too complex, and yet not applicable.

VII. Conclusion and Future Work

In this paper we had dealt with abstractions of test design techniques in the realm of automated test design based in the context of MBT. We had clearly identified the problem statement that today's approaches to MBT are not that comprehensible, repeatable and flexible as the existing academic and industrial literature in the last decade promised to be. In fact, a full understanding of the automated test design activities within an MBT approach requires access to the applied test generator(s), since they keep the knowledge of the applied test design techniques. This is the main challenge we addressed in this paper. The overall aim is to finally capture all test-relevant information independent of any implementation within the test model itself. This is the only way to ensure a seamless change of paradigms from document-centric to model-centric testing activities. Therefore, we analyzed the conceptual domain of test design compliant with the ISO 29119 standard. We described three issues of the ISO 29119 conceptual model on test design and mitigated them by introducing the notions of test design strategies and test design directives. These pure conceptual notions were subsequently mapped onto a concrete language framework realized as a UML profile. This profile is a most minimalistic realization of the conceptual model and integrates well with the UTP. The reason to go for a UML profile was a natural decision in our work in order to fill the conceptual gap of UTP regarding the specification of test design techniques. Finally, as a proof-of-concept we have described and illustrated how the language framework can be integrated into modeling environments. We therefore used our academic test modeling environment Fokus!MBT[5]. The illustrated proof-of-concept was kept minimal since it was not the scope of the paper to describe a full case study.

The experiences we made and results we obtained from the integration of SpecExplorer and Graphwalker on UTP

[4] http://www.smartesting.com
[5] http://www.fokusmbt.com

and the suggested language framework gave confidence that every of the listed test generators in this paper can be integrated with this approach. Another side-effect of the language framework is that the concatenation of different test generations engines can be achieved rather easily [11].

Future work in the realm of standardization will in particular be channeled towards the new major revision of the UML Testing Profile, which is currently undergoing. As said before, the absence of a facility to specify test design techniques on model level was already complained about. The requirements document of the new UTP version [16] requests precisely such a facility. Our contribution in this area will be minimal language framework presented in section *A Language Framework for Test Design*.

While writing this paper, we are already working on the integration of a usage-based test generator [8] and behavioral and data fuzzing generator [18] with the proposed language framework in the context of the EU project MIDAS[6] w

To conclude the achievements of our work, we have shown that different test generators are able to be integrated on an abstracted representation of commonly accepted (or proprietary) test design techniques. In addition, we think this language framework is flexible to allow for a fine-grained adjustment of applied test design techniques. Our work fills the conceptual gap of MBT approaches in this regards and, thus, allows for a more holistic and comprehensible approach for automated test design based on (semi-)formal models.

ACKNOWLEDGMENT

Most parts of the work presented in this paper were funded by the EU projects REMICS (no. 257793) and MIDAS (no. 318786).

REFERENCES

[1] Binder, R., Testing Object-Oriented Systems: Models, Patterns, and Tools. Addison Wesley, 1999.

Chow, Tsun S.: Testing Software Design Modeled by Finite-State Machines. IEEE Transactions on Software Engineering, Vol SE-4, No. 3, 1978. http://dx.doi.org/ 10.1109/TSE.1978.231496

[2] Dai, Zhen Ru: An Approach to Model-Driven Testing – Functional and Real-Time Testing with UML 2.0, U2TP and TTCN-3. Dissertation at the TU Berlin, 2006.

[3] Foster, M. and Graham, D., Software Test Automation. Addison-Wesley Professionals, 1999. ISBN: 978-0201331400.

[4] Fourneret, E. et al., "Model-Based Security Verification and Testing for Smart-cards," ares, pp.272-279, 2011 Sixth International Conference on Availability, Reliability and Security, 2011. http://dx.doi.org/10.1109/ARES.2011.46

[5] Friske, Mario; Schlingloff, Bernd-Holger; Weißleder, Stephan, Composition of Model-based Test Coverage Criteria, MBEES, 2008. 87-94.

[6] Grieskamp, W., Model-Based Testing in the Field: Lessons Learned, *Lecture Notes in Informatics*, Vol P-94 (2006), Pages 189- 196.

[7] Gurevich, Y., Evolving Algebras, 1993: Lipari Guide, *Specification and Validation Methods*, pages 9–36. Oxford University Press, 1995. http://dx.doi.org/10.1007/978-3-540-74284-5

[8] Herbold, S., Grabowski, J., Waack, S. (2011). A Model for Usage-based Testing of Event-driven Software. 3rd International Workshop on Model-based Verification & Validation: From Research to Practice (MVV). http://dx.doi.org/10.1109/TSE.2010.12

[9] International Organisation for Standardisation (ISO): ISO/IEC 29119, Software Testing Standard, http://www.softwaretestingstandard.org

[10] Jürjens, Jan, Secure Systems Development with UML. Springer-Verlag, 2005. http://dx.doi.org/10.1007/b137706

[11] Lackner, Hartmut; Schlingloff , Holger, Modeling for automated test generation - a comparison. MBEES 2012, p 57-70, 2012.

[12] Object Management Group (OMG): UML Testing Profile. URL: http://www.omg.org/spec/UTP

[13] Object Management Group (OMG): Unified Modeling Language. URL: http://www.omg.org/spec/UML

[14] Object Management Group (OMG): Business Motivation Model (BMM). http://www.omg.org/spec/BMM

[15] Object Management Group (OMG): Semantics of Business Vocabulary and Business Rules (SBVR). http://www.omg.org/spec/SBVR

[16] Object Management Group (OMG): UML Testing Profile 2, Request for Proposal (RFP), document number: ad/2013-12-08.

[17] Pretschner, A. and Philipps, J., Methodological Issues in Model-Based Testing. In: *Model-Based Testing of Reactive Systems*. Springer, 2004, Pages 281-29. http://dx.doi.org/10.1007/ 11498490_13

[18] Schneider, M., Großmann, J., Tcholtchev, N., Schieferdecker, I., & Pietschker, A. (2013). Behavioral fuzzing operators for UML sequence diagrams. In Ø. Haugen, R. Reed, & R. Gotzhein (Eds.) System Analysis and Modeling: Theory and Practice, vol. 7744 of Lecture Notes in Computer Science, (pp. 88–104). Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-36757-1_6

[19] Stachowiak, H.: Allgemeine Modelltheorie, Springer, Wien, 1973, ISBN-10: 3-211-81106-0. http://dx.doi.org/10.1007/978-3-7091-8327-4

[20] Utting, M.; Pretschner, A., Legeard, B.: A Taxonomy of Model-Based Testing. ISSN 1170-487X, 2006. http://www.cs.waikato.ac.nz/pubs/wp/2006/uow-cs-wp-2006-04.pdf. http://dx.doi.org/10.1002/stvr.456

[21] Utting, Mark; Legeard, Bruno, Practical Model-based testing – A Tools Approach, Elsevier, 2007.

[22] Wendland, M.-F., Großmann, J, and Hoffmann, A., "Establishing a Service-Oriented Tool Chain for the Development of Domain-Independent MBT Scenarios," 17th IEEE International Conference and Workshops on Engineering of Computer-Based Systems ECBS 2010, Oxford, England, IEEE, 2010, pp. 329-334. http://dx.doi.org/10.1109/ECBS.2010.47

[23] Wendland, Marc-Florian et al., Model-based testing in legacy software modernization: an experience report, in Proceedings of the 2013 International Workshop on Joining AcadeMiA and Industry Contributions to testing Automation (JAMAICA 2013). JAMAICA'13, July 15, 2013, Lugano, Switzerland. ACM 978-1-4503-2161-7/13/07. http://dx.doi.org/10.1145/2489280.2489291

[24] Wendland, M.-F.; Hoffmann, A., and Schieferdecker, I., Fokus!MBT - a multi-paradigmatic test modeling environment, in Proceedings of the workshop on ACadeMics Tooling with Eclipse (ACME 2013), ACME'13, Montpellier, France, ACM 978-1-4503-2036-8/13/07. http://dx.doi.org/10.1145/2491279.2491282

[6] http://www.midas-project.eu

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1585–1590
DOI: 10.15439/2014F263
ACSIS, Vol. 2

Automating Test Case Design within the Classification Tree Editor

Ute Zeppetzauer
Berner and Mattner Systemtechnik GmbH
Munich, Germany
Email: ute.zeppetzauer@berner-mattner.com

Peter M. Kruse
Berner and Mattner Systemtechnik GmbH
Berlin, Germany
Email: peter.kruse@berner-mattner.com

***Abstract*—This paper describes how the proven test design technique of the classification tree method is extended within the classification tree editor in order to contribute to current test design matters. The classification tree editor not only provides the tooling to use the method but also to apply new and helpful features in the test design process. This includes the automatic generation of test cases and test sequences according to desired test depth and focus, automated boundary value analysis, various tool couplings to integrate in each individual test process and supporting features like test evaluation or test coverage analysis amongst others.**

***Keywords*-classification tree method; classification tree editor; combinatorial interaction testing;**

I. Introduction

THE classification tree method [1] as well as the editor [2] have been developed at Daimler's research department for software technology in the 90ties. In the last 20 years both, method and tool became proven in practice around the world.

The classification tree method as a black box test design technique was introduced by Matthias Grochtmann and Klaus Grimm in 1993 [1]. The basic idea of the classification tree method is to separate the input data characteristics of the system under test into different classes that directly reflect the relevant test scenarios. The descriptive method covers the categories of test case design needed in industry, such as *equivalence class tests*, *boundary value analysis* [3], *interface testing*, as well as *combinatorial* [4] and *statistical testing* [5].

In the last 6 years where Berner and Mattner took over the further development of the classification tree method and the editor, it got extended by academic research results and industrial input. This includes additional statistical testing methods [5], a professional requirements tracing [6], a synchronization to test management, test evaluation, coverage analysis or support for website testing [7]. The classification tree editor supports thus the tester in the test design phase. It offers multiple functions to structure the test problem, to systematically generate test scenarios and to do all this within the scope of the testers environment.

The outline of this paper is as follows: Section II introduces the classification tree method, while Section III shows how it will prove to serve testers needs for a more automated way in designing test cases within the classification tree editor. Section IV the integration into the test process with requirements and test management tools are introduced. Section V details Excel import and combinatorial test coverage analysis, test evaluation is given in Section VI. Related work can be found in Section VII, while conclusion is drawn in Section VIII.

II. Test Case Design using the Classification Tree Method

A. How to create a classification tree

The basic idea of the classification tree method is to separate the input data characteristics of the system under test into different classes that directly reflect the relevant test scenarios (*classifications*) [1]. The main source of information is the specification of the system under test or a functional understanding of the system should no specification exist. Two significant steps must be performed to create a classification tree:

- The identification of relevant factors involves the determination and structuring of the relevant test scenarios and their interrelations to other parts of the system under test.
- The test specifications combine the relevant factors needed in order to achieve the desired test coverage.

The first phase begins with the identification of the classifications that are relevant for testing based on a functional description and understanding of the system under test. For each classification, there may be several input data that are all to be considered during testing, the *classes* of a classification.

Each classification should have a limited number of clearly defined input scenarios for the system under test. The input data characteristics are used to define the input ranges required for the relevant test scenarios. This is a classification in the mathematical sense: The set of all possible inputs is disjointly and completely classified into subsets - the classes. The separation based on the input data characteristics is done independently for each test scenario, and can therefore be done easily.

This approach applies the concept of equivalence class testing [3]: Testing with a data item that is representative of an equivalence class makes tests with all other elements of the same class redundant and therefore unnecessary because

978-83-60810-58-3/$25.00

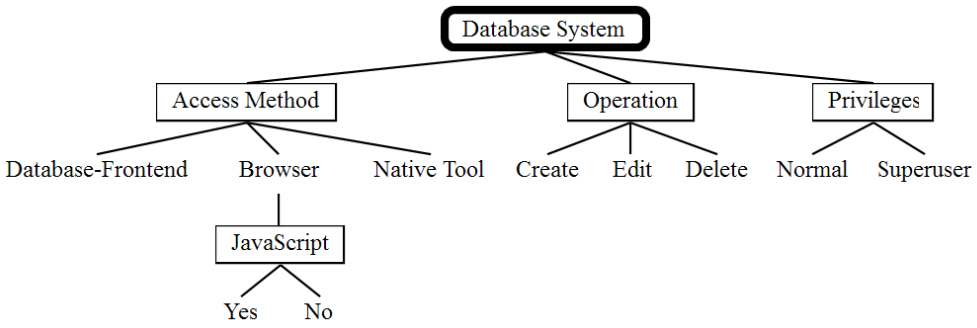


Fig. 1. Test Object *Database Management System*

there should be no difference in how they are handled during execution of the system under test.

Example. Figure 1 shows a classification tree for a database management system. Three aspects of interest (*Access Method*, *Operation* and *Privileges*) have been identified for the system under test. The classifications are partitioned into classes which represent the partitioning of the concrete input values. In our example the refinement aspect *JavaScript* is identified for the class *Browser* and it is divided further into two classes *Yes* and *No*.

B. Dependency Rules

Often, there are dependencies or constraints [8] between some classes of the classification tree. To overcome this design gap, the CTE offers to describe the relationships between the elements of the classification tree by defining dependency rules. The classification tree editor provides two mechanisms for defining dependency rules:

1) Logical dependency rules between the classes of a classification tree using propositional logic. The result is that test cases that would not fulfill the previously defined rules are not generated and vice versa [9].
2) Numerical dependencies between the classifications using logical and numerical operators. They serve to express mathematical dependencies between the elements of the classification tree [10].

C. Boundary Value Analysis

The boundary value analysis [3] is a helpful tool to run an analysis automatically with user specified input. There are two ways of applying the boundary value analysis within the CTE. The first is for an initial analysis, the second is to analyze and expand a classification tree with more possible parameters and intervals see Fig. 2.

For the initial analysis, the CTE supports to create parameters and intervals. For each interval it is possible to set the boundary borders so that CTE will create the corresponding boundary values. The classification tree is created from the specified values.

For the analysis or expansion of the tree, the CTE loads the existing data into its own data model. Then it is possible to add new parameters and intervals. The CTE then generates the boundary values and adds additional elements to the tree.

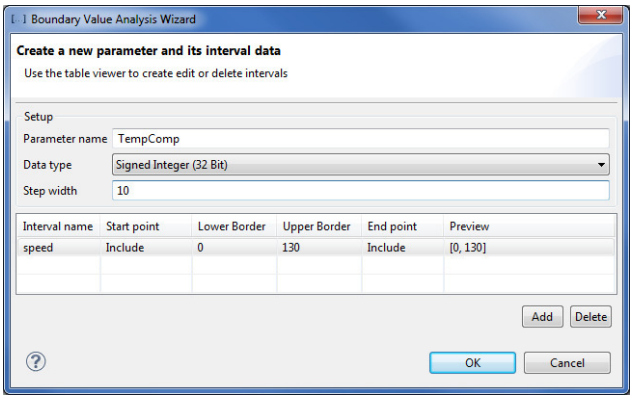


Fig. 2. Parameter and interval settings in CTE

III. Test Case Generation Facilities within the CTE

In practice, not every test generation mechanism is applicable to every test problem. Thus, there are numerous facilities available within the classification tree editor (CTE) in order to serve testers in many ways. The following test case generation mechanisms are implemented in the CTE.

- combinatorial test case generation
- prioritized test case generation
- test sequence generation

A. Combinatorial Test Case Generation

For combinatorial testing, the question is on how to achieve adequate test coverage without having to test too much. Complex software testing problems easily reach a maximum number of test cases of several billions. This is not feasible to handle. So a wise, structured and reproducible selection of test cases according to the current test problem saves effort, time and helps to actually better understand the testing focus. For this, combinatorial testing with all of its aspects is ideal. It brings variation to a test suite with a clear definition of the test intensity.

The CTE offers combinatorial test case generation facilities that support the structured variation [5], [9]. This includes pairwise combination, threewise combination, minimal test coverage and individual combinatorial rules.

The minimal combination $A + B$ ensures that each class of classification A and each class of classification B is considered in at least one test case.

The pairwise combination $(A_1, \ldots, A_n)$ ensures that each class of the classifications A_1 to A_n is combined pairwise with every other class in at least one test case.

The threewise combination $(A_1, \ldots, A_n)$ ensures that every possible combination of three classes of the different classifications A_1 to A_n is generated in at least one test case.

Higher strength interaction coverage can be achieved by practical testing approaches [11]. The key to unlocking better performance for higher strengths of interaction, seems to rely upon the use of constraints, which reduce the number of possibilities to be considered. In CTE, constraints are defined in terms of dependency rules.

Individual combination rules may be designed upon the needs of each test problem. All standard operators are available for that.

Logical expressions are used to formulate dependency rules. The CTE XL allows to specify any kind of logical dependency rules, containing {*AND*, *OR*, *NOT*, $\Rightarrow$, $\Leftrightarrow$, *XOR*, *NOR*, *NAND* }. Parentheses are used to formulate more complex expressions [9].

B. Prioritized Test Case Generation

In addition to the above mentioned classical combinatorial test case generation, the CTE offers the possibility to add weights to classes for a prioritized test case generation [5].

Weights on classes can be used to create test cases in an order corresponding to their relevance. For example, those tests that have revealed most of the failures in previous runs can be executed more frequently.

Within CTE, weights are distributed to classes in order to generate prioritized test cases according to occurrence, error or risk probability. The result is a test suite of tests covering the pairwise combination criterion and being sorted according to their relevance most relevant test cases at the top of the test suite, least relevant at the bottom [12]. By using the optimize menu, the tester can adjust the test suite individually by selecting the weighted coverage see Fig. 3.

C. Test Sequence Generation

Many software-based systems are state-based. Thus, test data used in test steps of one test sequence must provide a logical sequence to run the desired state transitions. The manual modeling of test sequences important for testing is a challenging task for the tester. Within the CTE this can be done automatically [13].

The tester defines simple state machines [14], modeling the behavior of the system see Fig. 4. Test cases are then derived from that.

A possible application for Hardware-in-the-loop testing has been discussed recently [15].

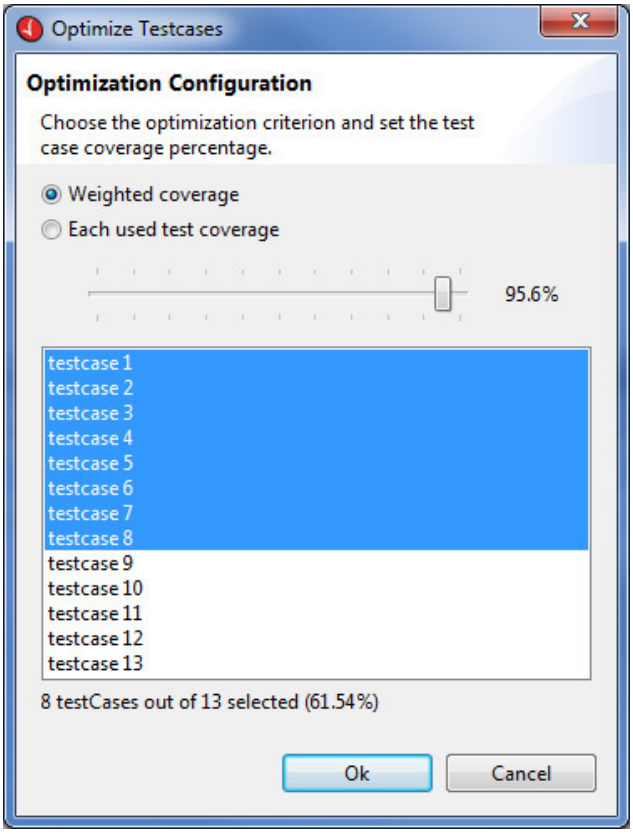


Fig. 3. Optimizing test suites according to weighted coverage in CTE

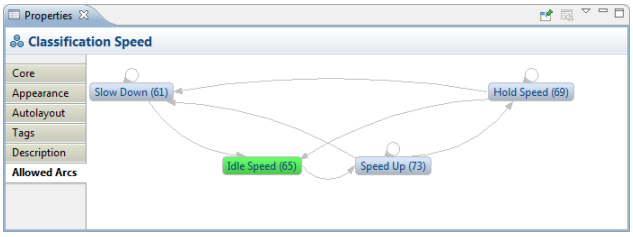


Fig. 4. Allowed arcs for the example classification speed in CTE

IV. Requirements Tracing and Test Management

A. Requirements Tracing within CTE

In the development process, test design is not the first action to take. Before even thinking about testing, the function must be specified. For this, there are several requirements management tools on the market which serve to define and monitor specifications.

For test design, a connection to those requirements might be a huge advantage with regard to traceability. Also, when linking requirements to test cases, gaps can be detected.

The CTE offers the tester to link tree elements and test cases to requirements from MS Access or IBM Rational DOORS [6]. The result is a requirements matrix that directly shows where there is still work to do (Fig. 5). Linked and not yet used requirements can be visualized.

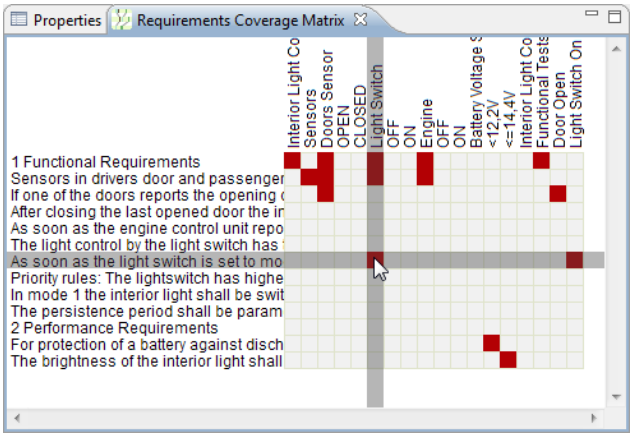


Fig. 5. Requirements matrix view in CTE showing linked CTE objects

Often, requirements change in the process. For this, synchronizing the changed database shows the tester in CTE where to modify the classification tree in order to be conform to the specification [6].

This connection is currently one way to get information from DOORS to use in CTE. In several practical projects, a bidirectional connection has been established, where the results from CTE were exported to the corresponding DOORS module.

B. Test Management in CTE

Parallel to all activities in the test process, test management is essential. For this, CTE synchronizes with HP ALM to commit or submit test cases from and to the test management tool. All information relevant to the tester will be downloaded from the server and also uploaded with updated data. This is supported by a comprehensive mapping mechanism in CTE (Fig. 6). The classification tree will be saved to the central folder on the server in order to make it available for other users for systematic testing.

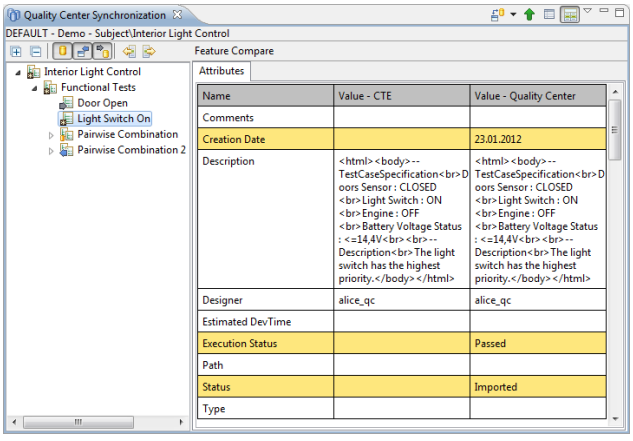


Fig. 6. CTE view of the HP ALM synchronization status after adding new values

Currently, the synchronization is made with the test plan module. For future work, it is planned to extend that connection to the requirements information in order to have a consistent and traceable process, and to extend it to read the information from the test lab which shows information to evaluate the performed tests.

V. Excel Import and Coverage Analysis

A. Excel Import

The most common tool for test design is Excel. There are many different ways in practice to use it within the testing process. With this background, an Excel import to continue with a systematic testing approach in essential.

All Excel sheets no matter if the column or the rows represent the test cases are imported into CTE, as well as the already defined test cases. So the classification tree is built, the test cases are imported and further combinatorial testing can be performed.

B. Coverage Analysis

In practice, the need to give a value for coverage increases. The tester needs to know how *good* the defined test cases are. Within the CTE, the coverage analysis gives a basic overview of the covered tuples see Fig. 7.

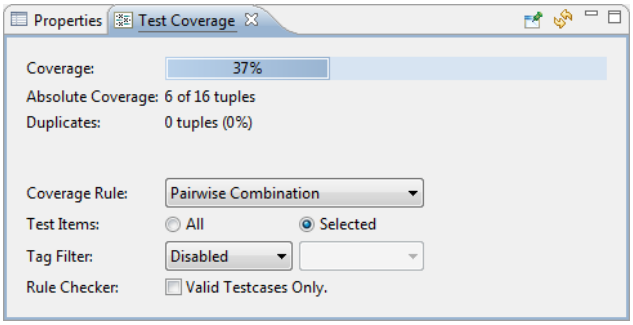


Fig. 7. Test coverage of a set of existing test cases to the coverage rule pairwise combination

If we refer to the Excel import above, it is hard to tell how good or bad the manually defined test cases that have been imported are. With the coverage analysis function, these imported test cases can be compared to specific generation criteria like the pairwise rule, or the minimal coverage rule. According to the result, test suites can then be complemented in order to reach the desired coverage criteria.

VI. Test Evaluation

Throughout the test process, test evaluation is the final step to determine the relevance of the test results. Within CTE, test results for test cases can be either imported or added and then evaluated. The CTE can track test results in terms of *Passed*, *Failed*, *Error*, *Not Executed*. By means of a root cause analysis, the error rate of single classes can be detected and thus gives important information on possible defects of the system under test.

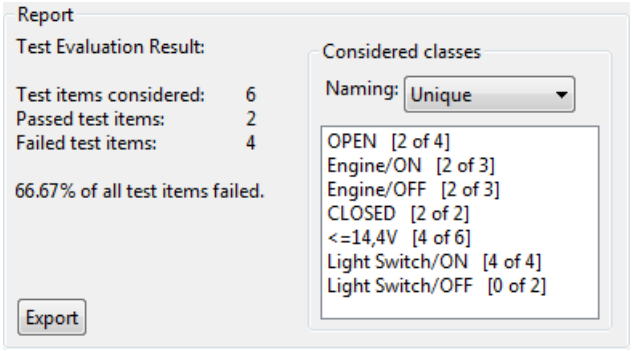


Fig. 8. Test evaluation report in CTE showing the considered test items and their failure rate

In the example given in Fig. 8, six test cases were considered for evaluation: Two passed while four failed. For all failed test cases, detailed results are given in the right part of the Figure, e.g. whenever *OPEN* was used in a test case, it failed two of four times. The last two entries *Light Switch/ON* and *Light Switch/OFF* are most interesting here: Whenever *Light Switch* was *ON*, the test case failed while with *Light Switch ON* there were no problems at all.

VII. Related Work

There is a broad body of work on combinatorial (interaction) testing [4], [16]–[25].

Elbaum et al. provide good overviews of existing prioritization approaches [26]. There has been some work on test case prioritization that considered limited resources [27], [28]. There are some known algorithm supporting prioritized test case generation. The first is an algorithm published in [29], which is an extension to [30]. For efficiency reasons, this algorithm does not consider constraints. The other approach supports constraints and is based on Binary Decision Diagram (BDD) [31]. CTE XL Professional also uses the latter [5], [12]. Search based solutions for test case generations have been presented for both, conventional [8], [32] and prioritized [33] test case generation.

A body of work on the application of the classification tree method and CTE can be found in recent work [34]–[37].

An introduction to root cause analysis can be found in [38], while there are several case studies available, e.g. [39]. The idea of combining the field of (combinatorial) testing with (root) cause analysis is not new, e.g. Cleve and Zeller try to find failure causes through automated testing [40]. The reduction of test suites down to failure-causing combinations is the background of [41], [42], so this approach might not fit too well for regression testing, where new errors can occur. The adaptation of combinatorial testing has also been discussed more recently in [43]. Beside widely positive works, Ramler et al. see limitations with the integration of cause analysis to combinatorial testing, although there is customer demand [44].

VIII. Conclusion

Regardless of the system under test, level of integration, test phase and domain, the classification tree editor is universally applicable [34]–[37]. This method has found a worldwide acceptance over the last two decades and has been used by commercial data processing, aviation and aerospace industries, and many others, for example, for examining the Hubble Space Telescope [45].

The systematic approach, starting from functional requirements, with understandable, reliable (intermediate) results, supported by an efficient, automatic test case generation, ensures that there are no gaps in the testing process and the resulting specifications.

Future work is twofold, first to ease the creation of classification trees, e.g. by importing logs [46] and second the transformation of test specification from the CTE to executable test cases.

Acknowledgment

This work is partly supported by EU grant ICT-257574 (FITTEST).

References

[1] M. Grochtmann and K. Grimm, "Classification trees for partition testing," *Softw. Test., Verif. Reliab.*, vol. 3, no. 2, pp. 63–82, 1993. [Online]. Available: http://dx.doi.org/10.1002/stvr.4370030203

[2] M. Grochtmann and J. Wegener, "Test case design using classification trees and the classification-tree editor CTE," in *Proceedings of the 8th International Software Quality Week*, San Francisco, USA, May 1995. [Online]. Available: http://www.systematic-testing.com/documents/qualityweek1995_1.pdf

[3] G. J. Myers, *The Art of Software Testing*. John Wiley & Sons, 1979.

[4] C. Nie and H. Leung, "A survey of combinatorial testing," *ACM Comput. Surv.*, vol. 43, pp. 11:1–11:29, February 2011. [Online]. Available: http://dx.doi.org/10.1145/1883612.1883618

[5] P. M. Kruse and M. Luniak, "Automated test case generation using classification trees," *Software Quality Professional*, vol. 13(1), pp. 4–12, 2010.

[6] J. Wegener and U. Herold, "Requirements and Test Case Tracing," in *Embedded Real Time Software and Systems 2012 ($ERTS^2$)*, 2012.

[7] P. M. Kruse, J. Nasarek, and N. Condori Fernandez, "Systematic Testing of Web Applications with the Classification Tree Method," in *XVII Iberoamerican Conference on Software Engineering (CIbSE 2014)*, 2014.

[8] M. B. Cohen, M. B. Dwyer, and J. Shi, "Interaction testing of highly-configurable systems in the presence of constraints," in *ISSTA '07: Proceedings of the 2007 international symposium on Software testing and analysis*. New York, NY, USA: ACM, 2007, pp. 129–139. [Online]. Available: http://dx.doi.org/10.1145/1273463.1273482

[9] E. Lehmann and J. Wegener, "Test case design by means of the CTE XL," *Proceedings of the 8th European International Conference on Software Testing, Analysis and Review (EuroSTAR 2000), Kopenhagen, Denmark, December*, 2000. [Online]. Available: http://www.systematic-testing.com/documents/eurostar2000.pdf

[10] P. M. Kruse, J. Bauer, and J. Wegener, "Numerical constraints for combinatorial interaction testing," in *Software Testing, Verification and Validation (ICST), 2012 IEEE Fifth International Conference on*. IEEE, 2012, pp. 758–763. [Online]. Available: http://dx.doi.org/10.1109/ICST.2012.170

[11] J. Petke, S. Yoo, M. B. Cohen, and M. Harman, "Efficiency and early fault detection with lower and higher strength combinatorial interaction testing," in *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering*. ACM, 2013, pp. 26–36. [Online]. Available: http://dx.doi.org/10.1145/2491411.2491436

[12] P. M. Kruse and I. Schieferdecker, "Comparison of Approaches to Prioritized Test Generation for Combinatorial Interaction Testing," in *Federated Conference on Computer Science and Information Systems (FedCSIS) 2012*, Wroclaw, Poland, 2012.

[13] P. M. Kruse and J. Wegener, "Test sequence generation from classification trees," in *Proceedings of ICST 2012 Workshops (ICSTW 2012)*, Montreal, Canada, 2012. [Online]. Available: http://dx.doi.org/10.1109/ICST.2012.139

[14] D. Harel, "Statecharts: a visual formalism for complex systems," *Science of Computer Programming*, vol. 8, no. 3, pp. 231–274, 1987.

[15] P. M. Kruse and J. Reiner, "Systematic design and automated execution of embedded system tests," in *Embedded Real Time Software and Systems ($ERTS^2$) 2014*, 2014.

[16] D. R. Kuhn, D. R. Wallace, and A. M. Gallo, "Software fault interactions and implications for software testing," *IEEE Transactions on Software Engineering*, vol. 30, pp. 418–421, 2004. [Online]. Available: http://dx.doi.org/10.1109/TSE.2004.24

[17] D. M. Cohen, S. R. Dalal, M. L. Fredman, and G. C. Patton, "The AETG System: An Approach to Testing Based on Combinatorial Design," *IEEE Transactions on Software Engineering*, vol. 23, pp. 437–444, 1997.

[18] Y. Lei, K. Tai, F. Inc, and N. Raleigh, "In-parameter-order: a test generation strategy for pairwise testing," in *Third IEEE International High-Assurance Systems Engineering Symposium, 1998. Proceedings*, 1998, pp. 254–261. [Online]. Available: http://dx.doi.org/10.1109/HASE.1998.731623

[19] S. Maity and A. Nayak, "Improved test generation algorithms for pair-wise testing," in *ISSRE*. IEEE Computer Society, 2005, pp. 235–244. [Online]. Available: http://dx.doi.org/10.1109/ISSRE.2005.23

[20] M. B. Cohen, J. Snyder, and G. Rothermel, "Testing across configurations: implications for combinatorial testing," *SIGSOFT Softw. Eng. Notes*, vol. 31, pp. 1–9, November 2006. [Online]. Available: http://dx.doi.org/10.1145/1218776.1218785

[21] M. Grindal, J. Offutt, and S. F. Andler, "Combination testing strategies: a survey," *Softw. Test., Verif. Reliab.*, vol. 15, no. 3, pp. 167–199, 2005.

[22] J. Czerwonka, "Pairwise testing in real world, practical extensions to test case generators," in *Proceedings of 24th Pacific Northwest Software Quality Conference*. Citeseer, 2006, pp. 419–430.

[23] R. C. Bryce and C. J. Colbourn, "The density algorithm for pairwise interaction testing: Research articles," *Softw. Test. Verif. Reliab.*, vol. 17, no. 3, pp. 159–182, 2007. [Online]. Available: http://dx.doi.org/10.1002/stvr.v17:3

[24] W. Grieskamp, X. Qu, X. Wei, N. Kicillof, and M. B. Cohen, "Interaction coverage meets path coverage by smt constraint solving," in *Proceedings of the 21st IFIP WG 6.1 International Conference on Testing of Software and Communication Systems and 9th International FATES Workshop*, ser. TESTCOM '09/FATES '09, 2009, pp. 97–112. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-05031-2_7

[25] A. Calvagna and A. Gargantini, "A formal logic approach to constrained combinatorial testing," *Journal of Automated Reasoning*, April 2010. [Online]. Available: http://dx.doi.org/10.1007/s10817-010-9171-4

[26] S. Elbaum, A. Malishevsky, and G. Rothermel, "Test case prioritization: A family of empirical studies," *IEEE Transactions on Software Engineering*, vol. 28, no. 2, pp. 159–182, 2002. [Online]. Available: http://dx.doi.org/10.1109/32.988497

[27] H. Do, S. Mirarab, L. Tahvildari, and G. Rothermel, "The effects of time constraints on test case prioritization: A series of controlled experiments," *IEEE Transactions on Software Engineering*, vol. 36, pp. 593–617, 2010. [Online]. Available: http://dx.doi.org/10.1109/TSE.2010.58

[28] K. R. Walcott, M. L. Soffa, G. M. Kapfhammer, and R. S. Roos, "Timeaware test suite prioritization," in *Proceedings of the 2006 international symposium on Software testing and analysis*, ser. ISSTA '06. New York, NY, USA: ACM, 2006, pp. 1–12. [Online]. Available: http://dx.doi.org/10.1145/1146238.1146240

[29] R. C. Bryce and C. J. Colbourn, "Prioritized interaction testing for pair-wise coverage with seeding and constraints," *Information & Software Technology*, vol. 48, no. 10, pp. 960–970, 2006. [Online]. Available: http://dx.doi.org/10.1016/j.infsof.2006.03.004

[30] C. J. Colbourn and M. B. Cohen, "A deterministic density algorithm for pairwise interaction coverage," in *Proc. of the IASTED Intl. Conference on Software Engineering*, 2004, pp. 242–252.

[31] I. Segall, R. Tzoref-Brill, and E. Farchi, "Using binary decision diagrams for combinatorial test design," in *Proc. of the 2011 International Symposium on Software Testing and Analysis*. New York, NY, USA: ACM, 2011. [Online]. Available: http://dx.doi.org/10.1145/2001420.2001451

[32] B. J. Garvin, M. B. Cohen, and M. B. Dwyer, "An improved meta-heuristic search for constrained interaction testing," *Search Based Software Engineering, International Symposium on*, vol. 0, pp. 13–22, 2009. [Online]. Available: http://dx.doi.org/10.1109/SSBSE.2009.25

[33] J. Ferrer, P. M. Kruse, J. F. Chicano, and E. Alba, "Evolutionary algorithm for prioritized pairwise test data generation," in *Proceedings of Genetic and Evolutionary Computation Conference (GECCO) 2012*, Philadelphia, USA, 2012. [Online]. Available: http://dx.doi.org/10.1145/2330163.2330331

[34] E. Puoskari, T. E. J. Vos, N. Condori-Fernandez, and P. M. Kruse, "Evaluating applicability of combinatorial testing in an industrial environment: a case study," in *Proceedings of the 2013 International Workshop on Joining AcadeMiA and Industry Contributions to testing Automation*. ACM, 2013, vol. 6, pp. 7–12. [Online]. Available: http://dx.doi.org/10.1145/2489280.2489287

[35] P. M. Kruse, N. Condori-Fernández, T. E. Vos, A. Bagnato, and E. Brosse, "Combinatorial testing tool learnability in an industrial environment," in *7th ACM / IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 2013. [Online]. Available: http://dx.doi.org/10.1109/ESEM.2013.49

[36] P. M. Kruse, O. Shehory, D. Citron, N. Condori Fernandez, T. E. J. Vos, and B. Mendelson, "Assessing the applicability of a combinatorial testing tool within an industrial environment," in *11th Workshop on Experimental Software Engineering (ESELAW 2014)*, 2014.

[37] N. Condori-Fernández, T. Vos, P. M. Kruse, E. Brosse, and A. Bagnato, "Analyzing the applicability of a combinatorial testing tool in an industrial environment," Technical report UU-CS-2014-008, Utrecht University, Tech. Rep., 2014.

[38] J. J. Rooney and L. N. V. Heuvel, "Root cause analysis for beginners," *Quality progress*, vol. 37, no. 7, pp. 45–56, 2004.

[39] M. Leszak, D. E. Perry, and D. Stoll, "A case study in root cause defect analysis," in *Proceedings of the 22nd international conference on Software engineering*. ACM, 2000, pp. 428–437. [Online]. Available: http://dx.doi.org/10.1145/337180.337232

[40] H. Cleve and A. Zeller, "Finding failure causes through automated testing," in *International workshop on automated debugging*, 2000, pp. 254–259.

[41] C. Nie and H. Leung, "The minimal failure-causing schema of combinatorial testing," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 20, no. 4, p. 15, 2011. [Online]. Available: http://dx.doi.org/10.1145/2000799.2000801

[42] L. S. G. Ghandehari, Y. Lei, T. Xie, R. Kuhn, and R. Kacker, "Identifying failure-inducing combinations in a combinatorial test set," in *Software Testing, Verification and Validation (ICST), 2012 IEEE Fifth International Conference on*. IEEE, 2012, pp. 370–379. [Online]. Available: http://dx.doi.org/10.1109/ICST.2012.117

[43] C. Nie, H. Leung, and K.-Y. Cai, "Adaptive combinatorial testing," in *Quality Software (QSIC), 2013 13th International Conference on*. IEEE, 2013, pp. 284–287. [Online]. Available: http://dx.doi.org/10.1109/QSIC.2013.22

[44] R. Ramler, T. Kopetzky, and W. Platz, "Combinatorial test design in the tosca testsuite: lessons learned and practical implications," in *Software Testing, Verification and Validation (ICST), 2012 IEEE Fifth International Conference on*. IEEE, 2012, pp. 569–572. [Online]. Available: http://dx.doi.org/10.1109/ICST.2012.142

[45] N. Tull, "Applications of specification-based testing in flight software development for hubble space telescope mission operations," 2005. [Online]. Available: http://terpconnect.umd.edu/~austin/ense623.d/projects05.d/NzingaTull-Final-Report.pdf

[46] P. M. Kruse, I. W. B. Prasetya, J. Hage, and A. Elyasov, "Logging to facilitate combinatorial system testing," in *Future Internet Testing*. Springer International Publishing, 2014, pp. 48–58. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-07785-7_3

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1591–1598
DOI: 10.15439/2014F152
ACSIS, Vol. 2

A Comparison of Three Black-Box Optimization Approaches for Model-Based Testing

Teemu Kanstrén
VTT, Oulu, Finland
UofT, Toronto, Canada
teemu.kanstren@vtt.fi

Marsha Chechik
University of Toronto
Toronto, Canada
chechik@cs.toronto.edu

***Abstract*—Model-based testing is a technique for generating test cases from a test model. Various notations and techniques have been used to express the test model and generate test cases from those models. Many use customized modelling languages and in-depth white-box static analysis for test generation. This allows for optimizing generated tests to specific paths in the model. Others use general-purpose programming languages and light-weight black-box dynamic analysis. While this light-weight approach allows for quick prototyping and easier integration with existing tools and user skills, optimizing the resulting test suite becomes more challenging since less information about the possible paths is available. In this paper, we present and compare three approaches to such black-box optimization.**

Keywords—model based testing; test automation; evaluation; test generation; optimization

I. Introduction

MODEL-BASED testing (MBT) [1] is a technique for generating test cases from a test model. As opposed to manual test design, where a test expert designs test cases one by one, in MBT the test expert designs a test model to represent the system behavior and uses an MBT generator tool to generate a set of test cases from it.

As the generator can potentially create a very large number of test cases for any non-trivial model, optimization approaches are often applied to choose which tests to include in the generated set. Many optimization solutions [2, 3, 4, 1, 5] use a custom notation combined with a specialized runtime environment to represent the test model, allowing for performing an in-depth white-box static analysis, e.g., symbolic execution, of the model. Tools such as constraint solvers can then be applied on this information to find test cases that yield short paths to reach given coverage targets [3, 6, 5]. Some use manually crafted scenarios to guide the generator towards specific paths [2].

In our work, we have aimed to provide light-weight solutions suitable for easy industry adoption and to support a fast iterative testing process through rapid prototyping of test models, test generation, and test execution, while still providing good test coverage and useful test results. We use a general-purpose programming language (Java) to represent our test models, allowing use of all language features, tools, libraries, and the standard (Java) virtual machine (JVM) runtime. This enables using existing development skills and toolsets such as test libraries and environments, integrated development environments, and continuous integration systems, along with any features they provide. Our open-source test generator called OSMO Tester [7] has been successfully applied, together with industry partners, to test large scale real industry systems.

Our test model is as an executable program, executed in different ways by the generator to produce test cases. We allow use of different, evolving versions of the language platform (virtual machine) and features to create the model and run the generator on it. As the model can make use of third-party libraries and networked services, we assume no access to source code or even all binaries for analysis. Such limitations are common in practical settings (e.g. [8]). In this context, we cannot apply approaches based on white-box static analysis but rather rely on dynamic analysis and information available at runtime. We call this type of test generation *model-based black-box test generation*. While there has been extensive research into optimizing test generation using static analysis based approaches, little work exists in optimizing model-based black-box test generation.

In this paper, we describe three algorithms for optimizing test generation in this type of an environment. All of them are based on generating a large set of potential tests in parallel and picking the most optimal ones based on given coverage criteria. One is targeted at *online testing* where test generation and execution are interleaved. Two others are targeted at *offline testing* where the test set is first generated and later executed in a separate phase. We compare the strengths and weaknesses of the three algorithms and make usage recommendations.

The rest of the paper is structured as follows. Section II presents background on our modelling notation and test generator. It also defines how we assess achieved test coverage over the test model. Section III presents our optimization algorithms. In Section IV, we evaluate each algorithm individually in terms of achieved coverage, generation time and test length. In Section V, we compare the different approaches to each other. In Section VI, we compare our results with related work. We conclude in Section VII with a summary of the paper and discussion of future research directions.

II. Background

To provide background for the following sections, we first briefly outline our modelling notation and model structure and describe our notion of test coverage, including how model elements are used to calculate coverage.

978-83-60810-58-3/$25.00

A. Modelling Notation

OSMO Tester uses a generic programming language (Java) as the modelling language. The models are executable and capture a set of possible test steps and how they can be combined to produce valid test cases. To generate test cases, the test generator executes different paths through the model which is often referred as a *model program* [9, 10]. In Figure 1, we illustrate our notation using a simple test model for a counter which can perform two functions: *increase* (the value of the counter by one) and *decrease* (the value of the counter by one). For illustration purposes, we will later use '+' to represent the increase step and '-' the decrease step.

In this model, the *system under test* (SUT) is represented by the *sut* variable. This example illustrates an *online* testing approach where the test steps are concretely executed against the SUT as they are generated. In an *offline* approach, the test steps yield a script which also includes the input for the SUT and the checks to perform against its states and output.

```
1: public class CounterModel {
2:  @Variable  //annotation to identify interesting model state to generator
3:  private int value = 0;
4:  private Counter sut = new Counter();
5:  private Requirements req = new Requirements();
6:  @BeforeTest
7:  public void start() {
8:    value = 0;
9:    sut.reset();
10: }
11:  @Guard("decrease")
12:  public boolean allowDecrease() {
13:    return value > 1; //when true, "decrease" is enabled
14:  }
15:  @TestStep("decrease") //enabled when above guard true
16:  public void decreaseCounter() {
17:    value--; //updates our model state
18:    sut.decrease(); //execute test step on SUT
19:    assertEquals(value, sut.value); //test oracle, check model vs SUT
20:    req.covered("decrease"); //user defined coverage requirement
21:  }
22:  @TestStep("increase") //has no guards, so is always enabled
23:  public void increaseCounter() {
24:    value++; //updates our model state
25:    sut.increase(); //execute test step on SUT
26:    assertEquals(value, sut.value); //test oracle, check model vs SUT
27:    req.covered("increase"); //user defined coverage requirement
28:  }
29:  @CoverageValue( public String zero() {
30:    return "" + (value == 0);
31:  }
```

Figure 1. Example counter model program.

The two basic model elements here are the methods annotated with *@TestStep* and *@Guard*. The *@TestStep* methods represent test steps that are executed by the test generator at different times to produce a test case. In MBT, these are also referred to as *actions* [11] and *action methods* [5]. Each test step invokes a function on the SUT, updates the model state, or checks the SUT state and output against the expected values (the test oracle). A sequence of these steps forms a *path* through the model, producing a test case.

To define the potential paths, i.e., the steps the generator can take in a specific model state, the test generator executes all *@Guard*-annotated methods (line 11 in Figure 1). These define rules for enabling test steps. When the guards for a step become true, the step is enabled, and the generator can choose to take any of the enabled steps. The guards are matched to steps based on the names given as annotation parameters. For example, in the beginning *value* is 0 and thus the guard for *decrease* is false, meaning the test can only start with the *increase* step.

B. Specifying Coverage Values

In our example, the current value of the counter is stored in the model as the *value* variable. The annotation *@Variable* identifies this variable to be of relevant for the generator to track for coverage. To provide a test oracle, the *value* variable is constantly updated to match the expected value as result of actions executed against the SUT (lines 17 and 24). These actions are the test steps invoking the *increase* and *decrease* functionality of the actual SUT (lines 18 and 25). The test oracle compares the actual value in the SUT with the expected value in the model (lines 19 and 26).

We also define a set of additional terms for elements of the model when calculating our test coverage. We use the term *step-pair* to refer to two steps following one another in a test case. For example, a path of '++-++' would have three unique step-pairs: '++', '+-', and '-+'.

Coverage *requirements* (lines 20 and 27) can be used to tag specific paths of interest through the model. Methods annotated with *@CoverageValue* annotation (line 29) can be defined to produce values to record as covered for specific paths. Typically, these record specific instances or combinations of interest for model state. In our example, it records whether the value `zero' is covered by a path. As such functions are typically related to model state, we refer to them as *user defined state* coverage. When two different values are observed in a sequence inside a test path, the term *state-pair* coverage is used. State-pairs are recorded similarly to step-pairs but with the user defined state values.

The annotations, variable names and values, and method information are accessed at runtime using the standard JVM reflection support, maintaining the black-box quality of our approach.

C. Coverage Calculation

In order to evaluate model coverage achieved by the generated test cases, our generator keeps track of any coverage values it has observed during test paths. Users can give weights to the different model elements to focus coverage where interesting. For example, one can set some weights to zero to ignore them, or tune state weight lower to avoid large state-spaces taking over all others.

The model elements used for coverage calculation and their default weights used by our tool are given in Table 1. The default weights are based on our experience with various test models and related test generation.

Using E to represent a model element and W its weight, the formula to calculate the coverage score is:

$$\sum_{N=1}^{7} (E_N * W_N)$$

N	Model Element	Default Weight
1	Variables	10
2	Variable values	1
3	Test steps	20
4	Step pairs	30
5	Requirements	50
6	User defined states	50
7	State-pairs	40

Table 1. Coverage elements.

That is, the number of unique values observed for each model element is multiplied by the weight for that element, and the sum of these forms the coverage score for a test case. For example, a test case TC1 with a path of '++-++' would cover 1 variable (*value*), 3 variable values (1,2,3 for *value*), 2 steps ('+','-'), 3 step-pairs ('++', '+-', '-+'), 2 requirements ('increase', 'decrease'), 1 user defined state ('false'), and a single state pair ('false-false'). The score of this path is thus 10*1+1*3+ 20*2+30*3+50*2+50*1+40*1 = 333.

The coverage score for a test suite (all test cases together) is calculated by adding up all items in all test cases and then applying the formula. For example, suppose a test suite consists of TC1 defined above and a test case TC2 covering the path `++--+'. If TC2 was the only test case, its value would be 10*1+1*2+20*2+30*4+50*2+50*2+40*3 = 492.

However, since TC1 is already in the suite, only the new elements added by TC2 are considered. These are a new step-pair '--', a new user-defined state 'true' and two new state-pairs 'false-true' and 'true-false'. Thus the suite score rises by 30*1 (step-pair) + 1*50 (state) + 2*40 (state-pairs) = 160. The final coverage score for this test suite is then 333+160 = 493. In general, our scoring function guarantees that the score of a test suite is independent on the order of adding test cases.

A common goal for testing is covering variance over important properties of the SUT [12]. The goal of our coverage algorithm is to provide a measure of overall model variance coverage for the optimization algorithms, which we expect to be designed to represent important aspects to test in the SUT.

III. Optimization Algorithms

In this section, we describe our optimization algorithms. First we describe the online optimization algorithm (Section 3-A), followed by the two offline algorithms (Sect. 3-B and 3-C).

A. Online Algorithm

Online testing interleaves test generation with test execution. Once the generator chooses a test step, it immediately executes it against the SUT; once the step has finished executing, it chooses the next step, etc. For us, online testing provides immediate test results and gives fast feedback to the test generation, modelling and evaluation process, and thus we want online test generation to be as fast as possible.

This type of real-time online test generation prevents us from doing optimization beforehand as we need to support cases where changes are made to the model, the generator is immediately invoked, and tests are generated and executed. To support this scenario, our online optimization algorithm explores sets of potential future steps while the previously chosen step is still executing.

The high-level algorithm is described in Figure 2. When the generator has chosen a step to execute (using the algorithm in Figure 2) but before it has executed it, it starts the exploration of the next step in parallel threads (or concurrently on a networked larger machine as discussed in [13]). It takes the sequence of steps so far executed (LS) for the current test case and the exploration depth (D) given by the user as input. The exploration depth defines how many steps the algorithm tries to look into the future in parallel to current step execution.

```
Input: current model instance and state CM used for concrete test generation,
       list of steps LS currently taken in test case, the exploration depth D.
Output: The next step to take.
1. execute all model guards on CM to construct the set of enabled steps ES
   Set the set of potential future paths PP to empty set ∅.
2. for each step S in ES
3.   create a new instance M of the model program
4.   set M in exploration mode
5.   execute all steps in LS on M to reach current test state for M
6.   execute S on M to reach a new state NS
7.   decrease D by 1
8.     if D > 1, repeat from 2
9.     else add this path for M to PP
10. set value for best path score B to 0, create new empty set of best paths SB
11. for each potential path P in PP
12.   calculate coverage score CS for P
13.   if CS > B, clear SB and set B as CS
14.   if CS == B, add P to SB
15. return a random choice from SB as the next step to take
```

Figure 2. Algorithm for online optimization.

The starting point is the current concrete model instance CM (and its state) used to generate test cases. To explore the potential future paths, the algorithm starts with the set of enabled steps ES, that is the set of steps in CM where no guard returns false, as explained in Section II-A. For each step S in ES, a new instance M of the model program is created. This is set to exploration mode by invoking *@ExplorationEnabler* annotated methods on the model. This should, for example, replace a reference to a real SUT with a mock version that invokes no real functionality and thus has no visible side-effects when steps in M are invoked. The current set of steps LS is executed on this M to achieve the current generation state for M. This is repeated for each S, so that each S has its own instance of M that is in the same generation state.

For each of these instances of M, the associated S is invoked. If the exploration has at this point reached depth D, the coverage score for each M is calculated and that is the score that is given to this path. If D is not yet reached, all these executed paths for the different instances of M are taken as the new sets of steps LS and the algorithm starts from beginning for all these, with the value of D reduced by one.

In the end, all the final paths are taken and the highest scoring ones are collected. If there are several, the one that has the highest score fastest (in fewer steps) is taken. If several are still equal, one is chosen at random. The next unexecuted step on this path is given as a choice for the generator. In some cases, the algorithm still has to wait some time for the parallel exploration to finish. This can be mitigated by forming an "exploration buffer". That is, exploring several future steps at a time when there is a chance and using that as a buffer.

B. *Offline Algorithms*

Our offline optimization algorithm has two different variations. One aims to optimize for a large test set covering a high variation of the test model elements, using the coverage formula presented in Section II.B. The other one aims to optimize for minimal test lengths to cover specific targets in the test model.

Greedy Variation Optimizer. The version targeting high variation is a form of greedy algorithm. It is described in Figure 3. The main arguments it takes from the user are the population size (PS) and the timeout value (TO). The number of parallel generators (P) to run can also be configured but defaults to the number of processing units for the system.

The greedy algorithm runs P versions of generators G in parallel. In Figure 3 This is represented as $G_{1\text{-}P}$ meaning there are P different instances of test generators. The different ones are referred to as G_x, where x stands for one value from 1-P, or one instance of the generator. Each G_x is configured with the test generator configuration (GC) provided by the user and automatically populated with a unique randomization seed (to produce different test cases). Each G_x generates a given number of new test cases (PS). The new generated set of tests GS for G_x is merged with the existing test suite TS (which starts empty). Every T in TS is then iterated and highest scoring tests are added to the existing test suite TS for each G_x. First, the test T that has the highest coverage score in GS is added to TS and removed from GS. This process repeats until GS is empty or no T gets a positive score. Same procedure is done for each G_x generating a new GS and merging with TS.

Finally, once the overall generation timeout TO is reached or TS has not changed throughout the entire iteration, the process is stopped for that G_x. Once all G_x are finished, the TS for all G_x are combined to form the final optimized test set OS which is returned. This is done similar to creating the set for a single generator, starting with the highest scoring test in the overall set, followed by the test that adds most to this test (suite), and so on.

Input: generator configuration GC, population size PS, timeout TO, degree of parallelism P.
Output: Generated test suite OS
1. create P instances of test generator as $G_{1\text{-}P}$, configured with GC
2. for each G_x, in $G_{1\text{-}P}$ run the following in parallel
3. create empty test suite TS (∅)
4. create unique randomization seed for G_x
5. use G_x to generate PS new test cases as the test set GS
6. add tests in TS to GS
7. clear TS, setting TS to empty set ∅
8. for each test T in GS
9. calculate added score AS for T when added to TS
10. add highest scoring T to TS and remove it from GS
11. if any T remains in GS with AS > 0, iterate from step 8
12. if timeout TO has been reached, stop generation with G_x
13. else if TS scores higher or is shorter than previous TS
14 continue from line 5
15. else stop generation with G_x
16. wait for all G_x to finish
17. merge results for all G_x using lines 8-11 as output set OS
18. return OS

Figure 3. Greedy algorithm for offline optimization.

Single Target Optimizer. The *single target* offline optimization approach, described in Figure 4, aims to generate a set of test cases where each coverage requirement is covered by a single test case of the shortest possible length. This can be useful if specific tests are needed but we want to have them generated from the model as opposed to manual scripting.

This algorithm starts by generating test cases as random walks through the model, using a set of P test generators $G_{1\text{-}P}$ running in parallel. Each G_x is used to generate a given number PS of test cases. When a G_x finds a new test case T that covers a previously uncovered requirement R, it changes the global generator state to target finding R and to only allow the steps in T. T now becomes the *reference test,* called BT, for covering R. Upon finishing their iteration of generating PS tests, each G_x reconfigures itself with the new global generator state. This means that the goal of every G is to find a shorter path to cover R. The set of R to find can be given to the algorithm or it can pick them up from the model as it generates tests from it.

To further help the generators find potentially shorter paths, the global state is modified after each iteration to look for only those tests which are shorter than BT. Each G_x is configured to allow each available step one time less than in BT. If any G_x finds a new test T with a shorter path to R, this T becomes the new BT for R and all G_x reconfigure to target shortening this new BT. After test timeout TO is reached or the path cannot be shortened any more (it only has instances of one step), the final BT for R is added to the final output set OS.

Finally, the search is restarted with the goal of finding a new test for an uncovered R. The previously covered requirements are ignored at this point. The process is repeated until all requirements have been covered or suite timeout SO is reached.

Input: generator configuration GC, population size PS, suite timeout SO, test timeout TO, degree of parallelism P.
Output: Generated test suite OS with one test case for reaching each requirement R
1. create P instances of test generator as $G_{1\text{-}P}$, configured with GC
2. create unique randomization seed for each G_x in $G_{1\text{-}P}$
3. while SO has not been reached
4. run each G_x in parallel to generate PS test cases as test suite TS
5. if any test T in any TS for any G_x reached a new uncovered R
6. set R as target to cover for each G_x
7. set best test BT for R to T
8. for each step S in BT
9. reconfigure all G_x to only allow the steps in (BT – S)
10. use G_x to generate PS new test cases TS2
11. if any test T2 in TS2 is shorter than BT, set BT to T2
12. iterate from line 8 until minimal BT achieved or TO is reached
13. add BT to final output test suite OS
14. iterate from line 1 until SO is reached or all requirements are covered
15. return OS

Figure 4. Single target algorithm for offline optimization.

IV. EVALUATION

In this section, we describe the evaluation of the algorithms. Due to space reasons, we cannot include all the details of all the test runs here. The detailed results are available in [14].

While we have experience with several industry models, we cannot describe these due to confidentiality reasons. Thus our evaluations use three publically available models. Two of these, a test model for a movie reservation system (ECinema), and for a GSM SIM card [1], are ported from the ModelJUnit MBT tool [15]. The third model was previously developed by us for a web application called iTrust, which is a role-based healthcare application [16]. While these are not actual industry models, in our experience the relevant complexity is similar in terms of relevant structural coverage elements of the model such as states, and test steps. The models are available as part of our tool website and repository [7]. Table 2 summarizes the sizes of these models in terms of their model elements (used in coverage calculations).

Model	Steps	Step-Pairs	States	State-Pairs	Reqs
ECinema	19	177	9	40	17
SIM	15	240	2.6k+	22k+	32
iTrust	40	800	47	1124	11

Table 2. Model sizes.

We used the default coverage weights described in Table 1 with one exception. For the SIM model we set the state weight to 5 and state-pair weight to 1 in order to avoid the huge state space taking over all other coverage criteria. This is a typical example of tuning the coverage weights per domain as discussed in Section II-B.

A. *Online Algorithm*

For this algorithm, we are interested in the improvements of coverage when compared to the baseline approach of random step selection and any delay it adds. That is, we are interested in the cost and benefit of exploring the next step(s) while the previous are executing.

To set our evaluation parameters, we did an initial scalability study in which we found that test length of 100 steps and a depth of 2 were the most suitable parameters. Thus we used these in the online algorithm evaluation.

Coverage. To evaluate the algorithm in terms of coverage, we ran the test generator against each of the three test models. We did this with the random algorithm, and with the exploration algorithm using depths of 1 and 2. We repeated each of these experiments 100 times and collected the minimum, maximum and average achieved coverage for each algorithm in each configuration for each model. The results indicate that exploration algorithm generally outperforms the random selection for all our coverage criteria.

Timing. To evaluate the timing aspect of the algorithm, we used the iTrust system as a test subject as it was tested through the web-based graphical-user interface (GUI), which in our experience is a good candidate platform for parallel optimization due to test execution delays. The experiment was run on our laptop system, also running the SUT, including the database server, and the webserver. Test execution used the Selenium WebDriver (Chrome) component, allowing control of an actual browser to simulate a test user using the application. Thus, the flow is the same as with an actual user, with all the requests going through the whole SUT from the browser to the backend database and back.

As random selection is practically instant, it was used as a baseline for comparison. To briefly summarize the interesting parts, exploration at depth of 1 (Expl-D1) is in all cases faster than the execution of the concurrent test step (Random). Exploration at depth of 2 (Expl-D2) is in most cases much slower than the step execution time. This means that, for our setup, the depth of 1 is very reasonable, while depth of 2 is slower, although it does also achieve a much higher exploration score that depth 1.

Our goal in this study was also to evaluate the general feasibility of the approach to achieve near real-time test generation and execution. Here, this is true for depth 1. With optimizations and faster computing resources (e.g., [13]) we believe this is within reach for depth 2. Of course, this also largely depends on concrete test execution speed, which may give us more or less time for exploration. In our experience, long delays are common in many cases, specifically, with GUI-based testing.

B. *Greedy algorithm*

For this algorithm, we were interested in evaluating how well it covers variation over the test model structure. We were also interested in the time it takes to achieve the coverage and length of the resulting test cases.

Coverage. To evaluate the algorithm in terms of coverage, we again used all three of our test models. Comparing the results against the baseline random selection algorithm and the online optimization algorithm, Greedy generally outperforms the other two over long term. However, while Greedy performs much better than exploration in the long run, it may sometimes perform slightly worse for some properties.

Figure 5 shows the overall coverage score evolution for one run of each algorithm as new tests are added to the test suite. The coverage score shown is the overall score of the test suite as shown in Section II-II.C. The obvious observation is how random selection yields much lower coverage score than the other algorithms. What is not so clearly visible is how the exploration algorithms score slightly higher than greedy in the beginning, with greedy surpassing depth 1 at test 30 and depth 2 at test 100.

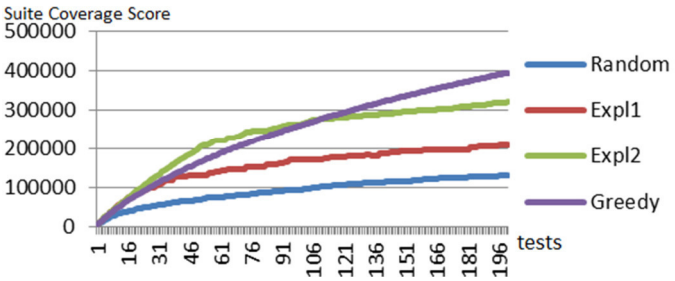


Figure 5. Coverage score evolution over 200 tests.

Timing. Unlike the online exploration algorithm, the greedy algorithm takes time to start and build the test suite for later execution. In this case running the greedy algorithm with our desktop configuration takes about 165 minutes. This is the delay one would have to wait before starting to execute the tests. Compared to the online exploration algorithm, the online version has no delay in the start and the overall generation time for the exploration at depth of 1 is about 20 minutes and for depth 2 about 105 minutes.

C. Single Target Algorithm

To evaluate the single target algorithm, we started by investigating the limits of the algorithm. To do this, we a model with 10 test steps, each always enabled. Thus the probability of taking any one of them with random choice is 10%. It has one requirement to cover, and the length of the path that needs to be taken to cover this requirement can be configured by the *target* parameter. Different values for this parameter were then applied with tests of different length.

The results indicated that the algorithm could always reduce the path to the shortest in all these configurations. Additionally, we used binomial distribution to verify these results. These calculations matched our experiments and thus confirmed the algorithm is useful for such model requirement coverage optimization. This also shows how this type of probability analysis can be used to tune the search parameters when we have an idea of the complexity of paths we need.

To further evaluate the performance, we used the SIM model as a test subject. It had a set of 35 coverage requirements defined by its original authors, and our target was to cover these with one test each. The results indicated that three requirements were not covered at all. In order to find out the reason, we manually analyzed the test model for the paths to reach these requirements. We found that due to complex ways the different test steps and model variables interact, it was not possible to reach those requirements in any way. While we are not the authors of the model, we assume that this is the result of model evolution where the requirements have not been rechecked, perhaps due to the complexity of this process. In any case, such knowledge in itself is valuable to better understand the test model and the generated test cases.

As for the 32 covered requirements, we found that the algorithm did find the minimal path to reach each of these requirements in the model. The results are good but we also realize that there can be more complex cases where the paths are more difficult to minimize due to dependencies between model steps and state variables. We leave further exploration of such models for future work. Again, we remind that the detailed results are available in [14].

V. Discussion

Out of the three algorithms that we have presented, the online optimizer and the greedy offline optimizer target similar goals. Both try to optimize coverage for a variation of chosen parts of the model for each test case and the overall test suite. The single target offline optimizer is different in trying to generate a set of test cases where each one reaches a specific path requirement. It is not looking for the overall variation using our coverage score but rather for a specific set of test cases, each with a minimal set of steps for reaching a given requirement.

As shown in Figure 5, the online exploration approach gains coverage faster in the beginning but loses to greedy over the long term. This is due to the online algorithm finding many uncovered steps, states and their combinations in the beginning but lacking vision in later test cases for how to achieve a state from which it can again gain more coverage score. It is also due to the states being distributed more after the initial set has been covered. The greedy version, on the other hand, selects from a large set of random tests, where the initial tests may not be as good as those in the online exploration version but where over time it finds more uncovered coverage elements in the larger set of random tests. For generating an overall regression test suite, a combination of these could be useful.

Overall, we prefer to use the online exploration algorithm when working on models for systems where test execution takes non-trivial time. One example is when we are evolving the models and trying to determine the impact of the modification on test results. In these cases, the algorithm quickly produces a variation over different parts of the model. If we want to focus this variation on specific parts of the model, we use scenarios to guide the generator to only consider these parts and the algorithm to produce a variation over these chosen parts. Scenarios are a way to tell the generator to ignore some steps completely, use a specific sequence to start every test, and to limit the number of times a test can contain some steps.

If we need the ability to execute large test suites, we have found the greedy or even the random approach to work better. When the tests execute fast (e.g., testing middleware or application logic), it is possible to run a large test set fast even on a single multi-core machine. In such a case, in the time it takes for the optimization approaches to produce the test set, the random selection and its online execution have very likely already achieved a high coverage score and also covered additional combinations. For slower to execute cases such as GUI-based testing, some of these benefits may also be achievable with large-scale testing tools such as Selenium Grid (e.g., overnight or over the weekend using a large set of machines) when we want to perform very extensive test runs.

One of the main considerations for us is the fast evolution of the test models. Modern software development is fast paced and changes are frequent, especially with agile software development practices. In such environments, we prefer the online approach. This allows us to version control only the test model (as opposed to thousands of tests), generate and execute the tests, modify the model and repeat. However, we also find it useful to generate an offline test suite for cases where we want a specific set covered every time, e.g., one larger set executed to provide a high overall model coverage once the external test interfaces have sufficiently stabilized.

This also brings us to the single target optimization algorithm. While the larger test suites generated by the greedy and online optimization algorithms generally also cover the requirements defined in those models, and in many cases a single test case covers many requirements, there can still be value in specific test cases. For example, they are useful to show management and domain experts how specific test requirements are covered by a concise set of test cases, or how specific paths through the model are formed by the generator. They can also be used as a smaller set of offline test cases to support a larger online test generation and execution process.

Our results indicate that using the single target algorithm, coverage requirements can often be more easily found by generating much longer test cases that the expected required length to cover the requirements, and then having the tool minimize them. As noted, the probability calculation presented in Section IV-C can be a useful tool to set these parameters.

One advantage of online testing compared to its offline counterpart is the ability of the former to adapt to responses and state changes of the SUT on the fly [1], e.g., when testing a non-deterministic system. While the algorithms we have presented mostly assume a deterministic test response, the online version can also adapt to non-determinism, with the exploration becoming the estimation of future paths, and each estimate is re-calculated for each new step.

As shown in Figure 5, the optimizations help achieve a more diverse coverage according to the defined score weights when compared to the random choice. We also found it useful, in practice, to add random variation to the generated test cases, to avoid expert bias, i.e., cases not considered by the human expert but exercised in practice. We find that the generation and optimization approaches based on our diverse coverage score (greedy and online look-ahead) work well to achieve such goals. Practically, they focus on achieving the given model coverage criteria and interleave these with elements of randomness in different parts. That is, the result may not be fully optimized to produce the smallest possible test sets, but generally, we prefer this over too aggressive optimization. This provides a good tradeoff between generating huge random test sets and only generating manually defined specific test sets, which do not make use of the large scale capacity of test generators and are subject to expert bias in what is tested.

VI. Related Work

Test optimization in MBT has been a popular research topic with various tools and approaches being implemented. Tools such as Conformiq [3], Uppaal [4], and Spec Explorer [5] use customized modelling languages, and static analysis-based approach with symbolic execution and constraint solving to optimize test sets. Mostly these approaches target offline test generation, while some work [6, 4] has also made use of two-phased test generation where static analysis is first performed and the resulting information is used to aid in online test generation. In contrast, our approaches are targeted to cases where such static analysis is not available due to the complexity of the modeling language and environment.

Various approaches using general-purpose programming languages for modelling have also been presented. Model programs similar to ours are used in MBT tools such as Spec Explorer [5], PyModel, NModel [10], and ModelJUnit [1]. All of these tools use variations of random search for test selection. Spec Explorer, PyModel and NModel support guiding test selection through user defined scenarios which slice the model to focus test generation around the specific parts of the model. Similar scenarios are also supported in our generator, and these can be used together with the optimization approaches we presented to focus testing on specific model parts.

A black-box optimization approach for test generation from models expressed using general purpose modelling languages is also available with ModelJUnit [15]. It supports executing the test model in a simulation mode as a pre-analysis phase before starting the actual test generation. Possible sequences of steps observed in these runs are collected and used to guide the actual online test generation in the following phase. However, this approach does not consider the state of the model, which defines what paths are available, and thus the set of paths assumed by the optimization analysis are different from the actual ones available during generation. The difference to our approaches is that we do not use a separate pre-analysis phase and produce accurate results where the exact paths and impacts on coverage are known.

Regression test selection, minimization and prioritization are topics related to optimizing test suites [17]. These typically consider the optimization in terms of different ways to cover the SUT implementation [17]. In this paper we have discussed optimization of test model coverage using a more varied set of coverage criteria at a higher abstraction level. However, investigating ways to use works such as described in [17] together with our approach is an interesting topic for future work.

The techniques presented in this paper can also be viewed as a multi-objective test optimization problem. We aim to cover the items defined for the scoring function in Section II, which can be seen as a fitness function in the terms of search-based testing (see e.g., [18]). The algorithms we apply for coverage score optimization can be seen as a dynamic multi-object optimization algorithm for the test suite. Various approaches in search-based optimization have been applied before [18] but none in our knowledge to model-based test generation with black-box constrains. In fact, we are not aware of previous work in MBT for optimizing a set of coverage criteria as diverse as the one we optimize in this paper.

A multi-objective approach to test suite optimization for product lines is presented in [19], targeting objectives such as cost of test case, cost of test requirement and product variants. Test coverage is considered as single coverage requirements on the test model as opposed to our extensive model variation coverage support. An interesting aspect to integrate with our work could be the association of cost to certain test targets as part of the coverage score function.

In software testing and verification, various combinations of static analysis and (random) test data generation have been used. For example, directed automated random testing (DART) [20] combines static analysis and observations about the running system with random inputs to guide it towards new paths. Java PathFinder (JPF) [21] enables generating test paths based on a combination of symbolic and concrete executions. While these are different testing approaches than MBT, combining the information from these types of different sources with test generation (similar to [6]) and the approaches presented in this paper could be an interesting future research topic. While it may be prohibitive in terms of wait time to perform such extensive pre-analysis, a possible approach could be to perform this as a background process during modelling similar to what is described for executing unit tests in [22].

In general, a lot of work in automated test generation targets the traditional code coverage as a test target. When additional test coverage targets are considered, these are typically specific coverage requirements such as labels on the test models [23]. In addition to the optimization algorithms we presented in this paper, the coverage scoring method itself extends these with wider generic model coverage criteria (the model structure elements), supported by domain-specific criteria (user defined state and variables).

Random testing has been shown to work well in various situations [24, 25], and some criticism has been voiced on the effectiveness of guided random testing approaches when the same coverage can be achieved by simply executing a large set of random test cases in the same time [25]. We share this view in noting that executing a very large set of random tests should yield the same or even higher coverage over time as our optimization approaches do. The aim of our approaches is to practically choose a reasonable subset of such a larger set under different use cases.

VII. CONCLUSIONS

In this paper, we presented three different approaches to test optimization in black-box model-based testing. We then evaluated their performance and provided comparisons on the strengths and weaknesses of each. Our evaluation highlights the benefits of these different approaches over the traditional random choice and their usefulness in different contexts. The online version works well to provide added coverage when prototyping slow to execute test cases. The greedy offline optimizer gives a test suite for higher overall model coverage. The requirements targeting optimizer can help provide specific test cases where useful. Combining potential benefits of these approaches with static analysis where possible would be an interesting future research topic. Domain-specific applications of these approaches, including customizations of algorithms and modelling languages, are also interesting future topics.

VIII. REFERENCES

[1] M. Utting and B. Legeard, Practical Model-Based Testing: A Tools Approach, Morgan Kaufman, 2006.

[2] W. Grieskamp, N. Kicillof, K. Stobie and V. Braberman, "Model-Based Quality Assurance of Protocol Documentation: Tools and Methodology," *Journal of Software Testing, Verification and Reliability,* vol. 21, no. 1, pp. 55-71, 2011. DOI: 10.1002/stvr.427

[3] A. Huima, "Implementing Conformiq Qtronic," in *Testing of Software and Communicating Systems*, 2007.

[4] M. Mikucionis, K. Larsen and B. Nielsen, "T-Uppaal: Online Model-Based Testing of Real-Time Systems," in *19th International Conference on Automated Software Engineering (ASE)*, 2004. DOI: 10.1109/ASE.2004.1342774

[5] M. Veanes, C. Campbell, W. Grieskamp, W. Schulte, N. Tillmann and L. Nachmanson, "Model-Based Testing of Object-Oriented Reactive Systems with Spec Explorer," *Formal Methods of Testing,* pp. 39-76, 2008.

[6] D. Ahman and M. Kääramees, "Constrain-Based Heuristic On-line Test Generation from Non-Deterministic I/O EFSMs," in *7th Workshop on Model-Based Testing*, 2012. DOI: 10.4204/EPTCS.80.9

[7] T. Kanstrén, "OSMO Tester Home Page," February 2013. [Online]. Available: http://code.google.com/p/osmo. [Accessed February 2013].

[8] A. Groce, A. Fern, J. Pinto, T. Bauer, A. Alipour, M. Erwig and C. Lopez, "Lightweight Automated Testing with Adaptation-Based Programming," in *IEEE International Symposium on Software Reliability Engineering*, 2012. DOI: 10.1109/ISSRE.2012.1

[9] M. Veanes, C. Campbell, W. Schulte and N. Tillmann, "Online Testing with Model Programs," in *ESEC/FSE-13*, 2005. DOI: 10.1145/1095430.1081751

[10] J. Ernits, R. Roo, J. Jacky and M. Veanes, "Model-Based Testing of Web Applications using NModel," in *Testing of Software and Communication Systems*, 2009.

[11] M. Utting, A. Pretschner and B. Legeard, "A Taxonomy of Model-Based Testing Approaches," *Software Testing, Verification and Reliability,* vol. 22, no. 5, pp. 297-312, 2012. DOI: 10.1002/stvr.456

[12] J. A. Whittaker, Exploratory Software Testing: Tips, Tricks, Tours, and Techniques to Guide Test Design, Addison-Wesley, 2009.

[13] T. Kanstren and T. Kekkonen, "Distributed Online Test Generation for Model-Based Testing," in *Asia Pacific Software Engineering Conf.*, 2013. DOI: 10.1109/APSEC.2013.43

[14] T. Kanstrén and M. Chechik, "A Comparison of Three Black-Box Optimization Approaches for Model-Based Testing," 5 June 2014. [Online]. Available: http://www.kanstren.net/appendix/ATSE2014_full.pdf. [Accessed 5 June 2014].

[15] M. Utting, "The ModelJUnit Model-Based Testing Tool," 2009. [Online]. [Accessed 17 May 2013].

[16] North Carolina State University, "iTrust: Role-Based Healthcare," 2013. [Online]. [Accessed 17 May 2013].

[17] S. Yoo and M. Harman, "Regression Testing Minimization, Selection and Priorization: A Survey," *Software Testing, Verification and Reliability,* vol. 22, no. 2, pp. 67-120, 2012. DOI: 10.1002/stvr.430

[18] P. McMinn, "Search-based testing: Past, present and future," in *3rd International Workshop on Search-Based Software Testing*, 2011. DOI: 10.1109/ICSTW.2011.100

[19] H. Baller, S. Lity, M. Lochau and I. Schaefer, "Multi-Objective Test Suite Optimization for Incremental Product Family Testing," in *IEEE Internation Conference on Software Testing, Verification and Validation (ICST)*, 2014.

[20] P. Godefroid, N. Klarlund and K. Sen, "DART: Directed Automated Random Testing," in *Programming Language Design and Implementation (PLDI)*, 2005. DOI: 10.1145/1064978.1065036

[21] C. S. Pasareanu and N. Rungta, "Symbolic Pathfinder: Symbolic Execution of Java Bytecode," in *Automated Software Engineering (ASE)*, 2010. DOI: 10.1145/1858996.1859035

[22] D. Saff and M. Ernst, "Reducing Wasted Development Time via Continous Testing," in *Proc. Int'l. Conf. on Software Testing and Analysis (ISSTA)*, 2003.

[23] S. Bardin, N. Kosmatov and F. Cheynier, "Efficient Leveraging of Symbolic Execution to Advanced Coverage Criteria," in *IEEE International Conference on Software Testing, Verification and Validation (ICST)*, 2014.

[24] I. Ciupa, A. Pretschner, M. Oriol, A. Leitner and B. Meyer, "On the Number and Nature of Faults Found by Random Testing," *Software Testing, Verification and Reliability,* vol. 21, no. 1, pp. 3-28, 2009. DOI: 10.1002/stvr.415

[25] A. Arcuri, M. Z. Iqbal and L. Briand, "Random Testing: Theretical Results and Practical Implications," *IEEE Transactions on Software Engineering,* vol. 38, no. 2, pp. 258-277, 2012. DOI: 10.1109/TSE.2011.121

3rd Workshop on Model Driven Approaches in System Development

FOR many years, various approaches in system design and implementation differentiate between the specification of the system and its implementation on a particular platform. People in software industry have been using models for a precise description of systems at the appropriate abstraction level without unnecessary details. Model-Driven (MD) approaches to the system development increase the importance and power of models by shifting the focus from programming to modeling activities. Models may be used as primary artifacts in constructing software, which means that software components are generated from models. Software development tools need to automate as many as possible tasks of model construction and transformation requiring the smallest amount of human interaction.

A goal of the proposed workshop is to bring together people working on MD languages, techniques and tools, as well as Domain Specific Languages (DSL) and applying them in information system and application development, databases, and related areas, so that they can exchange their experience, create new ideas, evaluate and improve MD approaches and spread its use. The intention is to target an interdisciplinary nature of MD approaches in software engineering, as well as research topics expressed by but not limited to acronyms such as Model Driven Software Engineering (MDSE), Model Driven Software Development (MDSD), and OMG's Model Driven Architecture (MDA).

1st Workshop on MDASD was organized in the scope of ADBIS 2010 Conference, held in Novi Sad, Serbia. From this year, MDASD becomes a regular bi-annual FedCSIS event.

Topics

Submissions are expected from, but not limited to the following topics:

- MD Approaches in System Design and Implementation – Problems and Issues
- MD Approaches in Software Process Models
- MD Approaches in Databases and Information Systems
- MD Approaches in Software Quality and Standards
- Metamodeling, Modeling and Specification Languages
- Model Transformation Languages
- Model-to-Model, Model-to-Text, and Model-to-Code Transformations in Software Process
- Transformation Techniques and Tools
- Domain Specific Languages (DSL) and Domain Specific Modeling (DSM) in System Specification and Development
- Design of Metamodeling and Modeling Languages and Tools
- MD Approaches in Requirements Engineering and Business Process Modeling
- MD Approaches in System Reengineering and Reverse Engineering
- MD Approaches in HCI development
- MD Approaches in GIS development
- MD Approaches in Document Engineering
- Model Based Software Verification
- Theoretical and Mathematical Foundations of MD Approaches
- Organizational and Human Factors, Skills, and Qualifications for MD Approaches
- Teaching MD Approaches in Academic and Industrial Environments
- MD Applications and Industry Experience

Event Chair

Luković, Ivan, University of Novi Sad, Serbia

Steering Committee

France, Robert, Colorado State University, USA, United States
Mernik, Marjan, University of Maribor, Slovenia
Ristić, Sonja, University of Novi Sad, Serbia
Tolvanen, Juha-Pekka, MetaCase, Finland

Program Committee

Amaral, Vasco, The New University of Lisbon, Portugal
Bryant, Barrett, University of North Texas, United States
Budimac, Zoran, Faculty of Sciences, Univ. of Novi Sad, Serbia
Chen, Haiming, Chinese Academy of Sciences, China
Erradi, Mohammed, ENSIAS, Mohammed-V Souissi University, Morocco
Fertalj, Krešimir, University of Zagreb, Croatia
France, Robert, Colorado State University, USA, United States
Gray, Jeff, University of Alabama, United States
Ivanović, Mirjana, University of Novi Sad, Serbia
Janousek, Jan, Czech Technical University, Czech Republic
João Varanda Pereira, Maria, Instituto Politecnico de Braganca, Portugal
Karagiannis, Dimitris, University of Vienna, Austria
Kardaş, Geylani, Ege University International Computer Institute, Turkey
Kollár, Ján, Technical University of Kosice, Slovakia
Kosar, Tomaž, University of Maribor, Slovenia
Krdzavac, Nenad, Michigan State University, United States
Kühne, Stefan, Universität Leipzig, Germany
Liu, Shih-Hsi Alex, California State University, United States
Maćoš, Dragan, University of Applied Sciences, Germany
Melo de Sousa, Simão, University of Beira Interior, Portugal
Mernik, Marjan, University of Maribor, Slovenia

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1601–1610
DOI: 10.15439/2014F144
ACSIS, Vol. 2

Grammar-Based Model Transformations

Galina Besova
Department of Computer Science
University of Paderborn
33098, Paderborn
Email: besova@mail.upb.de

Dominik Steenken
Department of Computer Science
University of Paderborn
33098, Paderborn
Email: dominik@mail.upb.de

Heike Wehrheim
Department of Computer Science
University of Paderborn
33098, Paderborn
Email: wehrheim@mail.upb.de

***Abstract*—Model transformation is a key concept in model-driven software engineering. The definition of model transformations is usually based on meta-models describing the abstract syntax of languages. While meta-models are thereby able to abstract from superfluous details of concrete syntax, they often loose *structural* information inherent in languages, like information on model elements always occurring together in particular shapes. As a consequence, model transformations cannot naturally re-use language structures, thus leading to unnecessary complexity in their development as well as analysis.**

In this paper, we propose a new approach to model transformation development which allows to simplify and improve the quality of the developed transformations via the exploitation of the languages' structures. The approach is based on context-free *grammars* and transformations defined by *pairing* productions of source and target grammars. We show that such transformations exhibit three important characteristics: they are *sound*, *complete* and *deterministic*.

I. Introduction

MODEL transformations are key to model driven engineering (MDE). Surveys on model transformations [1], [2] show their expanding application areas: model translation, model composition, refinement, and other.

In an MDE setting, the syntax of models is given in terms of *meta-models* which themselves conform to their own meta-models (e.g., MOF [3]). Meta-models define the *abstract syntax* of languages, abstracting away from the details of concrete syntax like keywords and ordering of elements. Model transformations thus operate on abstract syntax. While meta-models describe model elements and their direct relations, they fall short of describing more complex interrelations like sets of model elements always occurring together in particular shapes. In some cases, meta-models are enriched with OCL [4] constraints to enforce such shapes in models.

In contrast to MDE, traditional approaches to language definition (and translation) define languages by *grammars*, often given in an Extended Backus-Naur Form (EBNF) [5]. These translation techniques operate on concrete syntax. While the details of concrete syntax are in general unimportant (and thus make translation definition unnecessarily confusing), the *structural* information contained in the grammars is highly useful for defining translations. The productions of the grammars define the structures available in the languages, and by relating productions of grammars (as done in syntax-directed translation [6]) we can easily specify how language structures are mapped onto each other.

This work was partially supported by the German Research Foundation (DFG) within the Collaborative Research Centre "On-The-Fly Computing" (SFB 901).

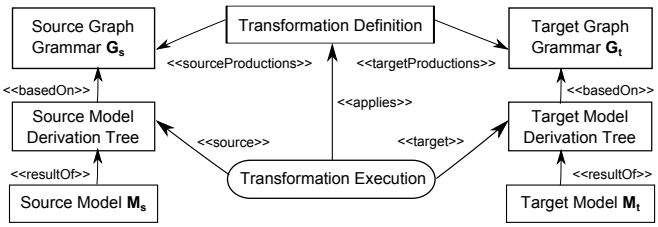


Fig. 1: Overview of our approach

An ideal approach for model transformation should thus combine these two approaches, taking the best of both: have language definitions with the abstract syntax of meta-models and the structures of grammars, and build model transformations on these definitions. An early approach following this idea, although not in the area of model transformations and not with meta-models but with graphs, is the one of Pratt [7]. Pratt defines *pair grammars* as a way of relating the grammars of two languages, thus obtaining a natural way of relating languages and building translations from one to the other.

The objective of this paper is to bring the idea of pair grammar based translation to the world of MDE and model transformations, lifting it to the level of abstract syntax while preserving its advantages. We also extend it in order to cover a broader variety of model transformations.

Fig. 1 gives an overview of our approach. The transformations we focus on are model-to-model transformations. Our models are given in abstract syntax and are generated by grammars. For this generation purpose we use a type of context-free graph grammars – *hyperedge replacement graph grammars* [8] – typed and constrained by meta-models. Transformation rules – like pair grammars – relate productions of the source with those of the target grammar. Model transformations are executed on the *derivation trees*: given a source model $\mathtt{M_s}$, its derivation tree in the source grammar is obtained by parsing, and used by the model transformation to produce a derivation tree in the target grammar, and thereby the corresponding target model $\mathtt{M_t}$.

We exemplify our approach on a transformation from activity diagrams to the process algebra CSP [9]. We also prove

important qualities of the transformations developed with our approach – *termination*, *soundness*, *completeness*, and *determinism*. Showing these quality properties for a transformation described using current state-of-the-art techniques is usually hard, if not impossible [10].

First, in Sec. II we give background on grammar-based language definition and show our source and target grammars. Then, we introduce our approach in Sec. III and in Sec. IV show the quality characteristics of the developed transformations. In Sec. V we demonstrate extensions of our approach, and in Sec. VI we evaluate it in comparison with the most closely related approaches in MDE. Finally, we survey related work in Sec. VII and conclude in Sec. VIII.

II. Background

There are two fundamentally different ways of specifying the syntax of a given language: with (context-free) grammars or with meta-models. Our approach is built on grammars generating instances of meta-models, i.e., graphs. In the following, we introduce the main concepts of the grammar-based language definition and show how they can be lifted to graph-based languages, enabling grammar-based definition of modelling languages and model transformations utilizing these definitions. We show how our example modelling languages for activity diagram and CSP can be described using such grammars. Finally, we introduce the transformation example used later to demonstrate our approach.

A. Grammar-Based Syntax Definition

In their original usage, grammars define languages of strings via a set of generative rules. We briefly review some general definitions of grammars for string languages [6] because our graph grammars is a natural extension of string grammars.

Definition 1 (Grammar)**.** *A* grammar $G = (N, \Sigma, P, S)$ *consists of a set of* non-terminal symbols N*, a set of* terminal symbols Σ*, a set of* productions P *and a designated* start symbol $S \in N$*. Each production* $p \in P$ *is of the form* $p = (l, r)$*, with* $l \in (\Sigma \cup N)^* N (\Sigma \cup N)^*$*, i.e., a string of symbols with at least one non-terminal, and* $r \in (\Sigma \cup N)^*$*.*

A grammar $G = (N, \Sigma, P, S)$ *is called* context-free *iff every production* $p \in P$ *has the form* $p = (n, r)$ *with* $n \in N$*.*

Applications of productions derive new strings from given ones by a process called *rewriting*. A string $s \in (N \cup \Sigma)^*$ is rewritten into a new string by a context-free production $(n, r) \in P$ by finding n in s and replacing it with r. In this way, a grammar defines the set of strings that can be *derived* from its start symbol S. This set is called the *language* of the grammar G.

Definition 2 (Language)**.** *Let* $G = (N, \Sigma, P, S)$ *be a grammar. A* sentence *of* G *is a string* $s \in \Sigma^*$ *of* terminal *symbols that can be derived from* S *using a finite sequence of applications of productions in* P*. The language* $\mathcal{L}(G)$ *of* G *is defined as* $\mathcal{L}(G) = \{s \in \Sigma^* | s \text{ is sentence of } G\}$

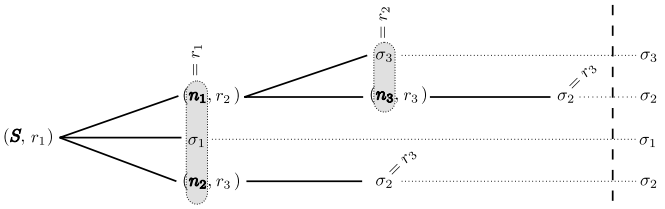


Fig. 2: A derivation tree deriving $\sigma_2\sigma_1\sigma_2\sigma_3$ from S

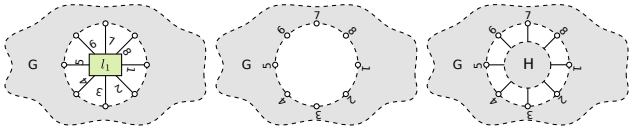
The advantage of context-free grammars is that parsing can be done efficiently [6]. Parsing a sentence in a language using a context-free grammar gives us a *derivation tree*, i.e., a structure showing the application of productions from the start symbol leading to the derived sentence. A derivation tree for a sentence defines its *structure* and confirms its language membership in $L(G)$ (i.e., syntactical correctness). The inner nodes of such a tree are labelled with productions. The root is labelled with a production that consumes the start symbol of the grammar. Every leaf of the tree is labelled with a terminal symbol. Fig. 2 shows an example derivation tree for the sentence $\sigma_2\sigma_1\sigma_2\sigma_3$.

In general, there can be multiple derivation trees for a sentence in one grammar that are not equivalent in their structure, making the grammar ambiguous. In our approach, we only consider *unambiguous* grammars for defining source and target languages, to ensure deterministic behaviour of the developed transformations (see Sec. IV for details).

Definition 3 (Unambiguous Grammar)**.** *A grammar* $G = (N, \Sigma, P, S)$ *is called* unambiguous *iff for every sentence* $s \in \mathcal{L}(G)$ *its derivation from* S *is unique if performed "leftmost derivation first".*

Because most models in MDE contexts are graph-based (since meta-models are graphs), we need context-free grammars producing graphs as sentences instead of strings. We use *hyperedge replacement graph grammars* (HR grammars) [8] which fulfil these requirements. HR grammars operate on *hypergraphs*, a generalization of graphs where edges, called *hyperedges*, can have more than two end points. These end points are called *attachment points* and their number is the arity of a hyperedge. Hyperedges take the role of non-terminal and terminal symbols. For replacing hyperedges in graphs by sub-graphs during rewriting, we need to specify how these sub-graphs are to be embedded. To this end, the replacing sub-graphs are equipped with *external nodes*, and rewriting proceeds by replacing the hyperedge with the sub-graph gluing together each external node with the attachment node of its corresponding attachment point.

More precisely, a *hyperedge replacement rule* $l_1 := H$ has three parts: a single n-ary non-terminal hyperedge l_1, a hypergraph H with $k \geq n$ external nodes replacing the hyperedge, and an (injective) *mapping* g of the k external nodes of H onto the n attachment points of l_1. Unlike string productions, graph replacement rules need to explicitly define how the new graph is attached to the remaining context. In HR grammars this is done via the mapping g. Fig. 3 shows a sketch of the replacement process (for $k = n$). First, the



(a) Matching (b) Removal (c) Gluing

Fig. 3: Rule application in hyperedge replacement grammars

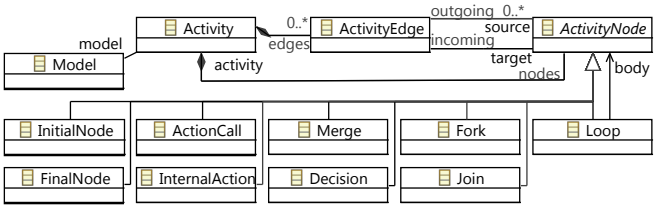


Fig. 4: Source meta-model: Activity diagrams

hyperedge l_1 is found by matching (a) and removed (b), then the graph H is inserted (c) by gluing all its external nodes with the attachment nodes of l_1 according to the mapping g. Due to lack of space, we refrain from giving a more formal definition here. All definitions of string grammars carry over to HR grammars: an HR grammar has the same parts as a string grammar (non-terminals, terminals, productions and a start symbol) and its language is a set of hypergraphs. The membership problem for HR grammars is decidable [8] which guarantees the existence of the derivation trees we are going to use.

Now, we can define both the source and target language we use to demonstrate our grammar-based model transformation approach in terms of hyperedge replacement grammars. We exemplify our approach on a transformation from activity diagrams to the process algebra CSP [9]. The HR grammars for these two languages are compliant with the respective meta-models, i.e., the graphs which our grammars generate are all instances of the meta-models. For this, we use (a simplified version of) the meta-model of UML activity diagrams [11] (see Fig. 4) and the CSP meta-model from [12].

The meta-model of activity diagrams only contains basic diagram elements, their hierarchy, and associations with multiplicities. It does not describe higher-level *syntactic structures of the language*. For example, in a well-formed activity diagram each *decision node* should be eventually followed by the corresponding *merge node* for the branches of that decision. Although these kinds of inductive language structures are intuitive to the transformation developer, they are usually not described in the meta-model.

Fig. 5 shows six out of eleven productions of our source HR grammar. Productions are given in abstract (plus some in concrete) syntax, in the form $l := H$, using bars to distinguish different right-hand sides of productions. Non-terminal hyperedges are depicted by dashed lines. Types and the number of attachment points of a non-terminal hyperedge are determined by the associations of the meta-model elements which it groups. The mapping between attachment points and external nodes is depicted by using the same numbers, one

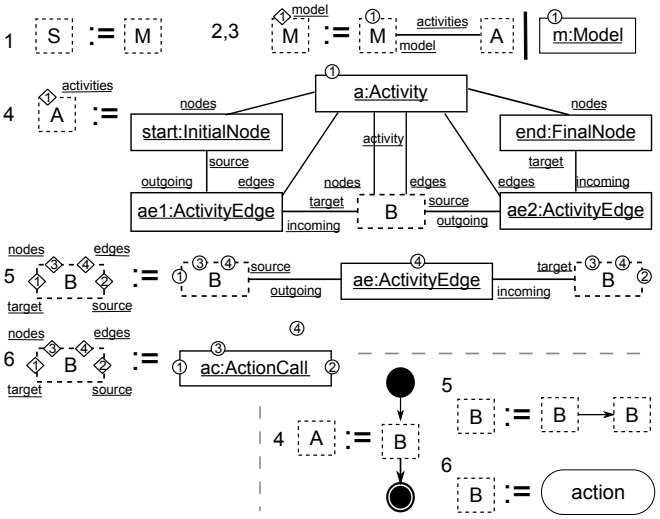


Fig. 5: HR productions subset for activity diagrams (abstract, concrete syntax)

given in a diamond and the other in a circle. Multiple external nodes can be mapped to one attachment point. Not all external nodes have to be connected in the graph (see production 6).

This grammar describes well-formed activity diagrams containing zero or more activities (productions 1 – 3) with exactly one *initial* and *final* node, and at most one recursively-defined high-level *syntactic structure block* called *B* connected to exactly one initial and final node (production 4). The non-terminal edge *B* can be replaced by one of the following structures: a *block sequence* (production 5), a *fork/join block*, a *decision/merge block*, a *loop*, an *internal action* or an *action call* (production 6). Note that our decision/merge constraint is now represented by the corresponding production which only allows to generate these elements together connected in a shape. Other language structures are also produced in this systematic way.

Fig. 6 shows the relevant subset of our meta-model compliant HR grammar for CSP. A *model* described by this grammar can contain a set of *processes* (productions 1 – 3) generated from non-terminal edges labelled *P*. A *process description* represented by non-terminal *PE* (production 4) can contain various expressions: a *sequential* or *parallel* process composition (productions 7 – 8), an *if-then-else* expression (production 9), an *event* followed by another process expression (production 10), *another process* (production 6), or an empty process *SKIP* (production 5).

B. Transformation Example

Both grammars we have defined will be used to describe our example transformation. We describe a transformation from activity diagrams to CSP, frequently employed for analysis purposes [13]. Alternatively, we could, for instance, use the activity diagrams to first-order logic transformation example from [14].

Fig. 7 shows our sample activity diagram of an enrolment and the corresponding CSP process. Here, we see the concrete

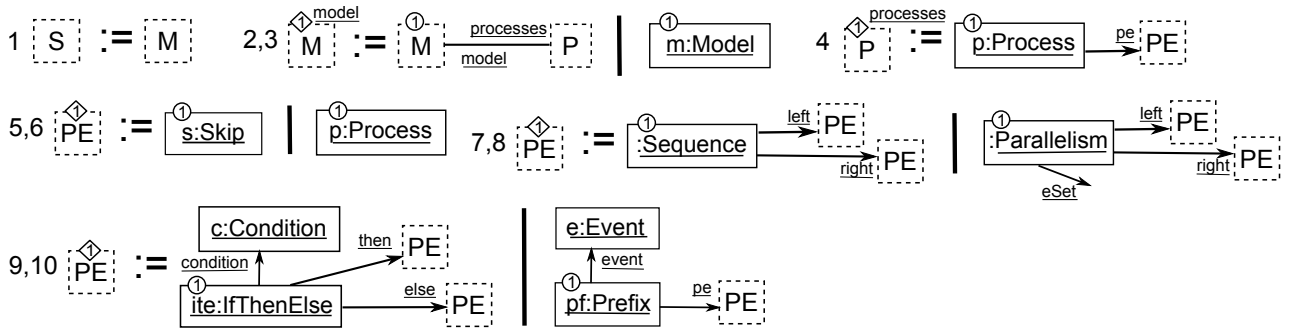


Fig. 6: HR productions subset for CSP (abstract syntax)

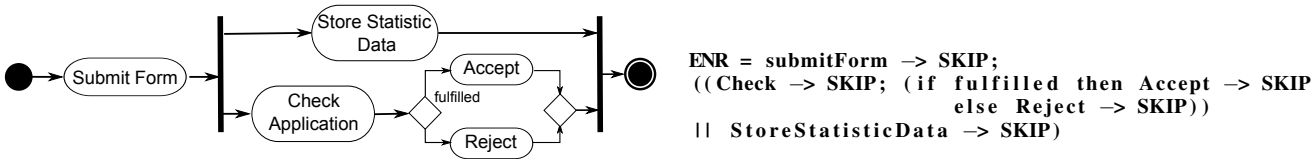


Fig. 7: Enrolment activity diagram (left) and its CSP process (`Check` abbreviates `CheckApplication` event) (right)

syntax of CSP: `->` is an event prefix, `;` a sequential composition, `SKIP` an empty process, `||` a parallel composition and `if-then-else` a conditional choice. We require every activity to be transformed into a process and every block sequence and fork/join block into a sequential and parallel process composition, respectively. A decision/merge block is to be transformed into an if-then-else expression and an action call into an event followed by a SKIP. Loops, although absent in this example, are transformed into recursive processes with conditions. Next, we show how this transformation logic can be structurally described using our approach.

III. Transformation Development

Our main goal is to allow an intuitive model transformation definition by mapping high-level syntactic structures in source and target languages onto each other. In terms of grammars, this means relating (or *pairing*) source and target productions creating corresponding structures. During execution of the model transformation these relations will be used to identify which target production is triggered for which source production.

Our model transformations operate on derivation trees of source and target models. Given a derivation tree for a source model, we incrementally construct a derivation tree for the target model by applying corresponding productions. Thereby, 1-to-1 correspondences between non-terminal edges in the source and target derivation trees help to keep track of related model structures.

Fig. 8 shows a sample transformation rule relating the source production for a sequence of activity blocks to the target production for a sequence of process expressions[1]. It states that when a block sequence is replacing a non-terminal hyperedge

[1]To simplify, we show transformation rules without HR grammar details.

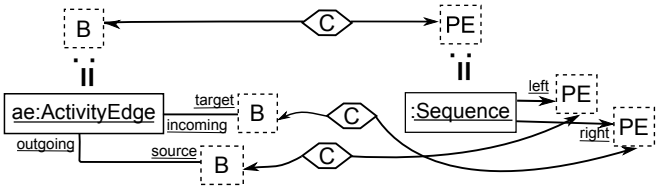


Fig. 8: Transformation rule: Block sequence to process sequence

of type *B* (block) in a source derivation tree, a sequence of process expressions should replace the corresponding hyperedge of type *PE* (process expression) in the related target derivation tree. In addition to relating productions, the transformation rule links source non-terminal edges of type *B* and target non-terminal edges of type *PE* via a 1-to-1 correspondence of type *C*. These correspondences determine which target edge will be replaced by the target production, when the linked source edge is replaced by the related source production. The notion of correspondence is inspired by triple graph grammars (TGGs) [15].

Fig. 9 shows further rules of the transformation definition for our example. It relates productions in the following way:

- Rules 1, 2: Productions for non-terminal (1) and terminal (2) edges representing *models* in the source and target grammars are related. Non-terminal edges in rule 1 are linked via a correspondence later required by rules 2 – 3.
- Rule 3: Production of a non-terminal edge of type *A* representing an *activity* is related to the production of a non-terminal edge of type *P* representing a *process*. The correspondence between model non-terminal edges is kept and other produced edges are linked for later application of rule 4.
- Rule 4: The production of an activity containing a *block* represented by a non-terminal edge of type *B* connected

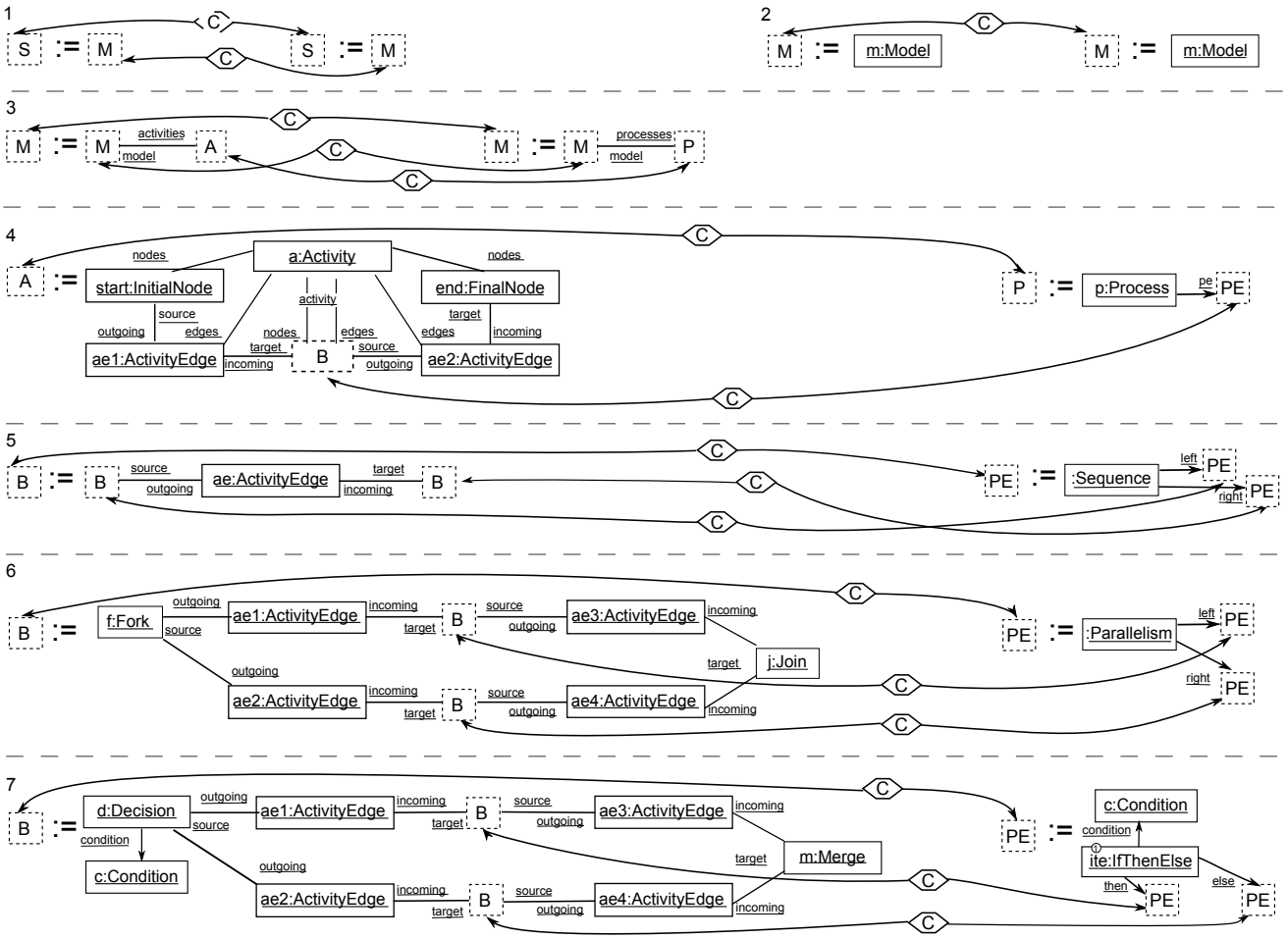


Fig. 9: Transformation definition (fragment): Activity diagram to CSP

to the initial and final nodes is related to the production of a process with a *process expression* represented by the non-terminal edge of type *PE*. The produced non-terminal edges are linked via correspondences which are required later for applying rules 5 – 7.

- Rules 5 – 7 relate different types of activity blocks to different process expressions: A *sequence of blocks* is related to a *sequence of process expressions* (5), a *fork/join* block to a *parallel composition* of process expressions (6), a *decision/merge* block to an *if-then-else* expression (7). Correspondences are created on the same principle as before.

In general, if the transformation definition between two HR grammars G_s and G_t has the following form (extended in Sec. V):

1) In each rule:
   a) a source production $p_s = (n_s, r_s)$ is related to a target production $p_t = (n_t, r_t)$;
   b) left-hand side non-terminals n_s and n_t are linked via a correspondence (start non-terminals S_s, S_t are always linked);
   c) each non-terminal edge in r_t has exactly one corresponding non-terminal edge in r_s,
2) For each source production $p_s = (n_a, r_a)$ in P_s and each correspondence between the edges of type n_a and n_b in some rule (or initial S_s to S_t correspondence), there is exactly one rule relating p_s to a target production p_t, where $p_t = (n_b, r_t)$, to cover all combinations of types of corresponding pairs (n_a, n_b),

and G_s is unambiguous, then the resulting transformation has some important characteristics which we show in Sec. IV. Now, we discuss how the transformation rules we have just defined are executed.

The transformation is executed on a source model in two steps: first, the source model is parsed to get its *leftmost derivation tree*[2] and then the transformation rules are applied to construct the target derivation tree (and the target model).

In the first step, a source model is parsed with respect to the source HR grammar. As HR grammar based parsing is decidable [8], for each source model we either get its leftmost derivation tree or a message that it is not parsable. If a model can not be parsed, it is not in the source language, and hence will be rejected by our transformation returning *not applicable*.

[2]Derivation tree representing leftmost derivation. Leftmost derivation in HR grammars is analogous to the one for string grammars (see [7]).

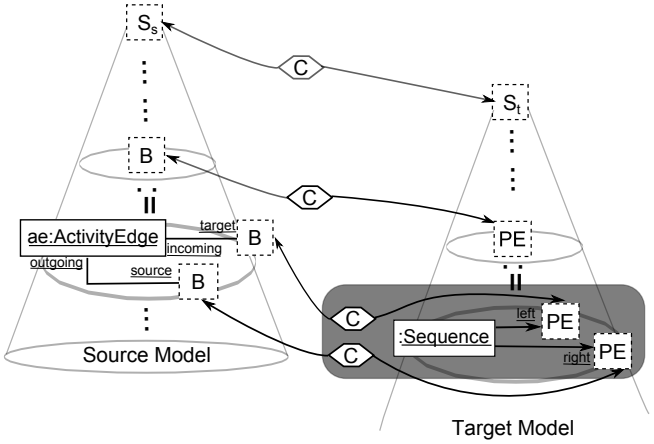


Fig. 10: Transformation execution sketch

In the second step, we build the target derivation tree by first initializing it by the edge of type S_t which has correspondence to the edge of type S_s in the source tree. Next, we iteratively construct the target tree in the following way: we traverse the source tree to find the next non-terminal edge e_s and the source production p_s that has rewritten it. Then, we consider each correspondence c of the edge e_s and find the transformation rule that pairs some p_t with the source production p_s, and where the left-hand side non-terminals are equal to the types of the edges linked by c. Finally, we apply the target production p_t to the target non-terminal edge linked to e_s through c, and create additional correspondences according to the transformation rule. The transformation terminates once the complete source derivation tree has been traversed and all correspondences have been considered. Due to the context-freeness of HR grammars [8], we can use any traversal method.

Fig. 10 sketches the transformation execution, highlighting a single production in the source derivation tree for our example applied to an edge of type *B* with the corresponding edge *PE* in the already created target tree fragment. The dark grey rectangle frames the result of applying the suitable transformation rule 5 (Fig. 9) to the corresponding edge *PE*.

IV. Transformation Properties

Besides allowing for a natural way of transformation definition by mapping logically equivalent concepts onto each other, our method for building model transformations exhibits some important properties. A transformation defined using our approach and fulfilling the criteria mentioned above is, by construction:

a) *terminating* – for any input model, the transformation terminates and returns either a resulting model or *not applicable*,
b) *complete* – all valid, i.e., parsable, models are transformable,
c) *sound* – a valid and transformable model is always transformed into a valid model (parsable in the target grammar),
d) *deterministic* – the output model and its derivation are fully determined by the input model.

For the last property, we require the source grammar to be unambiguous (see Def. 3). Due to the page limit, we only give proof sketches for each of the properties here. In the sketches, we emphasize the conditions on the transformation definition that are sufficient to guarantee these properties. In the following we refer to the source and target HR grammars as G_s and G_t, to the input (source) model as M_s, to the output (target) model as M_t, and to the transformation as τ.

*a) **Termination**:* As described in the last section, τ first parses the input model M_s, yielding a source model leftmost derivation tree T_s (or *not applicable*). Since our approach is based on *context-free* grammars, this is guaranteed to terminate. Then, a *single* G_t production applications is performed by τ for every G_s production application and correspondence c of the edge rewritten by it in T_s. Since these productions' applications are guaranteed to terminate, and the set of correspondences, and T_s *are finite*, the whole process is also guaranteed to terminate.

*b) **Completeness**:* For completeness, we have to show that if $M_s \in \mathcal{L}(G_s)$, $\tau(M_s)$ will not fail. If $M_s \in \mathcal{L}(G_s)$, the first step (parsing) will always succeed, and return T_s. In Sec. III, we have demanded that τ contains a transformation rule for *every production* in G_s and *every correspondence type* an edge rewritten by it might have. Hence, we can transform every production application in T_s into an application in T_t.

Next, we look at the derivation of the output target model M_t via T_t. For completeness, we need to show that this derivation does not fail, i.e., that all target productions are applicable at the place where the transformation wants to apply them. Both G_s and G_t are *context-free*, so the existence of the non-terminal edges ensured by the correspondences is all that is needed to ensure applicability of the target productions. When considering the next source production application, we apply the related target production to the non-terminal edge corresponding to the edge rewritten by the source production. Therefore, since the source production is applicable and τ contains rules for every type of correspondence that non-terminal edge rewritten by it might have, the target production is applicable too. Thus, τ always returns a model for a valid M_s and thus, τ is complete.

*c) **Soundness**:* It remains to be shown that $M_t \in \mathcal{L}(G_t)$. This can be reduced to the question whether the target tree T_t is complete, meaning that no non-rewritten non-terminal edges are left after the application of the transformation.

From requirements in Sec. III, every non-terminal edge produced by a target production is *linked by a correspondence* to a non-terminal edge in T_s. And since $M_s \in \mathcal{L}(G_s)$, all source non-terminal edges produced are eventually rewritten. This implies that all target non-terminal edges produced are also eventually rewritten by the related target productions, i.e., T_t is complete. Thus, we have $M_t \in \mathcal{L}(G_t)$.

*d) **Determinism**:* Since G_s *is unambiguous*, the production of T_s is deterministic. The tree T_s fully determines which rules are applied to construct the target tree, and where. This is because each production on the target side is uniquely determined by the source production and the correspondence

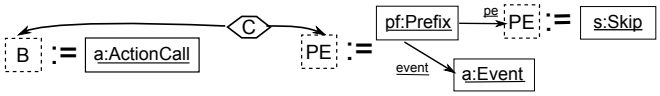


Fig. 11: Transformation rule: Action to an event followed by SKIP process

of the source non-terminal edge it rewrites. The target side of this correspondence also uniquely determines where the target production is applied. We thus, obtain exactly one target derivation tree, and consequently exactly one target model. Therefore, no two different target models can be the transformation result for a single M_s.

V. Extensions

As you might have noticed, we still have to describe a couple of rules to complete the definition of our sample transformation. The remaining rules need to transform: an action call (1) or an internal action (2) into an event followed by a SKIP process, and a loop (3) into an equivalent recursive process, since CSP does not natively support loops. In these two cases we need to relate *two or more* target productions that form a derivation sub-tree to a single source production. To allow this, we extend our approach by relating *derivation sub-trees* instead of relating *single productions* in transformation rules.

Fig. 11 shows such an extended rule for case 1 (2 is analogous): it relates the production for an action call to a sub-tree combining two target productions to create an event followed by a SKIP process. Since all non-terminal edges created by these target productions (one *PE*) have been consumed in the sub-tree, we only require a correspondence for left-hand side non-terminals of the first productions.

Fig. 12 shows the rule for case 3 and, thereby, the second extension of our approach – so called *non-local rules*. This type of rule allows the use of an *additional correspondence* between non-terminals, and uses its target non-terminal as a root of an *additional derivation sub-tree* containing one or more productions. During the transformation execution, when a correspondence of the defined additional type is found and the rule can be applied, the defined additional sub-tree is assigned to its target of this correspondence, replacing it.

Both extensions can be combined within one rule as seen in Fig. 12: it relates the production for a loop to the production for a process reference *L(c)* and uses a correspondence between two model non-terminals (*M*) to create the referenced process *L(x)*. This process consists of an *if-then-else* expression with the first non-terminal edge *PE* in the *then* branch linked to the body of the activity loop represented by the edge *B*, followed by the recursive call to itself, which terminates when the condition *x* fails. The new process is placed directly under the output model edge *M* which makes the result of this rule *non-local* and different from previous examples.

As for the conditions we imposed on our transformation for it to have the desired properties (see Sec. IV), they now need to be adjusted. A full explanation is out of the scope of this paper, thus, we only provide an idea of required adjustments.

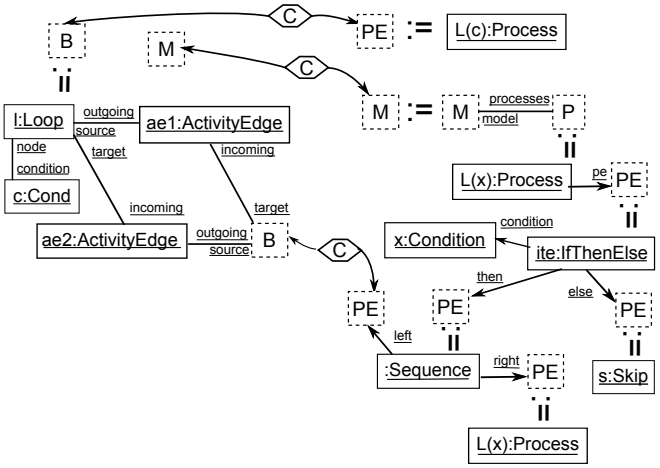


Fig. 12: Transformation rule: Loop to an equivalent recursive process

First, for multiple properties, when a rule relates derivation sub-trees, all non-terminal leaves of the target sub-tree have to have corresponding non-terminal leaves in the source sub-tree. Next, for determinism, the absence of two simultaneously applicable transformation rules need to be guaranteed: either by disallowing rules with the same correspondence and one source sub-tree being a sub-tree of the other, or by introducing rule priorities. Furthermore, for completeness, the presence of rules covering all possible source model derivation steps must be guaranteed: either via coverage analysis or forbidding to use source sub-trees with more than one production. In the case of non-local rules, to guarantee determinism, exactly one correspondence of the specified additional type must exist when the rest of the rule is applicable. Plus, the same correspondence condition as for the main target sub-tree must hold for non-terminal leaves of the additional target sub-tree.

VI. Tool Support & Evaluation

To support the approach we have designed a tool chain involving existing state-of-the-art tools: a graph grammar parser of the AGG [16] framework and the EMorF [17] transformation engine. The chain requires source and target grammars modelled as instances of our HR grammar meta-model extended with grammar-specific non-terminal and terminal type classes. All meta-models are specified in the *ECore* meta-modelling language of the *Eclipse Modelling Framework* [18]. The specified source and target grammars are used as input of our *Eclipse* based *Grammar-Based Model Transformation Framework* that allows the definition of transformation rules and validate their compliance with the conditions from Sec. III. The framework makes use of existing transformation languages and engines like TGGs [15] and EMorF engine by translating the created transformation into its declarative, language (and engine) specific implementation operating on trees. Other combinations of languages and engines like ATL [19] with its engine can also be used.

Fig. 13 shows the described tool chain with the used artefacts for the case when TGGs and the TGG engine (here

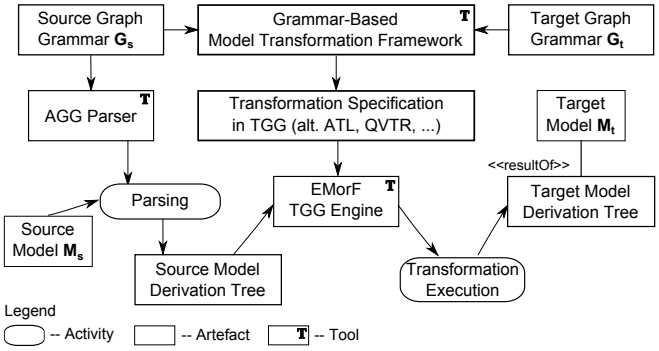


Fig. 13: Tool support for grammar-based transformations with TGGs and EMorF as implementation platform

EMorF) were chosen as the target execution platform for the developed transformation.

To evaluate our approach, we compare it on our running example with the most common transformation development practices in MDE, where meta-models are used in combination with declarative, imperative, and hybrid transformation languages. Meta-models may contain additional classes (e.g., *StructuralActivityNode* in UML [11]) to group (other) classes.

When a source meta-model does not include such structural classes, then the transformations that require this structural information defined on it have to use *imperative constructs*. Often they are also combined with recursion. In our example, such constructs are required to locate the related decision and merge activity nodes to build the corresponding sequence of processes in CSP.

```
rule Decision2IfThenElse extends DefaultNode2Skip {
from d  : AD!Decision
to   ps : ProcessExpression!Sequence (
   left  <- pe,
   right <- d.findMrg(d.outgoing.first().target,0)
 ),

 pe : ProcessExpression!IfThenElse (
   then <- d.outgoing->select(...).first().target,
   else <- ..., condition <- ...
)}

helper def:findMrg(n:AD!ActivityNode,i:Integer)
: AD!Merge =
if n.oclIsTypeOf(AD!Merge) then
 if i > 0 then
  thisModule.findMrg(n.outgoing.first().target,i-1)
 else
  n
 endif
else
 if n.oclIsTypeOf(AD!Decision) then
  thisModule.findMrg(n.outgoing.first().target,i+1)
 else
  thisModule.findMrg(n.outgoing.first().target,i)
 endif
endif;
```

Listing 1: ATL code fragment: Decision / merge block to if-then-else process sequence

Listing 1 shows implementation of this transformation step in a widely used general-purpose hybrid model transformation language ATL [19], which relies on meta-model based language definition. We use the meta-models from Sec. II-A to define the source and target languages of the ATL transformation. We use an ATL matching rule to transform a decision node d to an if-then-else process expression pe (rule *Decision2IfThenElse*). But to link the target expression pe to the next process expression, which should be the result of transforming the merge node corresponding to the decision node d, we have to use recursion to find that merge node. For this purpose, we implement a recursive search helper *findMrg* that follows the path starting in d and skips intermediate decision/merge pairs until it finds the right merge node.

To use this recursive search in ATL, we have to assume well-formedness of activity diagrams, ensured by extra OCL constraints. Such OCL constraints are not always defined and, if they are, they contain recursion and considerably complicate the complete meta-model based language definition. Furthermore, the need for imperative constructs forbids the use of declarative languages and significantly complicates readability and analysis of the transformations. In fact, most techniques aiming to guarantee transformation quality [20], [21] only consider declarative rules and are not applicable here.

Another solution to build structure-based transformations is to use transformation rules to create the required structures in a suitably defined (by meta-model containing structural classes) correspondence model as done in [22] using TGGs. As the previous imperative solution, the correspondence-based solution does not guarantee quality of the developed transformations, and it complicates their understandability and analysis.

If meta-models include structural classes, it is possible to define declarative structure-based rules in TGG-like format with better readability than in the previous solutions, and with more analysis possibilities. Still, as far as we know, there are no approaches showing completeness and determinism even for such declarative transformations. Unlike these common practices, our approach natively supports structure-based transformations keeping their rules graphical and concise (see Fig. 9, rule 7), and guarantees their quality.

Discussion: Transformation rules defined using our approach stay declarative (see Fig. 9) which brings multiple advantages: simpler and more intuitive transformation rules that are easier to understand and maintain especially when grammar productions are represented in concrete syntax; and guaranteed transformation quality properties discussed in Section IV. The only part that stays imperative is the source model parsing. Parsing causes most of the complexity in our method, but its mathematical foundations for HR grammars [8] guarantee termination and predictable worst-case run-time.

The conditions we currently put on transformations defined by our approach to guarantee their quality (Sec. III) might be too restrictive even with the use of extensions (Sec. V) for some purposes. In such cases, a developer can still define structure-based declarative transformations using our approach and employ existing testing and verification techniques to check their quality.

The choice of HR grammars for language definition currently limits our approach to the transformations between

context-free graph-based languages. We plan to address this limitation and take context-sensitive languages into consideration using grammars proposed in [23].

HR graph grammars, which we use to define source and target languages, in general, are more restrictive and complex than pure meta-models (with structural classes). However, when compared to meta-models with OCL constraints enforcing (when possible) the same structural well-formedness constraints on e.g., activity diagrams, HR grammars typed over meta-models present a more concise, intuitive, and powerful way to describe such constraints.

An a meta-model, an HR grammar only needs to be created once per language and, can then be used by any developer for any grammar-based transformation involving this language. The complexity of an HR grammar can affect the complexity of the grammar-based transformation using it, but this is also the case with meta-models.

VII. Related Work

For years, compiler construction benefits from syntax-directed translation [6]. This technique relates single string grammar productions via 1-to-1 relations and requires both grammars to have the same non-terminals, building a very basic version of our approach. Pratt [7] was first to propose to apply this technique to graphs and show that the resulting transformations are deterministic and reversible (under conditions). Our approach extends [7] to n-to-m relations between productions, n-to-m correspondences between non-terminals, and shows additional properties for the developed transformations. Thus, we consider a much larger scope of transformations than the approach from [7].

TGGs proposed in [15] were also inspired by Pratt [7] and were first to contain explicit correspondence nodes. However, the focus of TGGs was on relating context-sensitive productions to support data integration without the consideration of transformation quality properties. In MDE today, TGGs are defined on meta-models and relate model patterns gradually matched during the execution instead of grammar productions. The only approach we are aware of, that uses TGGs with meta-models to define structure based transformation in [22], has already been compared to in out evaluation in Sec. VI.

Halfway between grammars, as in our approach, and meta-models is the transformation development and validation approach proposed in [24]: it relates source model patterns with target productions and states extensive criteria for the transformation quality assurance. Target language in this approach is defined through graph grammar productions. However, it is unclear how the target grammars used there are defined and whether the related target productions can always be applied when the source pattern is found during the transformation execution. In [24] transformation execution strategy is defined manually whereas, in our approach it is automatically obtained during the source model parsing making it less error prone. Like us, authors of [24], consider transformation characteristics: termination, and confluence; however, they do not consider soundness and completeness.

Other approaches like [25], [26], [27] advocate transformation development by-example and by-demonstration, and do not directly focus on structure based transformations. Like us, these approaches recognize the problem of over abstraction of language definition through meta-models, but deal with it using examples to describe corresponding structures of the source and target languages, whereas we use graph grammar productions. By such example-based approaches it is not always clear whether the examples or the transformation should be adapted when the result is not yet satisfactory, how many examples are needed. Whether the developed transformation has desired properties is, as far as we know, not addressed by any of these approaches.

Several other approaches [28], [29] directly apply classic techniques to model transformations. Unfortunately, non of them considers properties of the developed transformations.

In [28] the authors attempt to use TXL [30] – a generic source transformation framework – to develop model transformations. They consider meta-model based languages, and transform them into TXL string grammars. TXL string grammars do not have the expressiveness and visualization advantages of graph grammars we use leading to very limited applicability of the TXL-based approach. Transformations described in TXL are fine-grained with explicit execution policy, which makes them flexible, but also complex and difficult to understand and maintain. This method can be placed halfway between the syntax-directed translation and our approach.

In [29] the authors attempt to simplify transformation development by eliminating the need to learn specialized languages. They regard models and meta-models as abstract data types – abstract structures with operations. On top of the types they define a minimal imperative model transformation language with formal semantics. This approach brings models and transformations into the world of programming, whereas our approach lifts translation techniques to graphs. Both last approaches use meta-models and none of them directly considers high-level structures in languages and transformations based on these structures, as we do. Transformation quality is not take under consideration either.

We consider our approach as the next step towards efficient and quality-aware transformation development that can be realized based on existing state-of-the-art including some approaches described above and the commonly used technologies like ATL [19] and TGG [15].

In general, use of alternative notations for modelling language definition – meta-model vs. graph grammar – raises the issue of integration and interoperability of approaches and tools respectively. Various methods address this issue by defining transformations between the alternatives [31], [32], [33], applying inference to obtain graph grammars [34], or combining them as different views for multi-level modelling [35]. The later option is the one we use.

Finally, we want to point out that the simplest version of our method has been recently successfully used in [36] for semantic-based machine translation in the field of computational linguistics.

VIII. Conclusion

In this paper, we have presented a grammar-based model transformation development approach that allows to naturally consider structures of involved language. We have employed HR grammars to specify source and target languages, and defined transformation rules by relating their productions and adding correspondences between non-terminals. We have shown that model transformations defined using our approach terminate and are sound, complete, and deterministic. We have also presented some extensions of the approach.

Future Work: Currently, we are working on the extension of initial case studies to evaluate our approach and continue improving the tool support. In the future, we look to support computation of attributes, while still keeping the desired transformation properties. When expressiveness of HR grammars is not sufficient, we plan to explore the decidable contextual graph grammars proposed by Drewes in [23].

References

[1] K. Czarnecki and S. Helsen, "Feature-Based Survey of Model Transformation Approaches," *IBM Systems Journal*, vol. 45, no. 3, pp. 621–646, 2006. doi: 10.1147/sj.453.0621

[2] D. D. Ruscio, R. Eramo, and A. Pierantonio, "Model Transformations," in *SFM*, ser. LNCS, M. Bernardo, V. Cortellessa, and A. Pierantonio, Eds., vol. 7320. Springer, 2012. doi: 10.1007/978-3-642-30982-3_4 pp. 91–136.

[3] "Meta Object Facility (MOF) Core Specification." [Online]. Available: http://www.omg.org/spec/MOF/

[4] "Object Constraint Language (OCL)." [Online]. Available: http://www.omg.org/spec/OCL/

[5] "Extended BNF," ISO/IEC 14977, Int. Organization for Standardization, 2001.

[6] A. Aho, M. Lam, R. Sethi, and J. Ullman, *Compilers: Principles, Techniques, and Tools*. Pearson/Addison Wesley, 2007. ISBN 0-321-48681-1

[7] T. W. Pratt, "Pair Grammars, Graph Languages and String-to-Graph Translations," *Journal of Computer and System Sciences*, vol. 5, no. 6, pp. 560 – 595, 1971. doi: 10.1016/S0022-0000(71)80016-8

[8] G. Rozenberg, Ed., *Handbook of Graph Grammars and Computing by Graph Transformation*. World Scientific Publishing Co., Inc., 1997, vol. 1. ISBN 98-102288-48

[9] C. A. R. Hoare, *Communicating sequential processes*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1985. ISBN 0-13-153271-5

[10] E. Syriani and J. Gray, "Challenges for Addressing Quality Factors in Model Transformation," in *ICST*, G. Antoniol, A. Bertolino, and Y. Labiche, Eds. IEEE, 2012. doi: 10.1109/ICST.2012.198 pp. 929–937.

[11] "Unified Modeling Language (UML)." [Online]. Available: http://www.omg.org/spec/UML/

[12] D. Varró, M. Asztalos, D. Bisztray, A. Boronat, D.-H. Dang, R. Geiß, J. Greenyer, P. V. Gorp, O. Kniemeyer, A. Narayanan, E. Rencis, and E. Weinell, "Transformation of UML Models to CSP: A Case Study for Graph Transformation Tools," in *AGTIVE*, ser. LNCS, A. Schürr, M. Nagl, and A. Zündorf, Eds., vol. 5088. Springer, 2007. doi: 10.1007/978-3-540-89020-1_36 pp. 540–565.

[13] G. Besova, S. Walther, H. Wehrheim, and S. Becker, "Weaving-Based Configuration and Modular Transformation of Multi-layer Systems," in *MoDELS*, ser. LNCS, R. B. France, J. Kazmeier, R. Breu, and C. Atkinson, Eds., vol. 7590. Springer, 2012. doi: 10.1007/978-3-642-33666-9_49 pp. 776–792.

[14] S. Walther and H. Wehrheim, "Knowledge-Based Verification of Service Compositions – An SMT Approach," in *Engineering of Complex Computer Systems (ICECCS), 2013 18th International Conference on*, July 2013. doi: 10.1109/ICECCS.2013.14 pp. 24–32.

[15] A. Schürr, "Specification of Graph Translators with Triple Graph Grammars," in *WG*, ser. LNCS, E. W. Mayr, G. Schmidt, and G. Tinhofer, Eds., vol. 903. Springer, 1994. doi: 10.1007/3-540-59071-4_45 pp. 151–163.

[16] G. Taentzer, "AGG: A Tool Environment for Algebraic Graph Transformation," in *AGTIVE*, ser. LNCS, M. Nagl, A. Schürr, and M. Münch, Eds., vol. 1779. Springer, 2000. doi: 10.1007/3-540-45104-8_41 pp. 481–488.

[17] L. Klassen and R. Wagner, "EMorF - A tool for model transformations," *ECEASST*, vol. 54, 2012.

[18] "Eclipse Modeling Framework (EMF)." [Online]. Available: http://www.eclipse.org/modeling/emf

[19] F. Jouault, F. Allilaire, J. Bézivin, and I. Kurtev, "ATL: A model transformation tool," *Science of Computer Programming*, vol. 72, no. 1-2, pp. 31–39, 2008. doi: 10.1016/j.scico.2007.08.002

[20] J. Cabot, R. Clarisó, E. Guerra, and J. de Lara, "Verification and Validation of Declarative Model-to-Model Transformations Through Invariants," *J. Syst. Softw.*, vol. 83, no. 2, pp. 283–302, 2010. doi: 10.1016/j.jss.2009.08.012

[21] F. Büttner, M. Egea, J. Cabot, and M. Gogolla, "Verification of ATL Transformations Using Transformation Models and Model Finders," in *ICFEM*, ser. LNCS, T. Aoki and K. Taguchi, Eds., vol. 7635. Springer, 2012. doi: 10.1007/978-3-642-34281-3_16 pp. 198–213.

[22] C. Lohmann, J. Greenyer, J. Jiang, and T. Systä, "Applying Triple Graph Grammars For Pattern-Based Workflow Model Transformations," *Journal of Object Technology*, vol. 6, no. 9, pp. 253–273, 2007. doi: 10.5381/jot.2007.6.9.a13

[23] F. Drewes, B. Hoffmann, and M. Minas, "Contextual Hyperedge Replacement," in *AGTIVE*, ser. LNCS, A. Schürr, D. Varró, and G. Varró, Eds., vol. 7233. Springer, 2012. doi: 10.1007/978-3-642-34176-2_16 pp. 182–197.

[24] J. M. Küster, "Definition and validation of model transformations," *Software and Systems Modeling*, vol. 5, pp. 233–259, 2006. doi: 10.1007/s10270-006-0018-8

[25] D. Varró, "Model Transformation by Example," in *MoDELS*, ser. LNCS, O. Nierstrasz, J. Whittle, D. Harel, and G. Reggio, Eds., vol. 4199. Springer, 2006. doi: 10.1007/11880240_29 pp. 410–424.

[26] P. Langer, M. Wimmer, and G. Kappel, "Model-to-Model Transformations By Demonstration," in *ICMT*, ser. LNCS, L. Tratt and M. Gogolla, Eds., vol. 6142. Springer, 2010. doi: 10.1007/978-3-642-13688-7_11 pp. 153–167.

[27] Y. Sun, J. White, and J. Gray, "Model Transformation by Demonstration," in *MoDELS*, ser. LNCS, A. Schürr and B. Selic, Eds., vol. 5795. Springer, 2009. doi: 10.1007/978-3-642-04425-0_58 pp. 712–726.

[28] H. Liang and J. Dingel, "A Practical Evaluation of Using TXL for Model Transformation," in *SLE*, ser. LNCS, D. Gašević, R. Lämmel, and E. Wyk, Eds., vol. 5452. Springer, 2009. doi: 10.1007/978-3-642-00434-6_16 pp. 245–264.

[29] J. Irazábal and C. Pons, "Model Transformation Languages Relying on Models as ADTs," in *SLE*, ser. LNCS, M. Brand, D. Gašević, and J. Gray, Eds., vol. 5969. Springer, 2010. doi: 10.1007/978-3-642-12107-4_10 pp. 133–143.

[30] J. R. Cordy, "The TXL source transformation language," *Sci. Comput. Program.*, vol. 61, no. 3, pp. 190–210, 2006. doi: 10.1016/j.scico.2006.04.002

[31] M. Wimmer and G. Kramler, "Bridging Grammarware and Modelware," in *Satellite Events at the MoDELS*, ser. LNCS, J.-M. Bruel, Ed., vol. 3844. Springer, 2006. doi: 10.1007/11663430_17 pp. 159–168.

[32] B. Hoffmann and M. Minas, "Generating Instance Graphs from Class Diagrams with Adaptive Star Grammars," *ECEASST*, vol. 39, 2011.

[33] B. Henderson-Sellers, "Bridging metamodels and ontologies in software engineering," *Journal of Systems and Software*, vol. 84, no. 2, pp. 301–313, 2011. doi: 10.1016/j.jss.2010.10.025

[34] A. Stevenson and J. R. Cordy, "Grammatical Inference in Software Engineering: An Overview of the State of the Art," in *SLE*, ser. LNCS, K. Czarnecki and G. Hedin, Eds., vol. 7745. Springer, 2012. doi: 10.1007/978-3-642-36089-3_12 pp. 204–223.

[35] C. Atkinson, R. Gerbig, and C. Tunjic, "Towards Multi-level Aware Model Transformations," in *ICMT*, ser. LNCS, Z. Hu and J. de Lara, Eds., vol. 7307. Springer, 2012. doi: 10.1007/978-3-642-30476-7_14 pp. 208–223.

[36] B. Jones, J. Andreas, D. Bauer, K. M. Hermann, and K. Knight, "Semantics-Based Machine Translation with Hyperedge Replacement Grammars," in *COLING*, M. Kay and C. Boitet, Eds. Indian Institute of Technology Bombay, 2012, pp. 1359–1376.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1611–1620
DOI: 10.15439/2014F239
ACSIS, Vol. 2

Extended Entity-Relationship Approach in a Multi-Paradigm Information System Modeling Tool

Vladimir Dimitrieski, Milan Čeliković, Slavica Aleksić, Sonja Ristić, and Ivan Luković
University of Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia
Email: {dimitrieski, milancel, slavica, sdristic, ivan}@uns.ac.rs

***Abstract*—In this paper we present a Multi-Paradigm Information System Modeling Tool (MIST) that supports Extended Entity-Relationship (EER) approach to database design. MIST components currently provide a formal specification of EER database schema specification and its transformation into the relational data model, or the class model. Also, MIST allows generation of Structured Query Language (SQL) code for database creation and procedural code for implementing database constraints. In addition, Java code that stores and processes data from the database, may be generated from the class model.**

I. Introduction

IN THE last few decades, a number of information system (IS) development approaches have emerged. Some of methods that are still in use include: Entity-Relationship (ER) data model proposed by Chen [8] with its further extensions, relational data model [9], form type (FT) model [15], and object-oriented model [22].

Throughout our previous research [2], [15], [16], [17], we have developed a tool named IIS*Studio, to allow a usage of FT concepts in the IS design process. IIS*Studio provides an approach to an evolutive and incremental IS development. The approach is purely platform independent and it strictly differentiates between the specification of a system and its implementation on a particular platform. IIS*Studio currently provides the following functionalities: (i) conceptual modeling of database schemas, transaction programs, and business applications of an IS; (ii) automated design of relational database subschemas in the 3rd normal form (3NF); (iii) automated integration of subschemas into a unified database schema in the 3NF; (iv) automated generation of SQL/DDL code for various DBMSs [2]; (v) conceptual design of common GUI models; (vi) automated generation of executable prototypes of business applications; (vii) modeling check constraints and untypical functionalities of business applications [16]; and (viii) reverse engineering of relational databases to FT models [1].

With the emergence of Model Driven Software Development (MDSD) paradigm and Eclipse Modeling Project (EMP) [10] with an appropriate tooling, we decided to implement some of the existing IIS*Studio functionalities using these technologies. The motivation came from our intention to provide designers of ISs a possibility to use Eclipse environment and thus reduce steep learning curve for new users of IIS*Studio. Therefore, we have developed a domain specific language (DSL) allowing a specification of IS form type models. A detailed specification of this language is presented in [7].

One of the main goals of IIS*Studio is to provide a designer conceptual modeling by creating platform independent models. As designers mainly use other approaches than FT for this purpose, we have decided to support not only FT concepts, but also Extended Entity-Relationship (EER) data model, as a commonly used, traditional approach. Therefore, we have developed a DSL for the specification of EER database schema specifications, named EERDSL. Together, FT and EER are the approaches of our new Eclipse-based tool both providing conceptual database schema modeling. The tool is named Multi-paradigm Information System modeling Tool (MIST). In MIST, both approaches may be used simultaneously. For both FT and EER models, we provide in MIST a transformation into a relational data model. In our previous research on database reengineering approaches [1], we have developed a transformation from a relational data model to a FT model, named relational to form type transformation (R2FT). R2FT is used to transform an EER model into an FT model via relational data model.

As EER approach is present in almost every book on databases, we believe that MIST may also be used for educational purposes, such as learning about: (i) EER concepts and developing a database specification at the conceptual level; (ii) transformations of EER to relational database schema specifications; (iii) transformations of EER to class models; and (iv) MDSD approach by means of the EER approach the students are familiar with, since it is extensively taught in the previous database courses.

In this paper we present the architecture of MIST. It comprises several components that support not only conceptual modeling with FT and EER approaches, but also code generation. The main focus in the paper is on tool components supporting EER approach and transformations from EER to relational and class models. A detailed presentation of the code generators is out of the scope of this paper. Let us just notify here that we have developed both SQL and Java code generators. The SQL Generator provides SQL statements for creating database tables and all basic types of constraints according to SQL ANSI standard. Besides, the code of inverse referential constrains, as they are defined in [3], is generated. As it has to be implemented in a procedural way, different

Research presented in this paper was supported by Ministry of Education, Science and Technological Development of Republic of Serbia, Grant III-44010.

generators are needed for each target database management system (DBMS). Our tool currently supports generation of PL/SQL statements for Oracle DBMS. Our Java code generator provides a generation of Java classes from a class model. Generated Java classes are used in Java programs for storing loaded data from a database.

Apart from Introduction and Conclusion, the paper is organized in four sections. In Section 2 we present the architecture of MIST, while in Section 3 we present EER, Relational and Class meta-models. The aforementioned meta-models are used in the following transformation specifications: (i) the EER data model to relational data model transformation, named EER2Relational and (ii) the EER data model to class model transformation, named EER2Class. These transformations are presented in Section 4. In the same section we present results of applying the aforementioned transformations in our example and the excerpts from generated SQL and Java code. In Section 5, we present related work.

II. The Architecture of MIST

In this section we present the architecture of MIST. Its global picture is depicted in Fig. 1. MIST comprises the following components: FTDSL, Synthesis, Business Application Generator, EERDSL, EER2Rel, EER2Class, SQL Generator, Java Generator, and R2FT. In the following text, we explain each of the components from Fig. 1.

FTDSL component allows designers to specify a platform independent model (PIM) of an IS. FTDSL comprises Ecore meta-model specification of FT PIM concepts and a textual DSL based on these concepts. With the DSL a designer may specify a database schema of an IS, business applications and their graphical user interfaces (GUIs). After an IS is specified at the PIM level, the Synthesis component is used to generate a model of a relational database schema. The Synthesis component implements an improved synthesis algorithm, as it is presented in [14]. First, the component takes a form type specification and transforms it to a Universal Relation Schema (URS) specification. URS and all its benefits are presented in [14]. The synthesis algorithm takes the URS specification and produces a relational database schema as an output. As the FT component may be used to specify business applications of an IS, the MIST architecture includes a Business Application Generator component. This component takes a FT model as an input and generates Java code of a modeled business application. As the specification is enriched with GUI details, the generated application prototype may be executed and used to perform basic CRUD operations over the database.

EERDSL component provides a conceptual specification of an IS database model. Unlike FTDSL, EERDSL is used to specify IS database models only, without specification of business applications and their GUIs. We have created both textual and graphical notations for EERDSL. The textual notation was developed using the Xtext tool, while the Eugenia tool was used to develop the graphical notation. One of the benefits of having a textual notation is that textual editors may be used as an alternative option for more experienced users, or when the Eclipse environment is unavailable. The textual notation also allows a usage of commonly used textual version control systems to provide a better collaboration inside the developer team. Most database designers, however, are using some EER graphical notation. Several different graphical notations for the EER approach exist. Here we have implemented the notation presented by Thalheim in [21]. Both graphical and textual notations may be used by a designer simultaneously, while specifying an EER model. By this, two different viewpoints over the same model are provided in MIST.

EER2Rel component of MIST provides a transformation of EER model to a relational data model. Models being transformed conform to the EER meta-model and relational meta-model, respectively. The meta-models are presented in the following section. The relational data model may be further used in the process of SQL code generation. For this purpose,

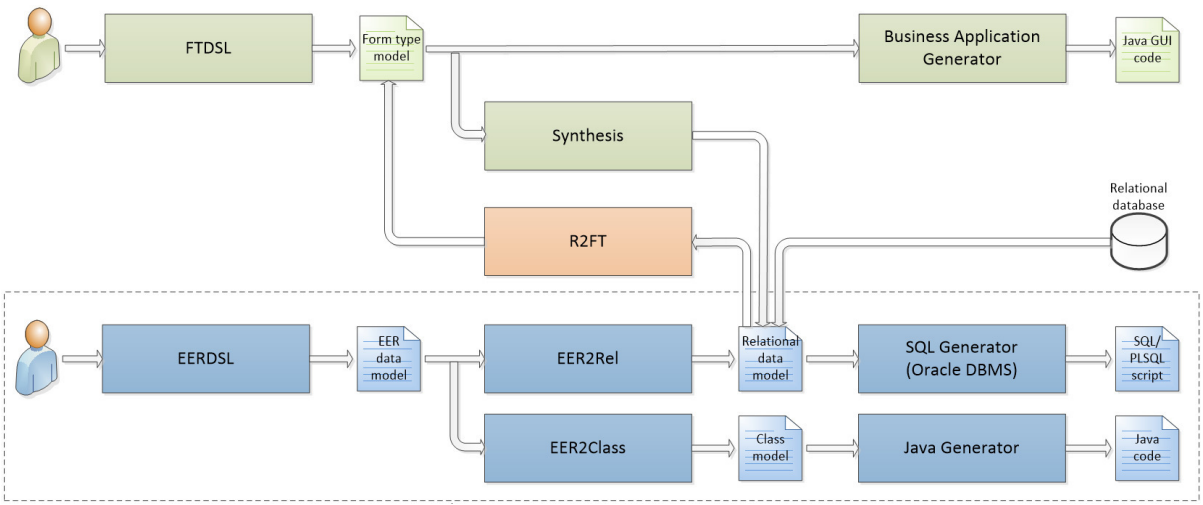


Fig. 1. The architecture of MIST

the SQL Generator component is developed. Currently, it provides generating SQL code for Oracle DBMS.

EER2Class component of MIST provides a transformation of an EER model to a class model. The class model conforms to the class meta-model presented in the Section 3. The class model may be used to generate code in some of the object-oriented programming languages. Our Java Generator component is used to generate Java code from the provided class model.

In order to provide reverse engineering of the relational database model to the FT model, R2FT component has been developed. The component comprises a transformation specification from the relational data model to the FT model. Also, this component may be used to transform an EER data model to the FT model through the relational data model.

III. EER, Relational, and Class Meta-Models

In this section we present in more detail meta-models used in EER approach. In Subsection A, we present our EER meta-model and introduce an example used throughout the paper. Our goal is to test the approach on this example. Further evaluations of the approach are also possible, such as comparing its quality and efficiency with the FT approach. However, this comparison is not presented in this paper due to the space limitations. In Subsection B, we present the relational meta-model, while in Subsection C we present the class meta-model.

A. EER meta-model and an example

In this subsection we present the EER meta-model depicted in Fig. 2. In the rest of this section we present the names of meta-model and model concepts in italic. This meta-model represents the abstract syntax of EERDSL used for specifying data models at the conceptual level. The root concept in our meta-model presented in Fig. 2 is *EERModel*. Each EER model comprises one or more *Entities* and zero or more *Relationships*, *Gerunds*, and *Domains*.

Entity concept is used to specify a class of observed real world entities in the IS being designed. In some approaches, the *Entity* concept is named as *Entity Type* concept. According to [21], we adopt the name *Entity*.

Each entity has zero or more attributes that are modeled by *Attribute* class. Attributes represent properties of real world entities that are of importance for the specified IS. For each attribute, a domain is specified. A domain represents a specification of possible values that can be assigned to an attribute and it is modeled using *Domain*. A domain must be based upon a primitive domain, such as integer, string, real, boolean, date, and time. The primitive domain is modeled by an enumeration *PrimitiveDomain*. An assignment of a domain to an attribute is modeled by *AttributeDomain*. For each attribute, length and default value may be specified. This way of restricting domains allows their reusability. Therefore, domains may be specified once at the level of EER model, and reused and further restricted at the level of attributes. An entity may have one or more keys modeled by *Key*. Each key comprises one or more attributes of the entity. Only one key may be declared as the primary key. This is modeled by *primaryKey* association.

In the meta-model from Fig. 2, different types of relationships between entities are modeled by: *Relationship*, *ISA*, *Categorisation*, and *IdentificationDependency*. An n-ary relationship between entities is modeled by *Relationship*. For each entity its role, minimum, and maximum cardinalities must be specified for each relationship. Minimum cardinality may be provided with the values of zero or one, while a maximum cardinality may be provided with the values of one or more. Entity role and cardinalities are modeled by *RegularEntity*. Each relationship may have zero or more attributes. IS-A

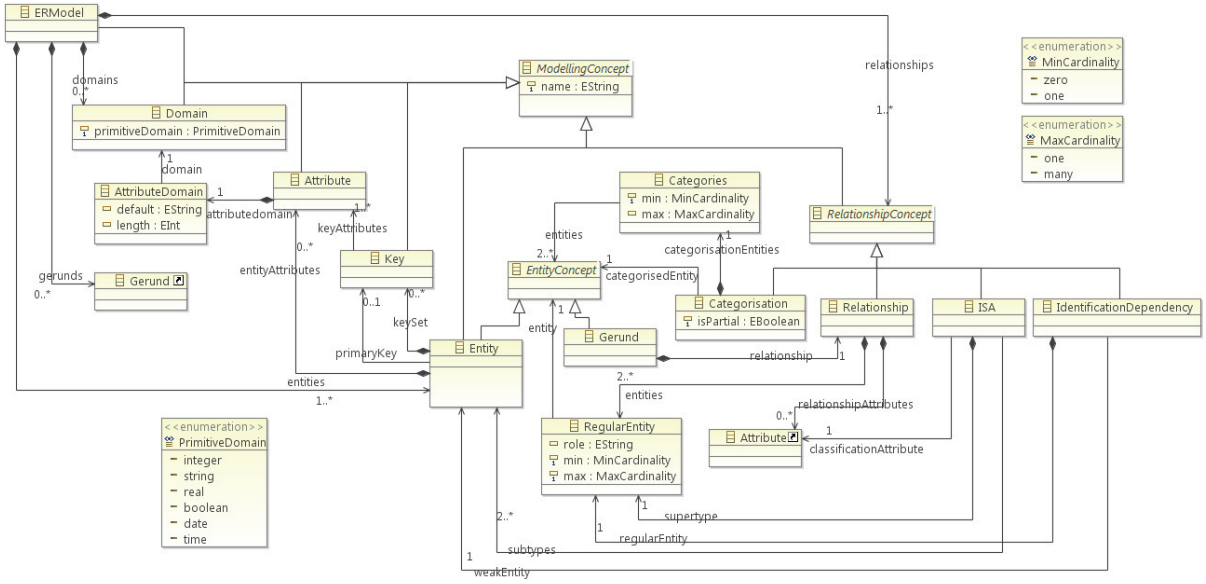


Fig. 2. EER meta-model

relationship, modeled by *ISA*, represents a specialization of entities. For each IS-A relationship a single supertype entity should be provided. This is modeled by *supertype* association. Also, for each IS-A relationship, subtype entities should be provided. This is modeled by *subtype* association. Categorization relationship represents a classification, i.e. typization relationship between entities. It comprises two or more category entities, modeled by *Categories*, and a single categorized entity. Identification dependency relationship concept is modeled by *IdentificationDependency*. A weak entity is modeled by *weakEntity* association, while regular entity is modeled by *regularEntity* association.

An n-ary relationship may participate as an entity in another relationship. Such relationship is called gerund and modeled by *Gerund*. A gerund may take a role of a regular entity in an n-ary or identification dependency relationship. Also, it may take a role of a specialized entity in a specialization, and category entity in a classification.

In Fig. 3, we present our example specified using a textual notation of EERDSL. The textual notation is chosen as it is rarely encountered while specifying EER models. However, due to the limited space, we omit repetitive constructs from the textual specification. We model a part of a Faculty IS, named *FacultySystem*, responsible for storing data about students and their grades.

Students are modeled with an entity named *Student*. Each student has four attributes: *studentID*, *studentsYear*, *studentName*, and *studentSurname*. Primary key named *keyStudent* comprises *studentID* attribute. Subjects are modeled with *Subject* entity having two attributes: *subjectID* and *subjectName*. Primary key named *keySubject* comprises *subjectID* attribute. Teachers of a faculty are modeled as *Teacher* entity and are described by *teacherTitle*, *teacherID*, *teacherName*, and *teacherSurname* attributes. *TeacherID* is the only attribute in *keyTeacher* primary key. Relationship between teachers and subjects they teach is modeled by *TeachesClasses*. Each teacher may teach one or more subjects, while a subject may be taught by one or more teachers. The relation between students and subjects is modeled as *Takes* relationship. Each student may enroll one or more subjects, while a subject may be enrolled by zero or more students. Relationship *Grades* models students' grades given by teachers. As only a teacher that teaches a subject may grade students enrolled on that subject, relationship *Grades* must relate relationships *Takes* and *TeachesClasses*. Therefore, relationships *Takes* and *TeachesClasses* must be represented as gerunds. Each student that takes a subject may be graded by exactly one teacher teaching the subject. A teacher teaching a subject may grade zero or many students of the subject. For each grading *examDate* and *grade* attributes are specified.

```
EERModel FacultySystem {
 domains {
  Domain int primitiveDomain integer,
  ... // omitted domains: varchar, Date, and Time
 }
 entities {
  Entity Student {
   attributeSet {
    Attribute studentName domain varchar(20),
    Attribute studentSurname domain varchar(20),
    Attribute studentID domain int,
    Attribute studentYear domain int
   }
   keySet {
    keyStudent (studentID)
   }
   primaryKey  keyStudent
  },
  ... // omitted entities: Subject and Teacher
 gerunds {
  Gerund Relationship Takes {
   entitiesInRelationship {
    Student (one,many),
    Subject (zero,many)
   }
  },
  ... // omitted gerund for relationship TeachesClasses
 relationships {
  Relationship Grades {
   entitiesInRelationship {
    Takes (one, one),
    TeachesClasses (zero,many)
   }
   attributeSet{
    Attribute grade domain int,
    Attribute examDate domain Date
   }
  }
 }
}
```

Fig. 3. Example of Faculty IS specification in EERDSL

B. Relational meta-model

In this subsection we present our relational meta-model. The root concept of the relational meta-model presented in Fig. 4 is *Database*. Each database schema comprises *Tables*. *SystemDataTypes* represent predefined data types built into each DBMS, while *UserDefinedDataTypes* represent user defined restrictions on existing data types. Each table comprises one or more columns, modeled with *Column*, and constraints inheriting the abstract concept *Constraints*. At the level of a relational database specification, for each table following constraints may be specified: (i) primary key constraint modeled by *PrimaryKeyCon*, comprising an array of columns; (ii) unique constraint modeled by *UniqueCon*, comprising an array of columns having a unique combination of values; (iii) foreign key constraint modeled by *ForeignKey*, comprising: an array of columns, referenced table, and primary key constraint or unique constraint of referenced table; and (iv) check constraint modeled by *CheckCon*, comprising a logical expression.

C. Class meta-model

In this subsection we present our class meta-model. We have modeled only the most basic concepts to specify elements for a representation of data in object oriented programs. The root concept of the class meta-model depicted in Fig. 5 is *ClassModel*. All concepts are grouped into packages modeled by *Package*. A package comprises zero or more *Classes* and

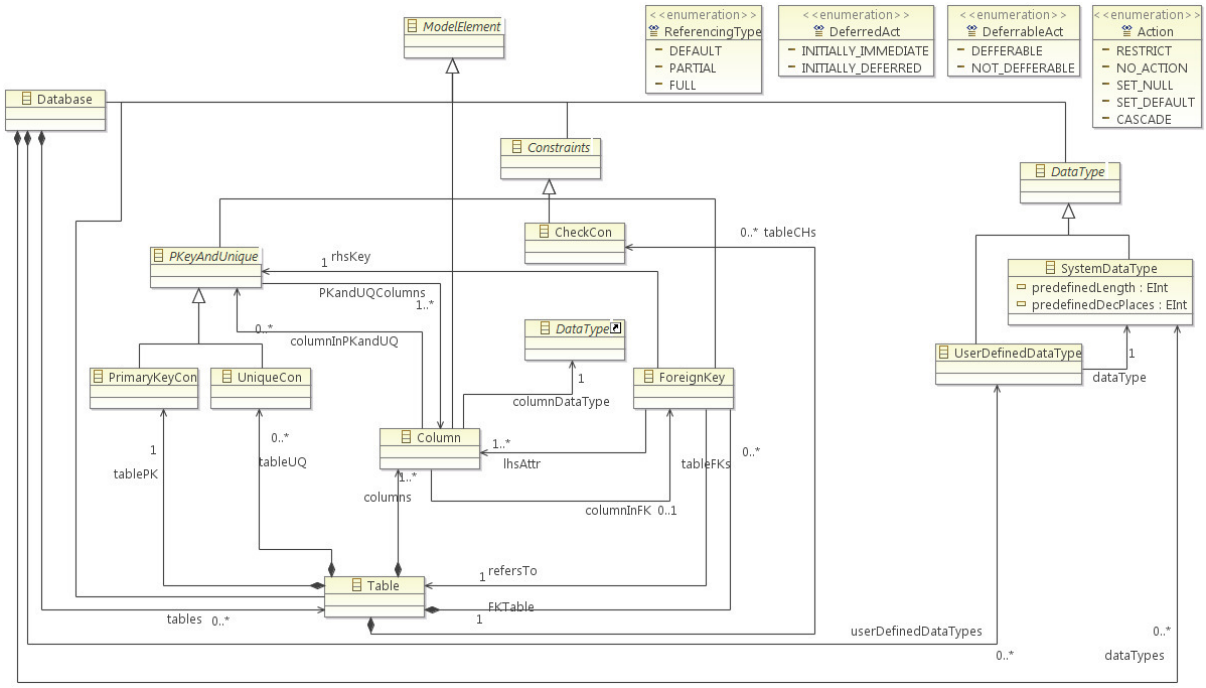


Fig. 4. Relational meta-model

other *Packages*. Each class comprises zero or more fields, modeled by *Attribute* concept, and methods modeled with *Function* concept. Data types are modeled by *Type*. Classes and types inherit the abstract *Classifier*. This concept is specified in order to allow attributes and functions to have both primitive and class types as their type, and the return type respectively. Finally, for each attribute, class, and function an access modifier should be provided. Possible access modifier values are modeled as an enumeration, which comprises following values: private, protected, default, and public.

IV. From EER Data Model to Generated Code

In this section we present two transformations: EER2Rel for transforming an EER model into a relational data model and EER2Class for transforming and EER model into a class model. Transformations are specified in ATL transformation language (ATL) [13]. We present ATL code for the most representative transformation rules. In this section we present the results of applying transformations on the example. Finally, generated code fragments of our example are presented at the end of this section.

Once an EER database model is specified using EERDSL, it may be transformed to the corresponding relational model. In Table I, the first two columns represent all of the corresponding concepts between the EER and relational meta-models. Based on these correspondences, concrete ATL transformation rules are specified.

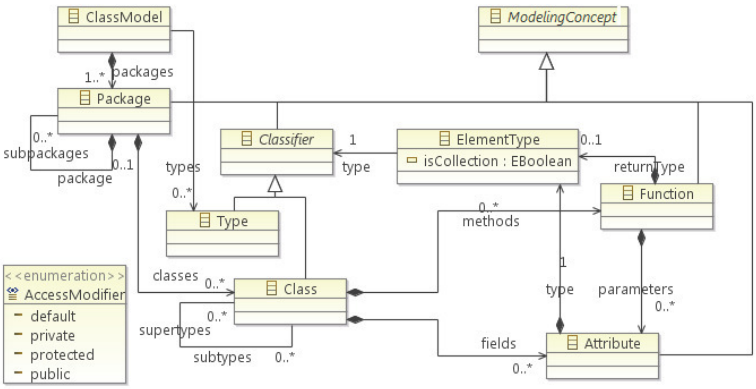


Fig. 5. Class meta-model

TABLE I
TRANSFORMATION ELEMENTS FROM AN EER MODEL TO RELATIONAL AND CLASS MODELS

EER model	Relational model	Class model
EERModel	**Database**	**ClassModel, Package**
Domain	**UserSystemType**, **SystemDataType**	**Type**
Entity	**Table**	**Class**, **Function** (Constructor)
Attribute	**Column**	**Attribute**, **Function** (Get/Set)
Key	**PrimaryKeyCon** (if the key is entity's primary key)	-
Relationship	**Table** (if all maximum cardinalities are *many*)	**Class** (if all maximum cardinalities are *many*), **Function**
Gerund	**Table**	**Class**, **Function**
Identification dependency	**Columns** of the propagated primary key are included into the primary key of the table created from a weak entity. **Foreign key** is created to reference the table created from regular entity.	For a class created from a weak entity, an **object-member** is created referencing a class created from a regular entity.
ISA	**Columns** of the propagated primary key are included into the primary key of the table created from a subtype entity. **Foreign key** is created to reference the table created from supertype entity.	A class created from a subtype entity **inherits** a class created from a supertype entity.
Categorisation	**Foreign key** is created to reference the table created from category entity.	For a class created from a categorized entity, **object-member** is created referencing a class created from a category.

Each EER database concept may be transformed directly to a relational database concept. A system data type is created from EER primitive domains and for each restricted domain, specific user system type is created. Tables are created from three different concepts in EER model: *Entity*, *Gerund*, and *Relationship*. All entities and gerunds are transformed directly into tables, while only the relationships that have maximum cardinality of *many* on both sides, are directly transformed into tables.

In Fig. 6 we present a transformation rule for transforming entities into tables. When transforming an entity to a table, the table name is the same as the name of the entity. For each entity, the following types of columns may be created: (i) columns created from attributes of the entity being transformed; (ii) columns created from attributes of a relationship having a maximum cardinality *one* on the transformed entity side; (iii) columns created from primary key attributes of related entities, where the entity being transformed is on *one* side of a relationship; (iv) columns created from primary key attributes of related categories in categorization relationships; (v) columns created from primary key attributes of related regular entities in identification dependency relationships where transformed entity plays a role of weak entity; and (vi) columns created from primary key attributes of related supertypes in ISA relationships. In this transformation rule, the aforementioned columns are created in both declarative and imperative way. The declarative part of the ATL rule may be used to create the first two types of columns. The creation of these columns does not require the creation of foreign keys and assigning the columns to a foreign key. Therefore, such columns do not require additional references to be set and as such they may be created in a fully declarative way. The imperative section of the ATL rule must be used for the next four types of columns. These columns represent copies of already created primary key columns in other tables. As such, foreign keys must also be created and *lhsAttr* and *columnInFK* references must be assigned to the created foreign keys and columns respectively. A foreign key must also reference a table containing original primary key columns with *refersTo*

```
rule Entity2Table {
  from e : ER!Entity
  using {--omitted variables used in do section }
  to t : RDBMS!Table (
    name <- e.name,
    columns <- e.entityAttributes -> union(
      e.getConnectedRelationshipsForEntity() ->
      collect(e1|e1.relationshipAttributes) ->
      flatten()),
    tablePK <- e.primaryKey
  )
  do {
    --FKs, columns from entities related with Mto1
    --This entity is on one side
    for(r in e.getConnectedRelationshipsForEntity())
    {
      t.tableFKs <- thisModule.CreateForeignKeys(
        r.getConnectedRegularEntity().entity, t,
        r.getConnectedRegularEntity().entity.
          getPrimaryKeyAttributes(),
        r.getConnectedRegularEntity().min=#one,
          r.name);
    }
    --FKs and columns from regular entites in ID rel
    ids <- e.getRegularEntitiesForWeak();
    thisModule.cols <- Sequence{};
    for (reg in ids) {
      t.tableFKs <- thisModule.CreateForeignKeys(
        reg, t, reg.getPrimaryKeyAttributes(),
        false,'ID');
    }
    pkcolumns <- pkcolumns->append(thisModule.cols)
      -> flatten();
    --FKs and columns from supertypes in IS-A rel.
    isaIds <- e.getSupertypes();
    thisModule.cols <- Sequence{};
    for (sup in isaIds) {
      t.tableFKs<-thisModule.CreateForeignKeys(sup,
        t, sup.getPrimaryKeyAttributes(), false,
        'ISA');
    }
    pkcolumns <- pkcolumns->append(thisModule.cols)
      -> flatten();
    --create PK from appropriate columns
    if (t.tablePK.oclIsUndefined()) {
      t.tablePK<-thisModule.Key2PKMtoN(pkcolumns,
        t.name);
    } else {
      t.tablePK.PKandUQColumns <- pkcolumns;
      for (pkc in pkcolumns) {
        pkc.columnInPKandUQ <- t.tablePK;
      }
    }
  }
}
```

Fig. 6. Entity to table transformation

reference. As the created foreign key is a part of the newly created table, the foreign key must be assigned to *FKTable* reference. The creation of these columns and foreign keys is provided in a form of the ATL called rule. We have specified a *CreateForeignKey* called rule which creates columns from primary key columns of a related table and a foreign key referencing that table. This rule adds both created columns and the foreign key to the newly created table and populates all of the aforementioned relationships between foreign key, tables and columns. Due to the space limitations, we have not presented *CreateForeignKey* rule in this paper, but its explicit invocation is presented in Fig. 6 Arguments that are provided are as follows: (i) entity transformed to a table being referenced with a foreign key; (ii) table containing the foreign key, i.e. referencing table; (iii) primary key containing attributes that are transformed to referenced columns; (iv) a boolean value specifying whether an inverse referential integrity should exist; and (v) a string value to be appended to the names of newly created columns in order to avoid name conflicts with previously created columns.

Besides being foreign key columns, columns created from a supertype and a regular entity in ID relationship must be a part of a primary key. In presented ATL rule, as to collect all primary key columns we use a global attribute helper named *cols* and a local variable named *pkcolumns*. *CreateForeignKeys* uses *cols* to return created columns as the return value is used to return a created foreign key. Returned columns are then appended to the other columns contained in *pkcolumns* variable which is local variable declared in the ”using” section of an ATL rule. At the end of an imperative section of *Entity2Table* rule, primary key of a table is set. If a primary key has already been created in the declarative part of the rule, *pkcolumns* are simply appended to *PKandUQColumns* of the existing primary key. The *columnInPKandUQ* relationship is set for each column referencing a primary key the column is a part of. However, a primary key may not be created in the declarative part of the rule. That may be the case if the entity from EER diagram does not have a specified primary key, e.g. a subtype in ISA relationship. For those entities a called rule *Key2PKMtoN* is invoked which creates a primary key and sets all of the appropriate references between columns and the primary key.

Unlike entities, which are all transformed into tables, only relationships that are not contained in gerunds and that have all maximum cardinalities of *many* are transformed into tables. ATL rule that transform such relationships into tables is presented in Fig. 7 Attributes belonging to the relationship are transformed into the table columns. However, when a relationship between two or more entities is transformed from EER to a relational specification it must reference all primary key columns from all related entities. Similarly to the aforementioned *Entity2Table* rule, former columns are created in a declarative way while latter ones are created in an imperative way.

Finally, tables are also created from gerunds. Each gerund encapsulates a single relationship and it may be a part of another relationship. As such, a gerund has all traits of relationships and entities. Therefore, *Gerund2Table* transformation rule is a combination of both *Entity2Table* and *Relationship2Table* rules. It should be noted that a gerund may not be a subtype of inheritance relationship, a weak entity of an identification dependency relationship, or a categorized entity in a categorization relationship. Therefore, the imperative code from *Entity2Table* creating foreign keys and columns from these kinds of relationships is not a part of *Gerund2Table* rule. The code of *Gerund2Table* is not presented in this paper due to the space limitations.

```
rule Relationship2Table {
  from
    --M:N relationship not contained in a gerund
    r : ER!Relationship (not r.isGerund() and
        r.areMaxCardinalitiesMany())
  using {--omitted variables used in do section }
  to
    t : RDBMS!Table (
      name <- r.name,
      columns <- r.relationshipAttributes
    )
  do {
    keys <- r.getConectedKeyAttributesSequence();
    thisModule.cols <- Sequence{};
    for (k in keys) {
      pks <- pks.append(k -> first() ->
        getPKForAttribute());
      t.tableFKs <- thisModule.CreateForeignKeys(
        k -> first() -> getParentConcept(),t,k,
          r.relationship.isIrc(k -> first() ->
            getParentConcept().name)
        k -> first() -> getParentConcept().name);
    }
    t.tablePK <- thisModule.Key2PKMtoN(
      thisModule.cols, t.name);
  }
}
```

Fig. 7. Relationship to table transformation

The transformation of an EER model to a class model is very similar to the *EER2Relational* transformation. In Table I, the first and the third column represent all of the corresponding concepts between the EER and class meta-models. Instead of tables, classes are created and instead of columns each class has attributes. One notable difference is the lack of constraint concepts. In the class model, primary keys are not created which eases the transformation specification. Instead of creation of foreign keys, an attribute of referenced class type is created. If a relationship has a maximum cardinality of *many* then the attribute is a collection of referenced classes. The second difference is the existence of functions. For each class a parametrized constructor is created and for each class attribute, get and set methods are created.

In Fig. 8 we present results of the transformation of our example. In the leftmost part of the figure, we present the same example as in Fig. 3 opened in the Eclipse “Sample Reflective Ecore Model Editor”. This model has served as the input model for both EER2Rel and EER2Class transformations.

Output of EER2Rel is presented in central part of Fig. 8. Each of three EER entities, *Student*, *Teacher*, and *Subject*, has been transformed into a table with the same name.

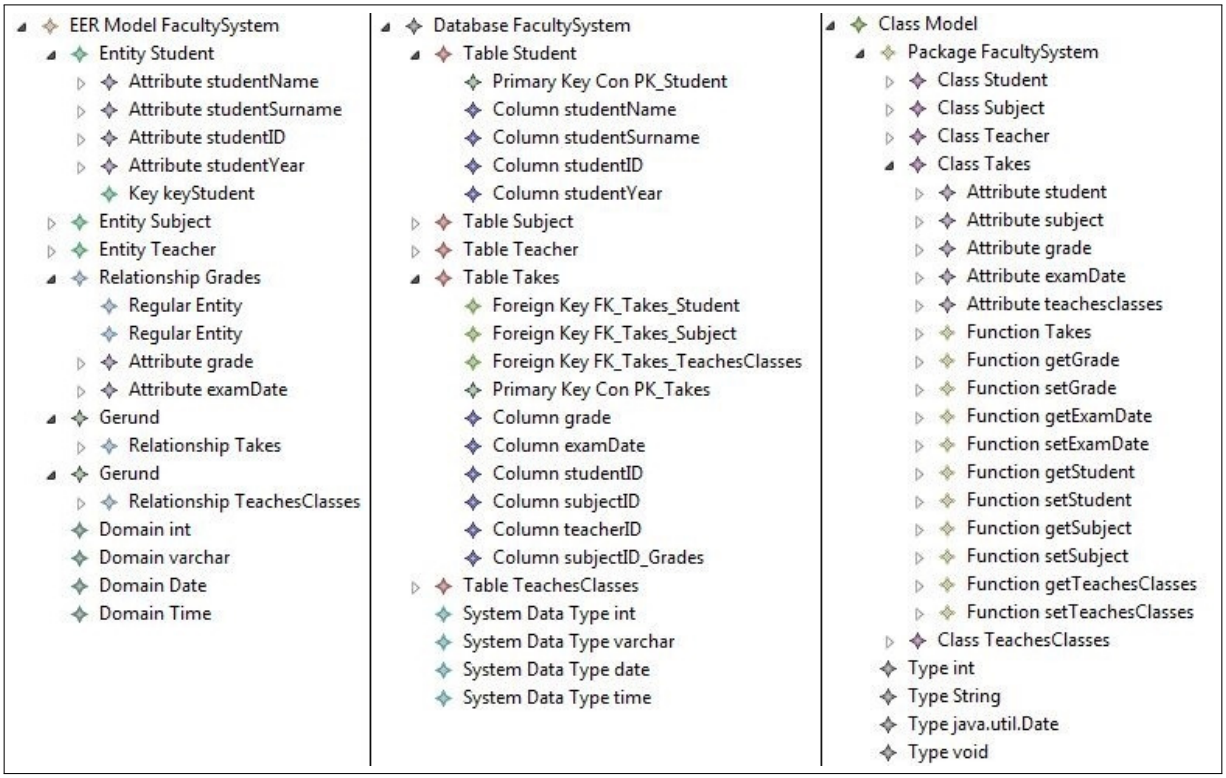


Fig. 8. Faculty IS example: the EER model, the relational database model, and the class model

Here, we present table *Student* in more details. The table contains columns created from *studentName*, *studentSurname*, *studentYear*, and *studentID* attributes of *Student* entity in EER model. From *keyStudent* in *Student* entity, *PK_Student* primary key is created. All the references between the key and attributes are preserved in a form of references between primary key and appropriate columns. Due to a lack of space, these references are not shown in Fig. 8. The gerund *TeachesClasses* is transformed into a table with the same name. Its columns are created from primary key columns of related *Teacher* and *Subject* tables. Appropriate foreign keys are also created for these columns. A primary key of *TeachesClasses* table represents a union of all primary key columns from *Teacher* and *Subject* tables. However, more interesting case of gerund to table transformations is the case of *Takes* gerund. Its relationship relates entities *Student* and *Subject*. Just like in the case of *TeachesClasses* gerund, the resulting table, named *Takes*, will have columns and appropriate foreign keys that reference primary key columns from both *Student* and *Subject* tables. These columns are *studentID* and *subjectID* with their foreign keys *FK_Takes_Student* and *FK_Takes_Subject*, respectively. Gerund *Takes* is related with the gerund *TeachesClasses* with the relationship *Grades*. This relationship is not to be transformed into a separate table as it has a maximum cardinality of *one* on the gerund *Takes* side. Instead, all of the relationship attributes from the *Grades* relationship will be created in *Grades* table as its own columns. Primary key attributes of *TeachesClasses* are to be referenced form appropriate columns in *Takes* table. Therefore, *Takes* table has *grade* and *examnDate* created from the attributes of *Grades* relationship. Columns *teacherID* and *subjectID_grades* with foreign key *FK_Takes_TeachesClasses* reference primary key attributes from *TeachesClasses* table. Let us note that *subjectID* from *TeachesClasses* table had to be renamed in *Takes* table as there was already a column with that name. We chose to append the name of a relationship through which the column was created in the table at the end of a column's name. As names of relationships are unique in the EER model, renamed attributes will all have unique names in their tables.

Output of EER2Class is presented in the rightmost part of Fig. 8. Only a detailed overview of a *Takes* class is given. Similarly to the relational model, a class model has five classes: *Student*, *Subject*, *Teacher*, *Takes*, and *TeachesClasses*. First three classes are created in a straightforward manner from the entities with the same name. *TeachesClasses* is created from *TeachesClasses* relationship. As relationship *Takes* relates *Student* and *Subject* entities, table *Takes* has two object-members: *student* and *subject*. For the same reasons as in the relational model, class *Takes* has two attributes from *Grades* relationship and an object-member representing *Teaches Classes*. These two attributes are grade and *examnDate* and an object-member is named *teachesclasses*. For object-members that represent a collection, a Boolean value of *isCollection* attribute should be set to *true*. For example, object member *teachesclasses* is of

the collection type, as it is created from *Grades* relationship that has a maximum cardinality of *many* on the side of *TeachesClasses*. In addition to attributes, a parametrized constructor and get and set methods are created. For each method a body is automatically generated in a form of a string.

In Fig. 9 we present excerpts from the generated SQL and Java code. In the left part of the figure we present a statement for creating *Takes* table as well as statements that add constraints to this table. In the same figure we present a part of the procedural code that handles inserting new tuples into tables on top of which an inverse referential constraint is enforced. In our example these two tables are *Subject* and *TeachesClasses*. In order to allow simultaneous insert in both tables, a view *View_Subject_TeachesClasses* is created that will allow such insert. An algorithm that handles such insert is a part of a generated trigger named *TRG_IRI_Subject_TeachesClasses_View*. Finally, in the right part of Fig. 9 we present generated Java code for the *Takes* class. We have omitted repetitive code of get and set methods.

V. Related Work

Since Chen proposed ER data model in [8], many papers have been published discussing ER data model, its features, extensions, and practical application. We found only one paper presenting EER data model implementation in the Eclipse environment using MDSD principles. In [12], the authors present EER meta-model and the EERCASE tool based upon it. The tool provides all of the EER concepts represented with Elmasri-Navathe's graphical notation [11].

Our tool is also integrated with the Eclipse environment, so as to provide beginners with an easy of use tool as they are already familiar with the environment. EERDSL component of our tool provides all concepts from the EER approach. All concepts are represented with widely used graphical notation presented also by Thalheim in [21]. Apart from graphical notation, provided by all of the aforementioned tools, our tool also provides EER modeling with a textual notation. Similarly to the aforementioned tools, our tool also supports generation of SQL and Java code from an EER model. Additionally, our tool allows generation of the procedural code for implementation of the inverse referential constraints. Currently, only a generation of PL/SQL code is provided.

There are numerous Computer Aided Software Engineering (CASE) tools to support EER approach, such as PowerDesigner [20], ERWin [5], SmartDraw [19], Oracle Designer [18], or Cameo Data Modeler [6] for MagicDraw. These are mainly commercial and widely used CASE tools and as such they provide proprietary graphical notation for EER usually supporting only selected concepts. In contrast to aforementioned CASE tools, EERDSL provides all of the theoretical EER data modeling concepts. Our tool also supports data modeling using the textual notation. EERDSL is the component of the MIST tool that also provides modeling using the FT concepts. MIST is the only tool that supports the usage of the FT concepts.

VI. Conclusion

During our previous research we developed FT components for our Multi-Paradigm Information System Modeling Tool (MIST). These components provide specification of an IS database schema, business applications and their graphical user

```
...
CREATE TABLE Takes (
  grade int NOT NULL ,
  examDate date NOT NULL ,
  studentID int NOT NULL ,
  subjectID int NOT NULL ,
  teacherID int NOT NULL ,
  subjectID_Grades int NOT NULL
  ,
  CONSTRAINT PK_TAKES PRIMARY KEY (studentID, subjectID)
);

ALTER TABLE Takes ADD (
  CONSTRAINT FK_TAKES_STUDENT FOREIGN KEY (studentID)
    REFERENCES Student (studentID),
  CONSTRAINT FK_TAKES_SUBJECT FOREIGN KEY (subjectID)
    REFERENCES Subject (subjectID),
  CONSTRAINT FK_TAKES_TEACHESCLASSES
    FOREIGN KEY (teacherID, subjectID_Grades)
    REFERENCES TeachesClasses (teacherID, subjectID)
  );

CREATE OR REPLACE VIEW
  View_Subject_TeachesClasses AS ...

CREATE OR REPLACE TRIGGER
  TRG_IRI_Subject_TeachesClasses_View
  INSTEAD OF INSERT ON
    View_Subject_TeachesClasses FOR EACH ROW
DECLARE ... BEGIN ... END;
...
```

```
package facultysystem;

public class Takes {
    protected Student student;
    protected Subject subject;
    private int grade;
    private java.util.Date examDate;
    protected java.util.ArrayList<TeachesClasses> teachesclasses;


    public Takes(Student student, Subject subject,
            int grade, java.util.Date examDate,
            java.util.ArrayList<TeachesClasses> teachesclasses) {
        this.student = student;
        this.subject = subject;
        this.grade = grade;
        this.examDate = examDate;
        this.teachesclasses = teachesclasses;
    };

    public int getGrade() {
        return this.grade;
    };

    public void setGrade(int grade) {
        this.grade = grade;
    };

    //omitted rest of the get/set methods
}
```

Fig. 9. Generated SQL and Java code

interface elements. However, designers widely use EER approach for database schema modeling. Therefore, we provided MIST components supporting EER approach, to offer designers a choice of two alternative conceptual level approaches to create IS specifications at the platform independent level. As both approaches allow a generation of relational database model from a conceptual specification, it is also possible to provide transformations between the two specifications via relational data model. Currently, we have developed a transformation from a relational to the FT model. It allows a designer to create a model using the EER approach, and then use the FT approach to enrich the specification with details of business applications and their GUI elements.

The MIST tool prototype is ready to be used in database and MDSD courses at our faculty. This should provide us with the practical experience and user feedback, allowing further improvement of the tool and new lessons to be learned.

In addition to the conceptual level meta-model and the transformation to the relational data model presented in the paper, we have developed several other model-to-model transformations and code generators. Our SQL Generator component provides generating SQL scripts and procedural code for inverse referential integrity constraints, from a relational model. Also, starting from an EER model, a class model of a database may be created and Java classes are generated. In this paper, our intention was not to give all the details about developed meta-models and transformations. Instead, we tried to focus just on those meta-model details that are necessary to recognize a general picture of the components supporting EER approach.

As components that support EER approach are public to a user and well documented, they may also be used in the educational purposes. It may be used in database courses to assist students in better understanding basic concepts of EER and the rules of EER to relational model transformations. A possible usage is in courses on domain specific languages and model driven software development, as students may familiarize themselves with new concepts using well-known EER concepts.

Several future research directions are possible, including a specification of MIST meta-models semantics and new features of the MIST tool. In order to formally specify semantics of our meta-models, one of the approaches presented in [4] should be used. This could allow us to fully automate the construction of tools supporting our language. Next, in order to fully support simultaneous conceptual specifications with EER and FT approaches, several further research directions are possible. One of them may include an implementation of EER2FT and FT2EER transformations that would allow transformations from one to another conceptual level specification. Also, another research direction would be to extend EERDSL with new concepts allowing more detailed specifications of data models. These new concepts should provide new constraint specifications at the conceptual level. For example, formal specification of database check constraints at the level of EER model is in many approaches poorly supported, or not supported, at all. As we already provide a conceptual specification of the check constraint at the level of FT models, we plan to create the appropriate formalisms for its specification at the level of EER model, too.

REFERENCES

[1] S. Aleksic, "Methods of database schema transformations in support of the information system reengineering process," Ph.D. dissertation, University of Novi Sad, 2013.

[2] S. Aleksic, I. Lukovic, P. Mogin, and M. Govedarica, "A generator of SQL schema specifications," *Computer Science and Information Systems*, 2007. [Online]. Available: http://dx.doi.org/10.2298/CSIS0702081A

[3] S. Aleksic, S. Ristic, I. Lukovic, and M. Celikovic, "A design specification and a server implementation of the inverse referential integrity constraints," *Computer Science and Information Systems*, 2013. [Online]. Available: http://dx.doi.org/10.2298/CSIS111102003A

[4] B. R. Bryant, J. Gray, M. Mernik, P. J. Clarke, R. B. France, and G. Karsai, "Challenges and directions in formalizing the semantics of modeling languages," *Computer Science and Information Systems*, 2011. [Online]. Available: http://dx.doi.org/10.2298/CSIS110114012B

[5] "CA ERwin." [Online]. Available: http://erwin.com/

[6] "Cameo Data Modeler." [Online]. Available: http://www.nomagic.com/products/magicdraw-addons/cameo-data-modeler.html

[7] M. Celikovic, I. Lukovic, S. Aleksic, and V. Ivancevic, "A MOF based meta-model and a concrete DSL syntax of IIS*Case PIM concepts," *Computer Science and Information Systems*, 2012. [Online]. Available: http://dx.doi.org/10.2298/CSIS120203034C

[8] P. P.-S. Chen, "The entity-relationship model toward a unified view of data," *ACM Transactions on Database Systems*, 1976. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-59412-0_18

[9] E. F. Codd, "A relational model of data for large shared data banks," *Communications of the ACM*, 1970. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-59412-0_16

[10] "Eclipse Modeling Project (EMP)." [Online]. Available: http://projects.eclipse.org/projects/modeling

[11] R. Elmasri and S. B. Navathe, *Fundamentals of Database Systems*. Addison-Wesley, 2010.

[12] R. N. Fidalgo, E. Alves, S. Espana, J. Castro, and O. Pastor, "Metamodeling the enhanced entity-relationship model," *Journal of Information and Data Management*, 2013. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-34002-4_40

[13] F. Jouault, F. Allilaire, J. Bezivin, and I. Kurtev, "ATL: A model transformation tool," *Science of Computer Programming*, 2008. [Online]. Available: http://dx.doi.org/10.1016/j.scico.2007.08.002

[14] I. Lukovic, "From the synthesis algorithm to the model driven transformations in database design," in *Proceedings of 10th International Scientific Conference on Informatics (Informatics 2009), Herlany, Slovakia*, 2009.

[15] I. Lukovic, P. Mogin, J. Pavicevic, and S. Ristic, "An approach to developing complex database schemas using form types," *Software: Practice and Experience*, 2007. [Online]. Available: http://dx.doi.org/10.1002/spe.820

[16] I. Lukovic, A. Popovic, J. Mostic, and S. Ristic, "A tool for modeling form type check constraints and complex functionalities of business applications," *Computer Science and Information Systems*, 2010. [Online]. Available: http://dx.doi.org/10.2298/CSIS1002359L

[17] I. Lukovic, S. Ristic, P. Mogin, and J. Pavicevic, "Database schema integration process a methodology and aspects of its applying," *Novi Sad Journal of Mathematics*, 2006.

[18] "Oracle Designer." [Online]. Available: http://www.oracle.com/technetwork/developer-tools/designer/overview/index-082236.html

[19] "SmartDraw." [Online]. Available: http://www.smartdraw.com/

[20] "Sybase PowerDesigner." [Online]. Available: http://www.sybase.com/products/modelingdevelopment/powerdesigner

[21] B. Thalheim, *Entity-relationship modeling: foundations of database technology*. Springer, 2000.

[22] R. S. Wazlawick, *Object-oriented analysis and design for information systems*. Morgan Kaufmann, 2014.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1621–1628
DOI: 10.15439/2014F278
ACSIS, Vol. 2

Stormgen - A Domain Specific Language to create ad-hoc Storm Topologies

Siddharth Santurkar, Abhishek Arora and K Chandrasekaran
Department of Computer Science and Engineering
National Institute of Technology, Karnataka, Surathkal, India
siddharth.santurkar@ieee.org; abhishekarora185@gmail.com; kchnitk@ieee.org

***Abstract*—Large-scale distributed data processing has gained significant momentum in research in the past decade. With the introduction of MapReduce, many frameworks have been developed that either implement MapReduce or provide additional functionalities useful in a larger domain. While the framework introduced in the MapReduce paper performs batch-processing of data, Apache Storm performs real-time computation on data. Storm does this with the help of Topologies, and the constituents of the Topology are developed using General-purpose Programming Languages (GPL). A Domain-specific Language (DSL) can provide a higher level of abstraction over GPLs and model the specialized features of a particular domain in a better way. In this paper, we propose the development of Storm Topology generator (Stormgen), a DSL for Storm Topology development, and show how the specifications of this DSL can be utilized during the code generation of exact Storm Topology components in Java. The parser and code generator for Stormgen's syntax are developed using the Eclipse Modelling Framework. The practical use of Stormgen is illustrated with a case study which considers the modelling of a Topology for the Word Count application.**

***Index Terms*—Domain-specific modelling, Languages, Eclipse Modelling Framework, Apache Storm, Code generation**

I. Introduction

THE ADVENT of big data analysis created a need for making systems that handled such data in a fault tolerant and distributed manner. Google's paper [1], which introduced the MapReduce programming model, and Yahoo! [2], with its Hadoop framework, succeeded in meeting this need. Several additions and improvements to Hadoop and other such frameworks have been made, owing to the diversity in the meaning and purpose of the data being analysed. In the case of real time, event-based and unbounded data, such a task must be approached in a manner different from that advocated by the frameworks named above. Apache S4 and Apache Storm, apart from a few other frameworks, have made this possible.

MapReduce eases the development of parallel batch-processing programs that conform to the map-reduce programming paradigm. Storm [3] can be used for real-time processing of data for a wide set of data processing use cases, some of which are listed below:

1) Stream processing, which includes processing of messages, updating of databases, etc.
2) Continuous computation, where data streams can be queried continuously and results can be streamed into clients.
3) Distributed Remote Procedure Calls (RPC), where the processing of a complicated query can be distributed and parallelized.

All these features can be easily implemented using the simple, yet, powerful primitives provided by Storm.

Storm involves the creation of a Topology [4] of computation, i.e. a graph with nodes and directed edges. The directed edges between nodes provide the paths that can be used by the event stream between the said nodes. Each node performs a stream transformation, i.e. accepts an input stream and emits a modified stream. There are also a special set of nodes that are dedicated to fetching the input data stream from an external source, if not creating the streams by themselves.

Hence, the developer has to write the code for each node in the Topology, and assemble the nodes with their connecting edges to finally obtain the graph. All this needs to be done using the Storm Application Programming Interface (API) [5], which is supported by multiple programming languages (Java Virtual Machine (JVM) based and non-JVM based) such as Java, Ruby, Scala, Python, etc. Commonly, various data mining and machine learning tools are deployed on Storm nodes, which generally fit the use cases mentioned above.

The Storm framework is developed under the Eclipse Public License, and is available to open use by companies and other organizations. Git and Altassian JIRA are used for version control and issue tracking, respectively, under the Apache incubator program. Some organizations that have employed Storm are Twitter, Groupon, Alibaba, The Weather Channel and FullContact.

DSLs [6] are small, simple and highly-focused specification languages developed for a clear and small problem domain. They are tailored to a specific application domain. They offer substantial gains in expressiveness and ease of use compared to the use of GPLs in the same domain. By employing well-known concepts, abstractions and notations derived from the problem domain, they are easy to learn, understand and use, both by developers and domain experts.

Further, DSLs facilitate the use and reuse of domain knowledge. They are not constrained to be like programming languages. One of the main advantages is that they transcend the boundaries of programming. They tend to be more descriptive and verbose than GPLs. Hence, using DSLs, domain experts can directly contribute to the development effort by validating, modifying and even independently developing DSL programs.

978-83-60810-58-3/$25.00

Every time a Storm Topology needs to be created, hundreds of lines of Java code will have to be written by the developer. Anyone who needs to deploy their application on Storm would first have to learn how to develop Storm Topologies using general purpose languages. This could waste a lot development effort and time, that could instead be used for improving the application at hand.

For these reasons, any DSL for the Storm domain should create a simple, quick and reliable means for assembling a Storm Topology and deploying the intended application. Of course, the domain knowledge of Storm and its concepts is still necessary, and the application should be deployable within the scope of the DSL. Further, the creation of the Storm Topology in an ad-hoc manner using the DSL can be used for testing the application on this framework.

XText [7] is an open-source framework for developing programming languages and domain-specific languages. It provides a parser, abstract syntax tree generator and a Java code generator. It is a part of the Eclipse modelling project. We used XText to create Stormgen.

The rest of the paper is organized as follows: the related work in this area is given in Section 2. A domain specific meta-model for Topology development in Storm, which is used for developing the abstract syntax of Stormgen, is discussed in Section 3. Based on this meta-model, the textual concrete syntax of Stormgen is presented in Section 4. The transformations from the specifications required for Code generation are elaborated in Section 5. A case study is presented in Section 6, where Stormgen is used to develop a Storm Topology for the popular "Word Count" problem. Finally the conclusions and possible future improvements are discussed in Section 7.

II. Related Work

XText, as a tool to develop DSLs, is gaining large popularity in the research community, due to its simplicity, stability and facilities. In [8], XText was used to create a DSL called SEA_L which is used in Semantic Web enabled Multi-agent Systems. This paper follows a similar analysis performed in [8] to present the DSL developed.

Other instances where DSLs have been created using XText includes [9]. In this work, the XText framework was used to describe the implementation of an assembler editor for the development of assembly code. The editor supports specific assembler instructions. The other features provided by this editor are content monitoring, detection of repeated instructions, and prediction to assist user input. Our DSL also comes with such an editor. XText comes with a customizable user-interface component, which can be used to create an Eclipse-like IDE for the DSL being created. This way, all the useful Eclipse development features can be provided to our DSL.

In yet another research work [10] selected dependability of multi-agent system (MAS) as the domain. The key requirement in this domain is an efficient verification of a Topology model of a power system. As a result, they developed a DSL as a reliability evaluation solution offering a significant rise in the level of abstraction towards MAS utilized by the system. They made use of Eclipse Ecore, as it becomes a common denominator, in which both meta-models and abstract syntax trees are defined. Eclipse Ecore is a meta-modelling framework, part of the Eclipse Modelling Project, and we have made use of this to prepare our meta-models.

Where distributed fault-tolerant systems are concerned, as discussed in [11], Pig Latin is a DSL that is used to create and execute MapReduce jobs on Hadoop. The Pig Latin syntax has the declarative style of SQL and the low-level, procedural style of MapReduce. This language is especially useful for an expert in RDBMS systems and SQL to perform MapReduce jobs without requiring knowledge of the MapReduce programming paradigm and Hadoop. Hence, it lies at a very high level of abstraction. However, our DSL does not provide that level of abstraction, as a user cannot fully exploit the various features of Storm, if restricted only to an SQL-like interface. Instead, a user with knowledge of the features of Storm can develop a Topology that can best fit the problem at hand. Esper is the Pig Latin equivalent in Storm, i.e. it provides streaming of SQL queries on top of Storm.

Other DSLs for Storm [4] include Redstorm [12], Scala DSL, Clojure DSL, etc. Each of these DSLs require knowledge of programming languages like Scala, Ruby and Clojure. Our DSL aims to alleviate the use of GPLs and provides simple constructs close to plain English to develop the Topology. We have evaluated Stormgen against Redstorm at the end of this paper.

III. Abstract Syntax

The abstract syntax of a DSL describes the domain concepts and their relations without any consideration of their meaning. In terms of Model-driven development, a domain model or data model represents the data we want to work with. The data model is generally independent of application logic. The meta-model is used to describe the structure of the domain model. The abstract syntax is described by a meta-model. This constitutes the analysis phase of the development of the DSL.

As mentioned earlier, a Storm Topology consists of a collection of nodes that do some processing and transformation on the incoming data stream. Broadly, there are 2 types of nodes that can exist in a Storm Topology:

1) **Spouts**
   These are nodes that create an input stream of data for the Topology either by generating it randomly on the fly, or by connecting to a third party source of events through a streaming API (for example, the Twitter streaming API [13]). The Spout collects the events from the source and emits them to the rest of the Topology. It can never have a stream input to it from any other node. In essence, it is the source vertex of the Topology. A directed graph, however, can have multiple source vertices. Similarly, a Topology can have multiple Spouts of different types.
2) **Bolts**
   These are nodes in the Topology that do some computation or processing on the incoming data stream(s) and emit the data to the downstream operators. Bolts at

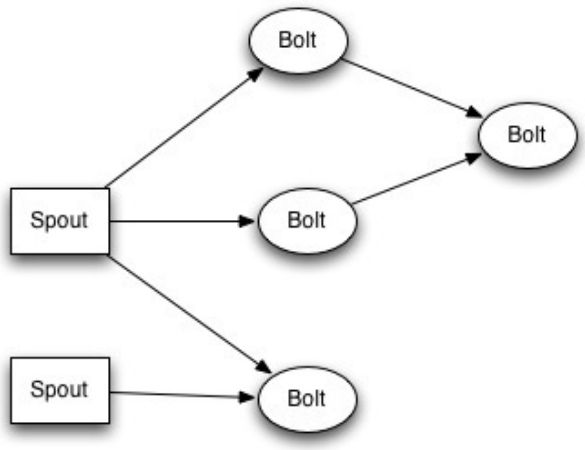


Fig. 1. Spouts and Bolts in a Storm Topology, as given in [4]

the end of the Topology do not emit data downstream. Bolts can do anything from filtering of data, execution of functions, communication with databases, database operations such as aggregations, joins, etc. They can perform either simple or complex stream processing.

A Topology can finally be assembled by adding Spouts and Bolts and the necessary edges between nodes. Fig. 1. shows a sample Topology.

Eclipse Modelling Framework (EMF) comes with two meta-models; Ecore [14] and Genmodel. The Ecore meta-model is used to store the information of the defined classes, whereas the Genmodel is used to store all the information required for code generation. EMF was chosen as the modelling platform mainly because it makes the domain model explicit, providing clear visibility. Moreover, EMF takes care of interface generation and the factory for object creation.

Ecore specifies a set of elements that could be used in the development of the meta-model. These elements are similar to some of the elements in the UML Class diagram:

1) EClass: This represents a class element. It can have zero or more attributes and references.
2) EAttribute: This represents an attribute, consisting of a name and type.
3) EReference: This represents an end-point of association between 2 EClasses.
4) EDataType: This represents the data type of an attribute. For example, *java.util.ArrayList*.

Our abstract syntax provides EClasses called Bolt, Spout and Topology. Fig. 2. shows the Ecore meta-model [14] for our model. The Storm API [5] provides interfaces to create Spouts and Bolts. For every Spout and Bolt a class has to be created and must override all the interface methods.

The aim of the final DSL is to provide very simple implementations of all the well-known characteristics of Spouts and Bolts. All the well-known features are captured in the *API* EAttribute. The Property EAttribute allows the user to define any data member (variable) and the Operation EAttribute lets the user define any function (method).

Some of the important features of the Bolt API are listed below

1) Prepare, which is used for the pre-deployment configuration of the Bolt.
2) Execute, which receives a tuple from the input stream, does some processing on the stream, and emits the result to the output collector.
3) Output Field Declarer (OPFields), which declares what logical type or key the fields in the emitted tuples assume.

Similarly, some of the important features of the Spout API are listed below

1) Open, which is used for the pre-deployment configuration of the Spout.
2) NextTuple, which connects to the source of the data stream, parses the stream into tuples, and emits them to the rest of the Topology.
3) Output Field Declarer (OPFields), which declares what logical type or key the fields in the emitted tuples assume.

Finally, the Topology assembly is done with the help of a TopologyBuilder class of the Storm API. The builder can either add Spouts or Bolts. As Bolts will have incoming edges, the adjacent upstream node needs to be specified as well. While adding a Spout to the Topology, the logical name and the instance of the class containing the source code should be provided. As any given node can be replicated multiple times during deployment, this number is provided as well while adding the Spout. While adding the Bolt, the same parameters as in the case of the Spout need to be provided. The logical name of the upstream operator needs to be defined using the "grouping" EAttribute.

IV. Textual Concrete Syntax

The textual concrete syntax of Stormgen is provided with XText [15]. XText is a language development framework to provide textual modelling languages. It can be used for creating a sophisticated Eclipse-based development environment. XText is based on Extended Backus Naur Form (EBNF) [16] rules. Hence, the design phase in the development of Stormgen constitutes the description of the EBNF rules.

As explained in the Related work section, we make use of the XText features in order to create an Eclipe-IDE user interface for Stormgen. This way, auto completion, syntax colouring, rename refactoring, bracket matching, auto edit,etc are provided for the syntax. By defining EBNF rules, the constraints discussed in the Abstract Syntax section of Stormgen's meta-model are realized. With these capabilities, the new DSL possesses both the structure and the static semantics of the Storm domain. The structure is defined by the method signatures and the semantics by the constraint code.

The rest of this section discusses the structure of the grammar used to specify Stormgen.

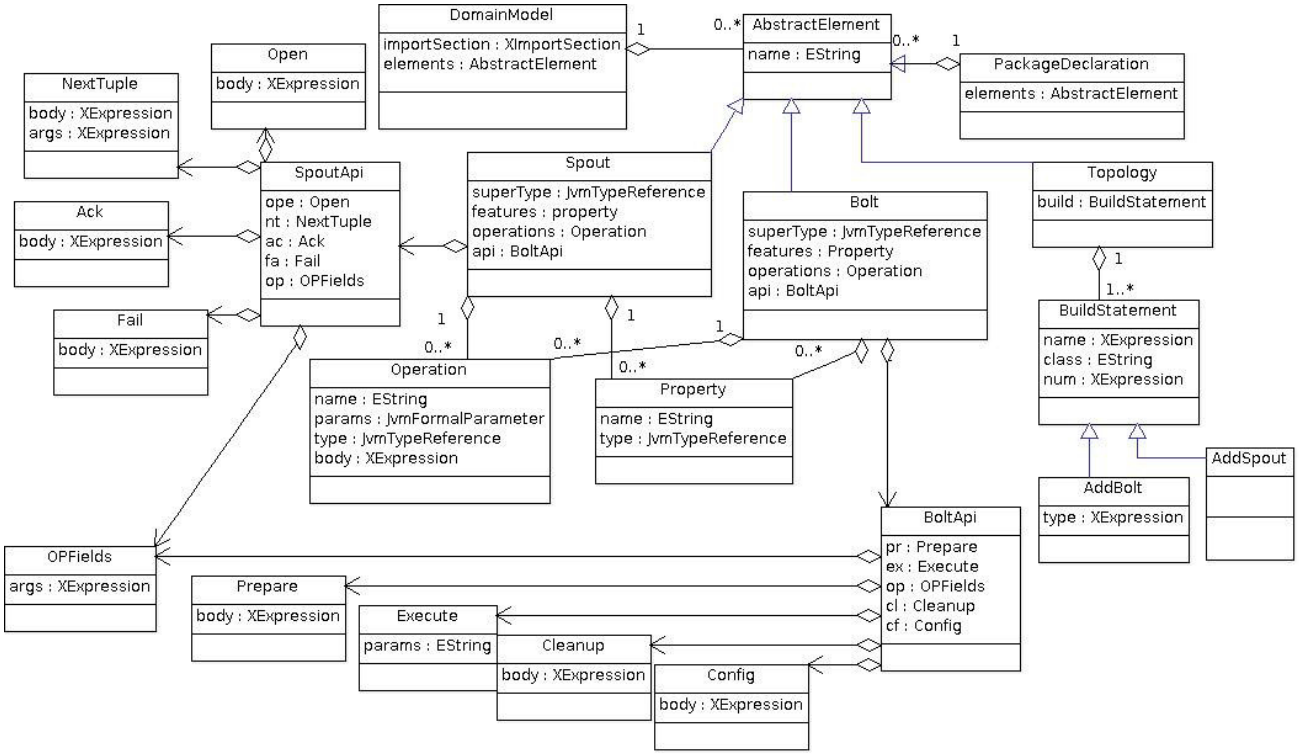


Fig. 2. Meta-model for Stormgen

Every Stormgen file should include the specifications of all the 3 major entities in the DSL, i.e. Spout, Bolt and Topology or they can contain a subset of the entire specification (For 1 or more entities).

In the beginning of the domain model all the non-Storm library related imports need to be specified. Due to the use of various JVM Elements here such as the Identifiers, Function parameters, and blocks of Java code, we made use of Xbase [17] to specify all the terminals in our grammar.

Once the imports have been specified, the rest of the file should model either a Spout, Bolt or Topology. Example 4.1 shows the implementation of this feature.

Example 4.1 (Imports and file header):

```
Domainmodel:
    importSection=XImportSection?
    elements+=AbstractElement*;
AbstractElement:
    PackageDeclaration | Bolt | Spout
    | Topology;
```

In the case of a Bolt, the meta-model constraints have to be followed. Every Bolt needs to have a well-defined name, followed by the type of Bolt primitive it is trying to implement. For example, it could implement the BaseRichBolt or IRichBolt types provided by the Storm Bolt API. The user can then optionally include basic members and methods following Java-like syntax. The rest of the code should include all of the abstract methods of the API. Example 4.2 demonstrates the Bolt syntax.

Example 4.2 (Bolt syntax):

```
Bolt:
'Bolt{'
  'Name:' name=ValidID,
  ('Type:' superType=JvmTypeReference)?,
  ('Members:' (features+=Property)+)?,
  ('Methods:' (operations+=Operation)+)?,
  ('Prepare:' pr=Prepare)?,
  'Execute:' ex=Execute,
  ('Cleanup:' cl=Cleanup)?,
  ('OutputFields:' op=OPFields)?,
  ('Config:' cf=Config)?
'}'
```

Within the Bolt API, while the implementation of the Execute feature is mandatory, the rest of the features are not. Within the Execute attribute, the emit function (outputting a tuple downstream) is called by defining the 'emit' keyword, followed by the sequence of parameters to emit.

Example 4.3 (Bolt API emit statement):

```
Execute:
'execute {'
```

```
    ('emit ('(params+=ValidID
       (',' params+=ValidID)*)')')
    '}';
```

The Spout syntax has a similar structure to the Bolt syntax, as shown in Example 4.4

Example 4.4 (Spout syntax):

```
Spout:
  'Spout{'
    'SpoutName:' name=ValidID,
    ('Type:' superType=JvmTypeReference)?,
    ('Members:' (features+=Property)+)?,
    ('Methods:' (operations+=Operation)+)?,
    ('Open:' ope=Open)?,
    'NextTuple:' nt=NextTuple,
    ('Ack:' ac=Ack)?,
    ('OutputFields:' op=OPFields)?,
    ('Fail:' fa=Fail)?,
  '}'
```

Like the Bolt, the Spout needs to be provided by name and type. The Spout could implement IRichSpout or BaseRichSpout types. The Spout can also have optional Properties and Features. Within the Spout API, the NextTuple feature needs to be included, but the rest are optional. The emit property is defined in the NextTuple function.

Finally, the grammar for building the Topology is given in Example 4.5.

Example 4.5 (Topology Syntax):

```
Topology :
    'Topology' name=ValidID '{'
     build+= BuildStatement*
    '}'
    ;
BuildStatement:
    AddSpout | AddBolt;

AddSpout:
   'addSpout{
    'SpoutName:name=XStringLiteral',
    'SpoutInstance:clas=QualifiedName',
    'Parallelism:num=XNumberLiteral
    '}';

AddBolt:
    'addBolt{
      'BoltName:name=XStringLiteral',
     'BoltInstance:clas=QualifiedName',
     'Parallelism:num=XNumberLiteral
     'Upstream:(Name:up=XStringLiteral,
      Type:(type = 'shuffle'|
      type= 'fields')+
     '}'
Config : ('default'|'custom')
```

The Topology grammar lets the user add any number of Bolts and Spouts to the Topology. Apart from specifying the primitive arguments of both Bolt and Spout, the logical name of the adjacent upstream node to every Bolt is provided. Additional configuration information needs to be applied to the Topology, such as the number of Worker processes per node, code concerning the submission of the Topology to the Storm daemon processes, etc. Using the default option with the 'Config' field (as shown in Example 4.5), the default configuration is generated for ad-hoc Topologies.

XText can generate EBNF rules from a given meta-model but we prefer to define EBNF rules manually to supply some preferred syntactical restrictions and constraints such as defining relations in a specific order (XText cannot extract the order from the meta-model because the meta-model does not have such an attribute by itself), defining at least one or more than one relation, etc.

V. Code Generation

It is not sufficient to complete the DSL definition only by specifying the notations and their representations. The complete definition requires that the semantics of the DSL's concepts are mapped to Java constructs. The mapping is provided through model to code transformations where the final executable software code for exact Storm Topology creation is obtained. Code generation for the instance models are provided by the Xtext Framework [15].

Many of the existing model driven engineering approaches accomplish code generation by writing strings to text files. XTend is a flexible and expressive dialect of Java, which compiles into readable Java compatible source code. XTend prepares a compiled output of Java source code that is similar to the equivalent hand-written code, both in structure and performance. Unlike other JVM languages XTend has no interoperability issues with Java.

Like XPand [19], XTend [18] is a template engine, which allows creating textual output using EMF models. XTend requires an EMF meta-model and one or more templates to translate the model into text. Once the requirements are provided and an EMF model [20] is defined, the code generator can be deployed. XTend traverses the abstract tree created by XText and generates the code along the way.

However, compared to XPand, XTend has the following additional benefits as explained in [21]

1) XTend is fast because XTend code is translated to Java code without adding overheads or dependencies at runtime.
2) XTend is debuggable as XTend code is translated to Java code. Hence, advanced Java debugging tools can be used. Additionally the Eclipse debugger provides the option to debug either the XTend source code or the generated Java code.
3) Better Integrated Development Environment (IDE) support.
4) As templates in XTend are expressions which yield some value, multiple templates can be composed and

the results can be passed around and processed.
5) Better extendibility.

Hence, the code generation for Stormgen is done using XTend.

Every EClass in the meta-model has a corresponding definition in the grammar. The grammar rules have to then be mapped during code generation into various components of the target generated program. As mentioned in the previous section, the 3 main components that have to be included in the Domain model are Spout, Bolt and Topology. The code generator has an "inferrer" defined for each of these 3 components.

The IRichBolt interface is used to implement the Bolt in the model. So during generation, the corresponding statement needs to be included, along with the required import. This is followed by the generation of the constructors.

Example 5.1 (Generating constructors):

```
members += element.toConstructor[
 for (feature : element.features)
 {
   parameters += toParameter
   (feature.name, feature.type)
 }
 body = [
   for(feature:element.features)
   {
     append ("this."+feature.name+
     "="+feature.name+";")
   }
  ]
]
```

Properties have to be translated into corresponding Java class member variables and Operations have to be translated to corresponding Java class functions

Finally the Bolt API attributes are translated into Java functions, which override the functions in the IRichBolt interface. All the attributes have to be appended with correct function call and arguments, having imported the corresponding arguments from the Storm library.

The Spout code is generated in a similar manner to that of the Bolt. The IRichSpout interface is implemented to provide all the necessary functions to override. The constructor of the Spout class is generated. All the instances of the Property EClass are translated into Java class members and all the instances of the Operation EClass are translated into Java class functions. Finally all the overridden methods are defined with the required prototype and the fully qualified arguments.

Topology has different grammar constructs from Bolt and Spout. Hence the code generation is slightly different. In order to build a Topology, a TopologyBuilder instance from the Storm API needs to be created. The Topology Builder instance can then be used to add Spouts and Bolts, constituting the assembly process. The adjacent upstream nodes for the Bolts are also specified. This builder instance is implicitly created, and the DSL provides EClasses like AddSpout and AddBolt to add the Spouts and Bolts to this implicit instance.

Over and above this, there is a lot of code that is used to configure the Topology for a local mode execution. All this is hard-coded into the generator, and generated for every possible Topology created using this DSL.

VI. Case Study: Word count

Word count [22] is a very popular problem that is especially used to understand the functioning of various distributed, fault-tolerant data processing systems, including Storm.

In word count, a very large document is provided as input to the framework and the expected output is a report of the frequency of every single word in the document. This is a fairly simple application to develop, and we will be using this example to underline the simplicity and power of Stormgen.

First we need to analyse the problem and understand how it can be modelled into the Storm domain. In essence, the Storm domain involves the creation of Topologies using Spouts and Bolts. So we now need to understand what components are required to develop a solution to this problem.

We need a real time data stream for processing. So we decided to simulate the same by constructing a RandomSentenceGenerator Spout which has a list of sentences, and randomly emits a sentence every time the 'NextTuple' routine is called by the framework.

Now that our data stream consists of Tuples, each being a randomly selected sentence, these sentences need to be split (removal of whitespaces) into words. To carry out this operation we construct a SplitSentence Bolt which accepts an input tuple containing a sentence, splits the sentence into words and emits each word downstream. This operation of splitting the sentence is written directly in Java.

The SplitSentence Bolt subscribes directly to the RandomSentenceGenerator Spout. As the process of splitting sentences doesn't need to be done at any specific instance of the SplitSentence Bolt, normal shuffle grouping is used to group the tuples to the instances of this Bolt.

Now that the SplitSentenceBolt generates a stream of words, these words need to be counted. For this, we construct the WordCounter Bolt, which accepts a word and increments the word's count, which is stored in a local Hash Map. This Bolt is the final Bolt in the Topology and does not emit any data subsequently. This Bolt subscribes to the SplitSentence Bolt. However, there should be a constraint enforced here. If different instances of the *same* word go to different instances of the WordCounter Bolt, then obviously the total count reported by each Bolt will be wrong. So it is important to ensure that *all* instances of the same word should go to the same instance of the Bolt to ensure a correct count. This can be achieved by using the Field stream grouping in Storm.

For convenience, in the DSL for our domain model, we have created 4 files:

1) RandomSentenceSpout.strgen
2) SplitSentenceBolt.strgen
3) WordCounterBolt.strgen

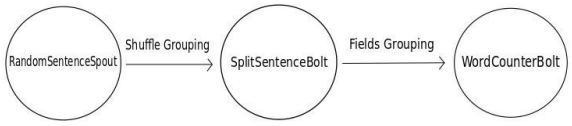


Fig. 3. The word count Topology

4) WordCountTopology.strgen

Note that all the entities could also be defined in a single Stromgen file.

Fig. 3. shows how the Topology looks like. In the WordCountTopology.strgen, we construct the Topology with the given Spouts and Bolts. The contents of the WordCountTopology.strgen file are shown in Example 6.1. The generated Java code for the Topology is given in Example 6.2. The examples of the other 3 code files are not included due to space constraints. Only if *all 4* of the source files are specified correctly will the correct Java application code will be generated.

Example 6.1 (Word count Topology Stormgen model):

```
package org.Topologies {
  Topology WordCountTopology {
    addSpout{
      SpoutName:"Spout",
      SpoutInstance:
        Spouts.RandomSentenceSpout,
      Parallelism:10
    }

    addBolt{
      BoltName: "split",
      BoltInstance:
        Bolts.SplitSentenceBolt,
      Parallelism: 5,
      Upstream: ("Spout",shuffle)
    }

    addBolt{
      BoltName: "count",
      BoltInstance:
        Bolts.WordCounterBolt,
      Parallelism: 5,
      Upstream: ("split",fields)
    }

    Config: default
}
```

Example 6.2 (Word count Topology generated java code):

```
package org.Topologies;

@SuppressWarnings("all")
public class WordCountTopology {
 public static void main(String[] args){
  backtype.Storm.Topology.TopologyBuilder
   builder = new backtype.Storm.Topology.
       TopologyBuilder();

  Config conf = new Config();
  conf.setDebug(true);
  builder.setSpout("Spout",
       new Spouts.RandomSentenceSpout(),
       10);
  builder.setBolt("split",
       new Bolts.SplitSentenceBolt(),
       5).
       shuffleGrouping("Spout");
  builder.setBolt("count",
       new Bolts.WordCounterBolt(),
       5).
       fieldsGrouping("split");

  if (args != null && args.length > 0){
    conf.setNumWorkers(3);
    StormSubmitter.submitTopology(args[0],
      conf, builder.createTopology());
  }
  else {
    conf.setMaxTaskParallelism(3);
    LocalCluster cluster =
       new LocalCluster();
    cluster.submitTopology("word-count",
       conf, builder.createTopology());
    Thread.sleep(10000);
    cluster.shutdown();
  }
 }
}
```

As it is evident from this case study, Stormgen greatly simplifies the assembly of a Topology. With the simple, verbose syntax of Stormgen we could generate the equivalent Java code for the whole Topology. Further, Stormgen is abstracted to capture just the domain concepts. Hence, only the Storm-API related specifications need to be specified using the DSL. The rest of the Java code can be generated once this specification is provided.

VII. Evaluation

As mentioned in the Related Work, there are several DSLs available for Storm, such as Redstorm, Scala DSL, Clojure DSL, etc. They provide a similar level of abstraction as Stormgen. However, the user is expected to know Scala, Ruby or Clojure, which are GPLs, but provide a simpler syntax than Java. Stormgen provides a simple, intuitive syntax for the development of the core components of the Topology. This syntax is not based on any GPL. Additionally, Stormgen permits the programmer to introduce Java code into the DSL whenever any specific functionality has to be incorporated.

The primary use case that Stormgen is developed for, is the deployment of third-party applications on Storm. As men-

tioned earlier, Storm is normally used for larger data mining or machine learning applications that need real-time, fault-tolerant and distributed data processing. Further, additional development effort has to be expended on understanding how to develop Storm topologies using Java. Instead, once the user has learnt the domain concepts, Stormgen can be directly used to create the necessary components of the topology and payload the various units of the third-party application on these components.

Redstorm [12] is a DSL for Storm developed using JRuby. All the components have to be written in Ruby. While JRuby does permit the direct use of Java code, it requires special configuration when access to non-bundled Java libraries is required. Non-bundled Java libraries will be required when a third-party application is being deployed on Storm. Stormgen solves this problem by permitting ordinary Java imports to be mentioned in the DSL file, which is added in verbatim to the generated code. This alleviates the need for any special configuration and the user can directly focus on the Topology development.

VIII. Conclusion and Future Work

In this paper we have presented our DSL, Stormgen, describing the motivation behind developing it, and further explaining in detail what went on in the whole software development process. In the analysis phase, the Abstract syntax was specified with the help of domain-specific meta-models. These meta-model concepts were then mapped to the concrete textual Syntax during the design phase. The next phase involved mapping the textual syntax to code by the construction of the code generator. Finally, we tested Stormgen with the popular WordCount application. All this was done entirely using tools such as Ecore, XText and Xtend, provided by EMF.

As explained throughout the paper, Stormgen allows the user to develop Topologies for Storm by simply applying domain knowledge and concepts to the domain model. Stormgen also allows the user to import external Java code, so that any other Java applications can be deployed seamlessly into Storm.

In future, Stormgen can be upgraded to incorporate support for additional Bolt/Spout interfaces apart from the BaseRich and IRich Bolt/Spout. While the existing implementation is sufficient to cover most use cases, providing the support for the other interfaces would cover each and every feature provided by Storm. Finally, using the Eclipse Graphical Modelling Framework (GMF) [23], a graphical DSL can be developed to supplement the textual DSL. The graphical DSL would provide a simple graphical user interface (GUI) to draw the graph of the Topology and a convenient technique to configure each node and stream (edge). This would provide a more intuitive visualization to the Storm Topology development.

References

[1] Dean, Jeffrey, and Sanjay Ghemawat. “MapReduce: simplified data processing on large clusters”, *Communications of the ACM* 51.1 (2008): 107-113. http://dx.doi.org/10.1145/1327452.1327492

[2] White, Tom. “Hadoop: The Definitive Guide: The Definitive Guide.” O’Reilly Media, 2009.

[3] Marz, Nathan. "Storm-distributed and fault-tolerant realtime computation." *Open Source Conference (OSCON)*. 2012.

[4] Marz, Nathan. "Storm wiki." https://github.com/nathanmarz/Storm/wiki (2012).

[5] Marz, Nathan. "Storm Javadoc." http://nathanmarz.github.io/storm/doc-0.8.1 (2012).

[6] Fowler, Martin. Domain-specific languages. Pearson Education, 2010.

[7] Eysholdt, Moritz, and Heiko Behrens. "Xtext: implement your language faster than the quick and dirty way." Proceedings of the ACM international conference companion on Object oriented programming systems languages and applications companion. ACM, 2010. http://dx.doi.org/10.1145/1869542.1869625

[8] Demirkol, Sebla, et al. "A DSL for the development of software agents working within a semantic web environment." Computer Science and Information Systems 10.4 (2013): 1525-1556. doi:10.2298/CSIS121105044D

[9] Kartalija, Sasa, et al. "One solution of implementation assembler editor on the Java platform using the XText framework." Telecommunications Forum (TELFOR), 2012 20th. IEEE, 2012. http://dx.doi.org/10.1109/TELFOR.2012.6419547

[10] Kowalski, Marcin, and Kazimierz Wilkosz. "A Domain Specific Language in Dependability Analysis." Dependability of Computer Systems, 2009. DepCos-RELCOMEX’09. Fourth International Conference on. IEEE, 2009. http://dx.doi.org/10.1109/DepCoS-RELCOMEX.2009.14

[11] Gates, Alan F., et al. "Building a high-level dataflow system on top of MapReduce: the Pig experience." Proceedings of the VLDB Endowment 2.2 (2009): 1414-1425. http://dx.doi.org/10.14778/1687553.1687568

[12] Superenant, Colin. "Red Storm" https://github.com/colinsurprenant/redstorm (2012).

[13] Benhardus, James, and Jugal Kalita. "Streaming trend detection in twitter." International Journal of Web Based Communities 9.1 (2013): 122-139. doi:10.1504/IJWBC.2013.051298

[14] Stephan, Matthew, and Michal Antkiewicz. "Ecore. fmp: A tool for editing and instantiating class models as feature models." University of Waterloo, Tech. Rep 8 (2008).

[15] Behrens, Heiko, et al. "XText user guide." DostupnÃľ z WWW: http://www.eclipse.org/Xtext/documentation/1_0_1/xtext.pdf (2008).

[16] Garshol, Lars Marius. "BNF and EBNF: What are they and how do they work." acedida pela ÃŽltima vez em 16 (2005).

[17] Efftinge, Sven, et al. "Xbase: implementing domain-specific languages for Java." ACM SIGPLAN Notices. Vol. 48. No. 3. ACM, 2012. http://dx.doi.org/10.1145/2480361.2371419

[18] Bettini, Lorenzo. Implementing Domain-Specific Languages with XText and Xtend. Packt Publishing Ltd, 2013.

[19] Klatt, Benjamin. "Xpand: A closer look at the model2text transformation language." Language 10.16 (2007): 2008.

[20] Budinsky, Frank, ed. Eclipse modeling framework: a developer’s guide. Addison-Wesley Professional, 2004.

[21] Five good reasons to port your code generator to Xtend - http://blog.efftinge.de/2013/06/five-good-reasons-to-port-your-code.html

[22] Dean, Jeffrey, and Sanjay Ghemawat. "Distributed programming with Mapreduce." Beautiful Code. Sebastopol: O’Reilly Media, Inc 384 (2007).

[23] Eclipse Consortium. "Eclipse Graphical Modeling Framework (GMF)(2007)."

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1629–1637
DOI: 10.15439/2014F255
ACSIS, Vol. 2

Study of Interoperability between Meta-Modeling Tools

Heiko Kern
University of Leipzig
Augustusplatz 10
04109 Leipzig, Germany
Email: kern@informatik.uni-leipzig.de

***Abstract*—Modeling is a fundamental concept in software engineering and other system development disciplines. Nowadays the modeling process is supported by powerful modeling tools. Generally speaking, tools which support the definition and usage of self-defined languages are called meta-modeling tools. An important requirement for meta-modeling tools is the interoperability among each other. For instance, interoperability helps to build complex tool chains covering the whole development process. Furthermore, interoperability can also avoid the vendor lock-in effect. Thus, interoperability facilitates the replacement of a tool by a new tool better fitting the customer needs. The objective of this paper is to investigate the current status of interoperability between meta-modeling tools. In more detail, we study the degree of model exchange between meta-modeling tools and look for typical exchange approaches. The study focuses on meta-modeling tools and approaches which are being used in practice or the real world, respectively.**

I. Introduction

MODELING is a fundamental concept in software engineering and other disciplines. A model represents a system in an abstract way. The abstraction helps to improve the understanding of a system and can facilitate the communication between different stakeholders. Beyond that, in modern development approaches (e.g. Model-Driven Software Development (MDSD) [22] or Domain-Specific Modeling [11]) models are increasingly used for automating development tasks such as code generation, model transformation or model-based testing.

Beside a theoretical foundation of modeling, a suitable tool infrastructure is necessary to enable the practical usage of MDSD approaches. Current modeling tools offer a variety of features which support the user during the modeling process. Modeling tools supporting the definition as well as the usage of self-defined languages are called meta-modeling tools. The modeling languages in these tools are generally defined by meta-models. Examples of such meta-modeling tools are MetaEdit+ [11], Generic Modeling Environment [15] or Microsoft Visio [4].

An important requirement for meta-modeling tools is the interoperability among each other. Interoperability is the ability of two or more tools to work together. For instance, often a tool is dedicated to a specific task. Tools have to work together or inter-operate to build complex tool chains covering the whole development process. Another issue is the evolution of a tool landscape. Interoperability can avoid the vendor lock-in effect. Thus, interoperability facilitates the replacement of a tool by a new tool better fitting the customer needs.

The objective of this paper is to investigate the current status of interoperability between meta-modeling tools. Although there are a variety of approaches, the current state of practice in the area of modeling is unclear. In more detail, we want to study the degree of model exchange between meta-modeling tools and look for typical exchange approaches. The study focus on meta-modeling tools and approaches which are used in practice or the real world, respectively. We mainly consider the import and export features of meta-modeling tools in order to exchange models and meta-models. The objective can be founded with the following two research questions:

– *Question 1: What is the degree of interoperability?* The first research question investigates the degree of interoperability. We want to analyze between how many of the involved tools an exchange of models is possible? Based on our experience, we assume that the model exchange between different meta-modeling tools is insufficient. This study will prove this assumption.

– *Question 2: What are the approaches to realize interoperability?* In order to give a satisfying answer to this question, research of approaches is necessary. There are already a variety of approaches in theory and literature. However, in this study we want to identify approaches used in practice.

The paper is structured as follows. In the subsequent section, we give a foundation of the interoperability concept. In section III we present a set of aspects which helps to scope the investigation. Afterward in section IV, we describe the methodology for the tool selection and analysis of these tools. In section V we present the results of the study and discuss the validity of these results. Finally, we conclude in section VI.

II. Interoperability

Interoperability is in research and in practice a subject of discussion since there are software systems. The word consists of two parts: “inter-operate” and “ability”. Inter-operate means that two systems can work together [6] and the suffix ability expresses “the ability of a system [...] to work with or use the parts or equipment of another system” [1]. A basis for interoperability is the capability to exchange information between two or more systems and to use the information that has been exchanged [2]. Furthermore, interoperability is

978-83-60810-58-3/$25.00

the basis to integrate systems. Integration is also an widely-used term in software and system development and can be defined as the combination and coordination of separate things, elements or units into a whole, so that they work together effectively [1], [3]. Regarding the concept of interoperability, integrated systems must be interoperable in any form, but interoperable systems do not need to be integrated. Interoperability extends the borders of already existing systems and enables the connection to other systems. Here, interoperability is often associated with loosely-coupled systems, where systems keep their autonomy [17]. In contrast to this, integration is characterized by a closely-coupled systems, where system are interdependent and difficult to separate from each other.

Another term in this context is migration. Generally, migration denotes processes of spatial movement. In information technology there are different application areas for migration, such as software systems, databases, application systems or hardware. A migration in the area of software systems is, for instance, updating from one major software release to the next highest version of the same software vendor. Already existing data, settings or specific extensions have to be transferred to the new software system.

The focus of this article is the interoperability of different meta-modeling tools. In this context, interoperability deals with the exchange of models and meta-models between meta-modeling tools. The exchange is realized as a migration of models and meta-models from one tool to another. The migration from one to another tool should be an isomorphic relation in order to preserve the structure and semantics of models. The terms integration, interoperability and migration can be used as synonyms in this paper.

III. Scope of Study

There are a variety of problems and solutions concerning the interoperability issue. This shows, for instance, the annotated bibliography from Wicks [26] which contains a huge amount of papers about the interoperability issue. In this section, we define a set of aspects or dimensions which helps to scope the research (questions) of this study. The finding and selection of these aspects are based on a theoretical study of different approaches and problems through literature analysis and the authors's knowledge. The dimensions and their properties are not eligible for completeness but fit the objective of this paper.

A. Unification Mechanism

One of the main reason for missing interoperability is heterogeneity between artifacts that have to be exchanged (e.g. models and meta-models). This heterogeneity between the models can be, for instance, of syntactic or semantic nature. To achieve interoperability between the participating models it is necessary to overcome this heterogeneity and to find a unification between different structures. We can distinguish between the following two fundamental unification mechanisms.

1) Common Structure: A mechanism to realize interoperability is to avoid heterogeneity a prior by defining a common structure. The definition can be regarded as a development process for a standard. Such a standard defines, for instance, a common structure of models and meta-models, their semantics and a specification for the exchange of models. If all systems conform to a selected standard, interoperability is guaranteed by this standard. Standards can address different aspects of the exchange of models and languages. In the domain of modeling, there are standards which define a whole language (syntax, semantic and pragmatics). One example is the Unified Modeling Language (UML) [7]. Additional to this, there are standards which define a whole meta-modeling environment (e.g. Meta Object Facility (MOF) [20] or Eclipse Modeling Framework (EMF) [23]) and a corresponding exchange format (e.g. XML Metadata Interchange (XMI) [21]). Meta-modeling and modeling tools which are implementing MOF, UML and XMI as serialization syntax can exchange models and meta-models without problems (theoretically).

2) Transformation: Another mechanism is the transformation of different models and meta-models. A transformation defines a mapping between different structures in order to overcome heterogeneity. Similar to standards, transformation can address semantic or syntactic issues. If there is no standard in order to exchange data between tools, transformations are a powerful approach to exchange models or meta-models. The mechanism of a common structure and transformations are not mutually exclusive. A proprietary meta-modeling environment can implement a standard by using transformations in order to create a model and meta-models conforming to this standard. But this is only possible up to a certain degree.

B. Modeling Level

A meta-modeling tool consists of a modeling and a language level. On the language level (also called as meta-modeling level) a language engineer can define different modeling languages by the creation of meta-models. These modeling languages can be used by a modeler at the modeling level to create models. Based on these two levels, we differentiate between the following two cases of model exchange.

1) Model Level: The exchange on the model level includes only the models themselves. The exchange of languages is excluded on this level.

2) Language Level: Additionally to the exchange on the model level, it is possible and necessary to exchange languages between meta-modeling tools. The exchange of artifacts on language level includes generally the exchange on model level. A sub-aspect on this level is the language preservation. Generally the source and target models and language should be isomorphic. Language preservation can relate to the meta-model and the concrete syntax of language.

C. Topology

Meta-modeling tools exchange their data by using different topologies. The concept of topology originally stems from the field of computer networks and is also used in software

and system integration (e.g. Enterprise Application Integration [16]). The topology concept can be transferred to the area of meta-modeling tool integration. We can distinguish the following topologies.

1) Point-to-Point: A simple topology is a point-to-point connection which connects two tools directly to each other. In this case a tool exports their models and meta-models via a file. The target tool imports this file and reads the models and meta-models. If an external transformation is part of this point-to-point connection, than we have an indirect point-to-point connection. Otherwise, we talk about a direct point-to-point connection.

2) Complex Topologies: If there are more than two meta-modeling tools involved in the integration, a point-to-point integration can be insufficient. For that reason, there are more complex integration topologies such as star or bus. A star topology is characterized by the fact that there is one common exchange format or interface to exchange models and meta-models between all participating tools. The realization of a star solution often requires an additional integration component, which is in the center of the integration architecture. This component controls the integration process and serves as a common structure for models and meta-models. Realizations of more complex structure are, for instance, Model Bus [9] or BPM-X-Converter [1].

D. Integration Layer

The integration of software requires access to artifacts that should be exchanged between the software systems. Generally there are different layers to access these artifacts. Integration layers describe on which layer the exchange of artifacts is realized. Based on the integration of software development tools [25], we can differentiate between the following typical integration layers.

1) Data: Models and meta-models can be represented as data in files or databases. Hence, the integration can be realized on the data layer. Many tools enable the export and import of models and/or meta-models as files. In this case, no complex infrastructure is necessary for building an integration solution. The disadvantage of this approach is that the serialization of models and meta-models as data are often complex. The processing of these complex data structures implies a complex solution (transformations and/or exchange formats).

2) Function: Above the data layer, many tools provide an API with different functions. The usage of functions are easier than operating directly on the data layer because complex operation are encapsulated. Typical functions are the selection and creation of model elements. There are integration approaches which uses the function layer instead of the data layer.

3) Presentation: The third layer is the presentation layer. The exchange on presentation layer often only considers the graphical representation of models. For instance, the export as image considers only the graphical representation. In doing so, the import is unsatisfying because the access of single model elements is impossible. A further example of exchange at presentation layer is the approach of Object Linking and Embedding (OLE).

[1] http://www.bpm-x.com/

TABLE I
SCOPE OF THE STUDY (GRAY =FOCUS)

Unification Mechanism	Common Structure		Transformation
Modeling Level	Model Level		Language Level
Topology	Point-to-Point		Complex Topology
Integration Layer	Data	Function	Presentation

E. Research Scope

Based on the dimensions described in this section, we setup the scope of the study as follows. Table I shows an overview of the selected investigation dimensions.

- *Unification Mechanism*: The study includes the investigation of common structures (standards) and transformation approaches.
- *Modeling Level*: We investigate approaches on model and language level but the focus lies on approaches on the languages level while preserving the language structure.
- *Topology*: We study interoperability approaches that realize a direct point-to-point connection between tools.
- *Integration Layer*: The study considers all layers but we focus on data layer.

IV. Selection and Analysis of Tools

A. Tool Selection

The selection of tools is an important aspect of this study because the tools form the basis for later analysis. For the tool search we mainly use the World Wide Web. Finding meta-modeling tools is a difficult task because the term meta-modeling tool is a theoretical term that is often used in the context of the Model-Driven Engineering community. Tool vendors uses different names for their meta-modeling tools. Beside the term meta-modeling tool, we use different synonyms such *meta-case tool* [24] or *modeling tool*. Most meta-modeling tools are denoted as modeling tools with the ability to define their own language. Hence, we also include in our search the general word *modeling tool*. Based on the initial set of tools, we filter the tools by the following criteria.

- *Maturity level*: The tools must fulfill a certain maturity level. A tool must be installable and usable for later analysis. Many tool vendors provide a trial version of their tool. The most tools are available as desktop applications but there are also web-based tools.
- *Concrete syntax*: A further requirement concerns the concrete syntax of models. We only select meta-modeling tools that enable the definition of graphical modeling languages.
- *Modeling domain*: The third criterion concerns the modeling domain of tools. Generally, meta-modeling tools are tools which allow the definition of domain-specific

languages in different domains. This implies that many tools have a universal/generic character. However, many modeling tools relate to a certain modeling discipline. In this study we focus on the following domains: software development, business process modeling, and data modeling.

– *Meta-modeling capability*: The last criterion is the meta-modeling capability. We can distinguish between the heavyweight and lightweight approach [8]. The heavyweight approach enables the creation of a language through the definition of a complete new meta-model (e.g. MetaEdit+). The lightweight approach adapts an already existing meta-model (e.g. UML profile mechanism). A tool must support the heavyweight meta-modeling approach by using a three-level model hierarchy consisting of model, meta-model and meta-metamodel.

Table III in the appendix shows the modeling tools we found during our search. Each tool in this list fulfill the first three selection criteria. The third column in Table III shows the meta-modeling capability of each tool. We only include tools that support the heavyweight meta-modeling approach. The last column indicates that a tool is included in this study. Overall the study includes 20 tools which fulfill the defined criteria.

B. Tool Analysis

The analysis starts with the installation of each tool. Afterwards, we investigate the import and export functionality. Typically, there are different interface layers: user, function and data interfaces. We concentrate on the user interface and especially on the tool menus. Often tools provide a menu entry for import and export of modeling artifacts. Some tools have no extra import and export menu because they offer this functionality in the load and save menu.

In addition to the user interface analysis, available documentation is used to find out exchange possibilities. For instance, we use product information and manuals because a lot of tool vendors emphasize and describe their import and export capabilities.

Some tools provide import and export capabilities in their programming interface or provide a generator component that can be used to implement export and import scripts. These functions are excluded in the study. We only include export and import interfaces that already exist. The exchange possibilities have to be ready to use without any programming of generators or functions.

Furthermore, we restrict the analysis of the exchange possibilities of a tool. We test the exchange mechanism in order to understand the approach used, but we do not investigate the quality of the model exchange. Table IV in the appendix shows the result of this analysis. The table contains the import and export capabilities of each tool. Based on this raw data, we derive the results in the next section.

V. Results

A. Unification Mechanism

1) Common Structure: Many tools use the approach of a common structure in order to exchange their meta-models and models. But the used formats are often a proprietary definition which realize the saving and loading of models instead of the model exchange between different tools. However, some tools use the Visio format in order to exchange models and language elements. In addition to the proprietary formats, there are standards which allow only the exchange of models. These models must conform to a certain (standard) language. For instance, some tools enable the export and import of BPMN-XML [19]. But these standards do not allow the exchange of languages or meta-models. Finally, there is no common format – with the exception of the Visio format – that is used to exchange meta-models and models between different tools.

2) Transformation: The exploration of transformation approaches are difficult because most tools implement their import and export process as a black box. Hence, we can only investigate transformations which are explicit or visible during the import or export. We found some tools which support transformations during the exchange. The first approach is used in Agilian, Visual Paradigm for UML and Business Process Visual Architect. The transformation is realized by a wizard which allows to configure a mapping between Visio models and models of this tool. The mapping can only be applied to certain modeling languages. A further tool which supports a mapping during the exchange process is ARIS. This tool allows a configurable import of Visio models. The description of the configuration is realized in XML on language level. Generally the transformations are not comparable to powerful transformation approaches such as Eclipse Epsilon Transformation Language (ETL) [14] or Atlas Transformation Language (ATL) [10].

B. Modeling Level

1) Model Level: Some meta-modeling tools contain pre-defined modeling languages. Based on these pre-defined languages, some tools offer a language-specific exchange. An example is Agilian that allows the export of BPMN-XML and Business Process Visual Architect that enables the import of BPMN-XML. These specific languages are often standards in a certain domain. Regarding the unification mechanism, this approach follows the strategy of a common structure to exchange models. The limitation to a certain language is unsatisfying in the context of meta-modeling tools.

Additional to this, there is an approach allowing the generic exchange of models between tools. The approach can exchange each model as a generic graph or tree. But the interpretation of models which conform to this generic format is unclear because of the missing language definition.

For instance, the following tools support exchange on the model level:

– Agilian, Visual Paradigm for UML, Business Process Visual Architect: These tools allow the import of Visio

stencils, but the imported masters of a stencil are not part of a target language. Masters are transformed into a separated icon library. Thus, these tools allows only the import of models conforming to a certain language. Please note, a Visio stencil can be regarded as a meta-model [13].
- ARIS: This tool allows the import of Visio models. It is not possible to import stencils. ARIS offers a mapping function which enables a mapping between Visio stencil elements and certain ARIS language elements. Based on this mapping, ARIS imports Visio models.
- Edraw Max: This tool enables the import of Visio models. It is impossible to import stencil elements.
- Lucidchart: This tool allows the import and export of Visio models. The export was not testable in the free version. The import transforms only graphical elements and no stencil elements.
- *Dia:* Dia allows the import and export of Visio models without stencils.

C. Language Level

In contrast to the model level, there are tools which allow the exchange of modeling languages and models conforming to these languages. In this case, the exchange approach transforms the source language into the target language. Based on this transformation, all models conforming to the source language are transformed into models, which are conformed to the target language. Some tools provide the reverse order of these transformations, that is, the tool imports model elements and after that the tool creates the corresponding language elements. But this reverse order leads to the problem that the import only considers a certain set of language elements.

- *ConceptDraw:* ConceptDraw enables the import of Visio models. Additionally to the import of models, ConceptDraw can import language elements (Visio masters). ConceptDraw indirectly imports the master elements via the model import. ConceptDraw also allows the export of models to Visio but there is no stencil support.
- *iGrafix:* This tool allows the import of Visio models by using the clipboard and the tool also imports stencil elements which are used by the imported model elements.

D. Degree of Interoperability

Table II shows the export and import connections of the investigated tools. The vertical axis is the source tool and the horizontal axis is the target tool. The source tool exports models and the target tool imports models. For instance, there is a directed connection from Visio to Concept Draw. That is, Visio can export models and ConceptDraw can import these Visio models. We differentiate the connections depending on their modeling level. The rectangle (□) stands for an approach that supports the exchange of languages (meta-models) and models. The plus sign (+) represents an approach which supports only the exchange on model level. This includes the approach for the exchange of language specific-models and the approach for the generic exchange of models. We combine both signs (⊞), if a tool supports exchange of languages and models as well as language-specific exchange or generic exchange of models.

The diagonal in the matrix shows that each tool allows the exchange of their own languages and models because each tool can save and load their own language definitions and models. Additionally, many tools allow the exchange on the model level (language-specific or generic model exchange). There are only three connections which allow the transformation of languages and models. The matrix also shows that Visio plays a key role because a lot of tools enable the import of Visio models.

All in all, there are $20 \times 20 = 400$ directed connections in this matrix. We assume that the export and import between the same tool is a basic feature in order to save and load models and languages. For this reason, we exclude the diagonal in our calculation. Thus, we have $20 \times 20 - 20 = 380$ possible connections. Out of these connections, there are 30 connections between different tools. This leads to a ratio of 7.9%. There are 27 connections that allow the exchange on model level, a ratio of 7.1%. Regarding the exchange on the language level, there are only three connections. This is a ratio of 0.8%. Hence, the degree of interoperability can be considered low.

E. Further Observations

Generally we identified different data formats for realizing the exchange of model data. One known exchange format for models is XML Metadata Interchange (XMI) [21]. We consider XMI in our study but we are not focus on XMI because their close relationship to MOF, EMF and UML. Other meta-modeling tools do not use XMI for realizing the exchange of their models and meta-models. Another mechanism to exchange models is the usage of graph formats such as Graph Exchange Language (GXL) [27] or GraphML [5]. Graph formats are suitable for exchange models because models and meta-models can be regarded as graphs. For instance, MetaEdit+ uses an adapted version of GXL for serializing their models and meta-models, but no other tool in the study can import this graph format. yEd can import the GraphML format. A further observation is that some tools provide Excel exports. This is not for exchange reasons, but rather than a format to make reports. Beside the possibilities to exchange meta-models and models, there are a lot of language-specific formats, depending on the tool's domain. For instance, in this study many tools support typical formats from the business modeling domain such as BPMN-XML, BPEL [18] or XPDL [28].

Another observation is that there are more tools allowing import than tools allowing the export of models. This could be a strategic reason. Most tools support the import because tool vendors want to increase their usage and often it is necessary to import data in order to replace other tools. The export is undesirable because the tool vendors try to bind their customers to a certain tool.

The last observation concerns the transformation capabilities. Some tools allow the definition of mappings between

TABLE II
MODEL AND LANGUAGE EXCHANGE (□=LANGUAGE LEVEL, +=MODEL LEVEL)

Tools	Agilian	ARIS BA	Atom^3	Business Process VA	ConceptDraw	Cubetto Toolset	Dia	Edraw Max	Enterprise Architect	GME	iGrafix Process	Lucidchart	Maram Meta-Tools	MetaEdit+	Microsoft Visio	PowerDesigner	ViFlow	VP for UML	VMSDK	yED
Agilian	⊞	+		+					+							+				
ARIS BA		⊞							+											
AToM^3			□																	
Business Process VA				⊞																
ConceptDraw					□										□					
Cubetto Toolset						□														
Dia							□								+					
Edraw Max								□												
Enterprise Architect	+	+		+					⊞							+				
GME										□										
iGrafix Process				+							□									
Lucidchart												□			+					
Maram Meta-Tools													□							
MetaEdit+														□						
Microsoft Visio	+			+	□		+	+	+		□	+			□	+		+		
PowerDesigner	+								+							⊞				
ViFlow																	□			
VP for UML	+	+		+					+							+		⊞		
VMSDK																			□	
yED																				□

models or languages in a simple and limited way. The tools do not provide powerful transformation languages such as ATL or ETL.

F. Threats to Validity

The interoperability between meta-modeling tools is between 0.8 and 7.9%. If we take a look at our study restrictions, the interoperability could be higher than these measured values. We focus on a limited set of tools in a selected domain. Maybe other tools in other domains have a higher interoperability degree. Furthermore we only looked for interoperability mechanisms that are provided by the tool itself. Some meta-modeling tools provide a generator which allows to generate every exchange format. Furthermore, we exclude external tools, such as BPM-X-Converter, which allows the migration of models between meta-modeling tools.

In contrast to this, we can argue for a lower value of interoperability. We only investigated the opportunity to import and export models. We cannot say anything about the quality of these exchange mechanisms. Furthermore, some tools relate very close to Visio. Hence, the import and export is easy for these tools. This is maybe similar to MOF-implemented meta-modeling tool. If we would not consider the Visio imports on language level, the interoperability would go against zero.

VI. CONCLUSION

In this paper we presented a study about interoperability between meta-modeling tools. The study included 20 tools in the area of software development, business process modeling and data modeling. In the first part of this paper we defined the investigation scope. In the second part, we analyzed the tools and presented the results.

Regarding the first research question about the degree of interoperability, we can give the following answer. Depending of the approach considered, the degree of interoperability is low with values between 0.8% and 7.9%. The answer to the second question is more complicated because of the different approaches. Regarding the modeling level (section V-B), we identified the following three approaches: (1) the exchange of models which conform to a specific modeling language, (2) the generic exchange of each model without their modeling language, and (3) the exchange of models with their language. Besides the aspect of the modeling level, there are many other dimensions including many other approaches for interoperability.

In our future work we want to increase the degree of interoperability between meta-modeling tools. We assume a main reason for the missing interoperability is the heterogeneity between the different meta-modeling languages of the tools. Despite the heterogeneity, there are common concepts which can be mapped to each other. These similarities and differences are described in a comparative analysis between different meta-modeling languages [12]. Based on this, we can develop a transformation-based adapter approach in order to enhance the model and meta-model exchange.

APPENDIX

TABLE III
MODELING AND META-MODELING TOOLS (●=YES, −=NO)

Name	Vendor	Version	Meta-modeling approach (light/heavy)	Included in study
Agilian	Visual Paradigm	4	●/●	●
Altova UModel	Altova	2012	●/−	−
ArgoUML		0.34	●/−	−
Archi	University of Bolton	2.3	−/−	−
ARIS Business Architect	Software AG	7.1	●/●	●
ARIS Express	Software AG	2.3	−/−	−
Artisan Studio	Atego	7.4	●/−	−
Astah	Astah	6.6.3	●/−	−
AToM3	McGill University	2008	−/●	●
bflow* Toolbox		1.2.5a	−/−	−
Bizagi Process Modeler	Bizagi	2.3	−/−	−
BOUML	Bruno Pagès		−/−	−
Business Process Visual Architect	Visual Paradigm	5	−/●	●
Cadifra UML Editor	A. & F. Buehlmann	1.3.3	−/−	−
CaseComplete	Serlio Software	7.0 (2012)	−/−	−
ConceptDraw	CS Odessa	9	−/●	●
Cubetto Toolset	Semture	1.7.1	−/●	●
Database Design Tool		1.5	−/−	−
DB Wrench	Nizana Systems	2.3.0	−/−	−
dbConstructor	DBDeveloper Solutions		−/−	−
DbSchema	Wise Coders Soluations		−/−	−
Dia		0.97.2	−/●	●
Edraw Max	EdrawSoft	6.3	−/●	●
Enterprise Architect	Sparx Systems	9.3	●/●	●
ER Creator	modelCreator Software	3.0	−/−	−
ER/Studio Software Architect	Embarcadero Technologies	1.1.0	●/−	−
ER/Studio Business Architect	Embarcadero Technologies	1.7.0	−/−	−
Generic Modeling Environment	Vanderbilt University	10.8	−/●	●
Gliffy	Gliffy		−/−	−
Grapholite	Perpetuum Software	1.6.0.7	−/−	−
iGrafix Process	iGrafix	2011	−/●	●
Intalio BPMS Designer	Intalio	6.1.12	−/−	−
Lucidchart	Lucid Software		−/●	●
MagicDraw	NoMagic	17.0.2	●/−	−
Maram Meta-Tools	University of Auckland	–	−/●	●
MetaEdit+	MetaCase	5.0	−/●	●
Microsoft Visio	Microsoft	2010 (14)	−/●	●
Modelio	Modeliosoft	2.1.1	●/−	−
NClass	Balazs Tihanyi	2.04	−/−	−
Objecteering	Objecteering Software	6.1	●/−	−
objectiF	microTOOL	7.1	●/−	−
Open ModelSphere	Grandite	3.2	●/−	−
ORM Designer	Inventic		−/−	−
Poseidon for UML	Gentleware	8	−/−	−
Papyrus	Eclipse	1.12	●/−	−
PowerDesigner	Sybase	16.1	●/●	●
Process Modeler	itp commerce	5	−/−	−
RISE	RISE to Bloome Software	4.5	−/−	−
Select Architect	Select Business Solutions		●/−	−
SemTalk	Semtation	4	−/−	−
Signavio Process Editor	Signavio	6.0	−/−	−
SmartDraw	SmartDraw Software		−/−	−
Topcased		5.2	●/−	−
UML Lab	Yatta Solutions	1.4.3	●/−	−
UMLet		11.5.1	−/−	−
ViFlow	ViCon		−/●	●
Violet UML Editor		0.21.1	−/−	−
Visual Paradigm for UML	Visual Paradigm	9	●/●	●
Visual Use Case	TechnoSolutions	4.069 (2009)	−/−	−
Visualization and Modeling SDK	Microsoft	VS2012	−/●	●
WinA&D	Excel Software		−/−	−
Xcase	Resolution Software	9.1	−/−	−
yED	yWorks	3.9.2	−/●	●

TABLE IV
IMPORT AND EXPORT FORMAT OF META-MODELING TOOLS

Agilian	
Import	Rational Rose (mdl) files, Rational DNX files, BizAgi project file, specific XML, XMI (1.2, 2.1), Eclipse UML2 (XMI 2.1), Visual Paradigm project file, MS Excel file with spcific schema, Visio drawings, Visio ERD, Visio drawing/stencils into Agilian Stencil, NetBeans 6.x UML diagrams, Telelogic System Architect, Telelogic Rhapsody, PowerDesigner project file
Export	BPMN2.0-XML, specific XML, XMI (1.2, 2.1), Eclipse UML2 (XMI 2.1), Visual Paradigm project file, MS Excel file with specific schema, VPP (ZIP project archiv)
ARIS Business Architect	
Import	XML with specific schema, UML (XMI1.1)
Export	ADF (ARIS filter), XMI, XML, Visio (VDX), BPEL, ADB (ARIS database)
Business Process Visual Architect	
Import	BizAgi project file, XML, BPMN2.0-XML, XPDL2.1, Telelogic System Architect, Excel, Visio
Export	BPMN 2.0 XML, XML, BPMN2.0-XML, XPDL2.1, Excel
ConceptDraw	
Import	Visio (VDX), MS PowerPoint
Export	CDX file (XML), Visio (VDX), MS PowerPoint
Cubetto Toolset	
Import	–
Export	ETZ format
Dia	
Import	Visio models, Dia, Dxf (specific XML file), SVG, Xfig
Export	Visio models, Dia, Dxf (specific XML file), SVG, Xfig
Edraw Max	
Import	Visio
Export	
Enterprise Architect	
Import	Database Schema, specific Visio models (Communication, Activity, Class, Object, Compontent, Deplayment, Custom), Doors, XMI (UML 1.1, 1.3 or 2.x), ARCGIS, ODM (OWL/RDF), Rhapsody, Rational Software Architect (EMX/UML2)
Export	XMI 1.0 (UML1.3), XMI 1.1 (UML1.3), XMI 1.2 (UML1.4), XML 2.1 (UML2.0), MOF1.4 (XMI1.2), MOF1.3 (XMI1.1), specfic XML, Ecore, OWL/RDF, BPMN2-XML
iGrafix Process	
Import	Visio models and metamodels
Export	BPEL XML, XPDL, XML
Lucidchart	
Import	Visio models (vdx, vsd, vsdx)
Export	Visio models (vdx)
MetaEdit+	
Import	GXL-adapted (models and meta-models)
Export	GXL-adapted (models and meta-models)
Microsoft Visio	
Import	–
Export	–
PowerDesigner	
Import	Excel, ERwin, XMI, Rational Rose (MDL), SIMUL8 file, specfic Visio models
Export	UML2, XMI2.1 XML schema files
Visual Paradigm for UML	
Import	ERWin Data Modeler project files, BizAgi project file, System Architect business process diagram, XMI (1.2, 2.1), Excel, Visio, Visio ERD, Visio diagram to Stencil, Rational Rose (MDL) files, Rational DNX files, Rational Software Architect files, PowerDesigner project file, Telelogic Modeler
Export	BPEL, XPDL, JPDL, BPMN2.0-XML, XMI (1.2, 2.1), Excel, SCXML
yEd	
Import	Graph Markup Language (GRAPHML), yWorks Binary Graph Format, Graph Modeling Language (GML, XGML), Trivial Graph Format (TGF), Gedcom Data (GED)
Export	–

REFERENCES

[1] *Webster's Third New International Dictionary*. Merriam Webster, 1986.

[2] *IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries*. IEEE, 1991.

[3] *Longman Dictionary of Contemporary English*, 5th ed. Langenscheidt ELT, February 2009.

[4] B. Biafore, *Visio 2007 Bible*. Wiley Publishing, April 2007.

[5] U. Brandes, M. Eiglsperger, I. Herman, M. Himsolt, and M. Marshall, "GraphML Progress Report Structural Layer Proposal," in *Graph Drawing*, ser. Lecture Notes in Computer Science, P. Mutzel, M. Jünger, and S. Leipert, Eds. Springer Berlin Heidelberg, 2002, vol. 2265, pp. 501–512. [Online]. Available: http://dx.doi.org/10.1007/3-540-45848-4_59

[6] David Chen (ed.), "Practices, principles and patterns for interoperability (Deliverable D6.1)," Network of Excellence - Contract no.: IST-508 011, Tech. Rep., May 2005.

[7] M. Fowler, *UML Distilled: A Brief Guide to the Standard Object Modeling Language*, 3rd ed. Addison-Wesley, September 2003.

[8] D. Frankel, *Model Driven Architecture: Applying MDA to Enterprise Computing*. Wiley Publishing, January 2003.

[9] C. Hein, T. Ritter, and M. Wagner, "Model-Driven Tool Integration with ModelBus," in *First International Workshop on Future Trends of Model-Driven Development, FTMDD*, 2009, pp. 35–39.

[10] F. Jouault and I. Kurtev, "Transforming Models with ATL," in *Proceedings of the 2005 International Conference on Satellite Events at the MoDELS*. Berlin, Heidelberg: Springer, 2006, pp. 128–138. [Online]. Available: http://dx.doi.org/10.1007/11663430_14

[11] S. Kelly and J.-P. Tolvanen, *Domain-Specific Modeling: Enabling Full Code Generation*. Wiley-IEEE Computer Society, March 2008.

[12] H. Kern, A. Hummel, and S. Kühne, "Towards a Comparative Analysis of Meta-metamodels," in *Proceedings of the Compilation of the Co-located Workshops on DSM'11, TMC'11, AGERE!'11, AOOPES'11, NEAT'11, VMIL'11*, ser. SPLASH '11 Workshops. New York, NY, USA: ACM, 2011, pp. 7–12. [Online]. Available: http://doi.acm.org/10.1145/2095050.2095053

[13] H. Kern and S. Kühne, "Integration of Microsoft Visio and Eclipse Modeling Framework Using M3-Level-Based Bridges," in *2nd ECMDA Workshop on Model-Driven Tool and Process Integration at Fifth European Conference on Model-Driven Architecture Foundations and Applications 2009*, Enschede, Netherlands, 2009.

[14] D. Kolovos, R. Paige, L. Rose, and F. Polack. (2014, April) The Epsilon Book. [Online]. Available: http://www.eclipse.org/epsilon/doc/book/

[15] A. Ledeczi, M. Maroti, A. Bakay, G. Karsai, J. Garrett, C. Thomason, G. Nordstrom, J. Sprinkle, and P. Volgyesi, "The Generic Modeling Environment," in *Workshop on Intelligent Signal Processing*, 2001. [Online]. Available: http://www.cs.virginia.edu/~rp2h/home/research/ReadingList/gmepaper.pdf

[16] D. S. Linthicum, *Enterprise Application Integration*. Addison-Wesley, 1999.

[17] A. Molina, H. Panetto, D. Chen, L. Whitman, V. Chapurlat, and F. Vernadat, "Enterprise Integration and Networking: Challenges and Trends," *Studies in Informatics and Control*, vol. 16, no. 4, pp. 353–368, 2007.

[18] *Web Services Business Process Execution Language Version 2.0*, OASIS Std., April 2007. [Online]. Available: http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html

[19] *Business Process Model And Notation (BPMN), Version 2.0*, Object Management Group Std., January 2011. [Online]. Available: http://www.omg.org/spec/BPMN/2.0/PDF

[20] *Meta Object Facility (MOF) 2.0 Query/View/Transformation Specification, V1.1*, Object Management Group Std., January 2011. [Online]. Available: http://www.omg.org/spec/QVT/1.1/

[21] *XML Metadata Interchange (XMI) Specification, Version 2.4.2*, Object Management Group Std., April 2014. [Online]. Available: http://www.omg.org/spec/XMI/2.4.2

[22] T. Stahl and M. Völter, *Model-Driven Software Development*. Wiley, May 2006.

[23] D. Steinberg, F. Budinsky, M. Paternostro, and E. Merks, *EMF: Eclipse Modeling Framework*, 2nd ed., ser. The Eclipse Series. Addison-Wesley, December 2008.

[24] J.-P. Tolvanen, "Incremental Method Engineering with Modeling Tools: Theoretical Principles and Empirical Evidence," Ph.D. dissertation, University of Jyväskylä, 1998.

[25] A. I. Wasserman, "Tool integration in software engineering environments," in *Software Engineering Environments*, ser. Lecture Notes in Computer Science, F. Long, Ed. Springer, 1990, vol. 467, pp. 137–149. [Online]. Available: http://dx.doi.org/10.1007/3-540-53452-0_38

[26] M. N. Wicks, "Tool Integration within Software Engineering Environments: An Annotated Bibliography," Heriot-Watt University, Tech. Rep., August 2006. [Online]. Available: http://www.macs.hw.ac.uk/cs/techreps/docs/files/HW-MACS-TR-0041.pdf

[27] A. Winter, B. Kullbach, and V. Riediger, "An Overview of the GXL Graph Exchange Language," in *Software Visualization: International Seminar Dagstuhl Castle, Germany, May 20–25, 2001 Revised Papers*, ser. Lecture Notes in Computer Science. Springer, 2002, pp. 324–336.

[28] *Process Definition Interface – XML Process Definition Language, Version 2.2*, Workflow Management Coalition Std., August 2012. [Online]. Available: http://www.xpdl.org/standards/xpdl-2.2/XPDL%202.2%20(2012-08-30).pdf

Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 1639–1645

DOI: 10.15439/2014F267
ACSIS, Vol. 2

Alvis Virtual Machine

Piotr Matyasik
AGH University of Science and Technology
Department of Applied Computer Science
Al. Mickiewicza 30, 30-059 Krakow, Poland
Email: {ptm}@agh.edu.pl

***Abstract*—Alvis is a formal modelling language. It combines graphical modelling of communication schema and a high level programming language to describe behaviour of individual system entities. An Alvis model can be verified formally by using methods based on a system state space. The paper presents the design and the command list of the Alvis Virtual Machine. The aim of the project is to provide an execution environment for Alvis language. Moreover, one of the goals is to allow different hardware units to run Alvis models. Thus, a virtual machine was chosen as a solution.**

I. Introduction

ALVIS [1], [2], [3] is a formal modelling language designed to provide a user friendly method for developing concurrent systems, especially embedded ones. *Agents* are basic entities of Alvis models. Usually they run concurrently and communicate with one another. From a user point of view a model consist of two layers. The *code layer* provides a high level programming language used to describe agents behaviour. It's syntax is similar to C, Java or Pascal and it provides high level constructions as loop or conditional statements. The *graphical layer* (called a communication diagram) is a visual hierarchical language used to define communication channels between agents [1]. The language is being developed at AGH-UST in Krakow, Department of Applied Computer Science. An on-line manual and software supporting modelling with Alvis can be found at the project web site *http://fm.kis.agh.edu.pl*.

States of a model and transitions among them are represented using a labelled transition system (LTS graph [4]). An LTS graph is used to verify the corresponding model formally with model checking techniques [5]. The Alvis Compiler allows users to write LTS graphs in different formats.

Aldebaran format is used to export LTS graph to the CADP Toolbox [6]. Thus, system behaviour requirements can be provided by using μ calculus [7], [8] or XTL [9] and the CADP Toolbox can be used to check whether the model satisfies them. All things considered, the result of developing concurrent systems with Alvis is an easy to understand model with properties verified formally. Moreover it creates an open environment that allows using other tools based on system state graph semantics. The only thing that is required is an appropriate export function in Haskell.

This paper addresses the problem of executable Alvis models by using virtual machine. The paper presents binary code organisation and the virtual machine design and operation. The presented solution allows execution of a formal Alvis models.

The paper is organised as follows. Section II provides an overview of developing formal models with Alvis and associated tools. The AVM design assumptions and operation are presented in section III. The binary code organisation is described in section IV. Section V deals with details about all the AVM instructions. The paper is summarised in the final section.

II. Alvis Environment

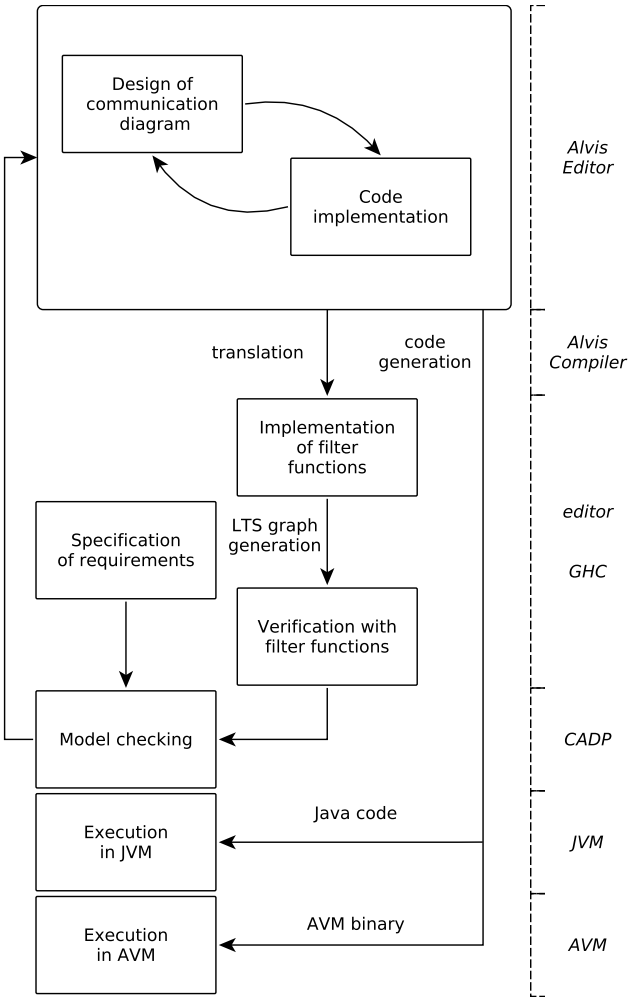


Figure 1. The modelling and the verification process with Alvis

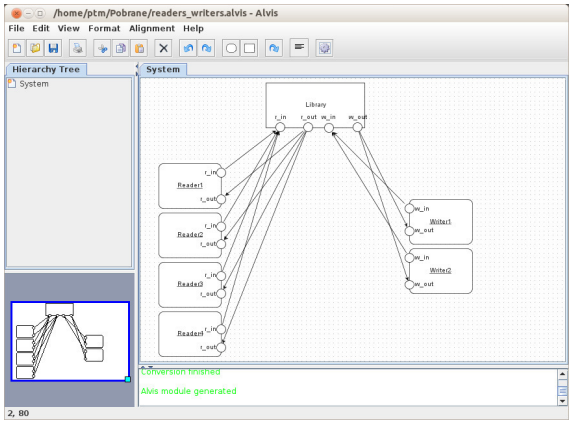


Figure 2. Alvis Editor

The scheme of the modelling, verification and execution process with Alvis is shown in Fig. 1. From a user's point of view, the process starts from designing a model using a prototype modelling environment called *Alvis Editor* (shown on Fig. 2). A system can be designed using a hierarchical representation of Alvis model. With *Alvis Editor* user can collapse and expand any valid part of a system back and forth. It is a very useful feature from the human point of view. However, before applying transformation methods, it has to be flattened. This Alvis system representation proved to be the best one for machine processing.

It is worth mentioning, that from the very beginning of Alvis all model transformations meant to be fully automatic. It was so to remove any "ideas" that user may introduce to system during human-powered translation.

A flat model is translated into Haskell [10], [4] source code and its Haskell representation is used to generate the LTS graph. A designer is able to define additional Haskell functions (called *filtering functions* [4]) that search an LTS graph for some states or parts of the graph that meet given requirements. The source code is compiled with the GHC compiler [11]. The results of the received program execution are the LTS graph for the given model and the report of the model verification with filtering functions. Further verification is performed with the CADP Toolbox [6] and the μ calculus [7], [8] or XTL [9]. Furthermore other tools that support LTS system specification might be used with little effort from the user.

For testing and educational purposes LTS graph can be exported to the DOT format and visualized. It is a useful feature during learning to model with Alvis. Unfortunately, it can be used only to fairly small systems (in terms of generated state space).

Alvis model can also be translated to an executable form. Currently Java target and a dedicated virtual machine, presented here, are supported. As it is shown on Fig. 1 it is not required for a model to be formally validated before execution. Any syntactically correct one can be translated to executable form, however it is wise to perform this step.

For more details on the Alvis syntax see [1] and the on-line manual at the project web site. The formal semantics for Alvis can be found in [2].

III. AVM DESIGN AND OPERATION

Alvis Virtual Machine was designed to run Alvis models without modifying its structure. Common problem in using formal methods in real cases is translation from a model to code. Even properly designed system can be ruined during implementation phase. Thus an *executable specification* concept was introduced and becomes more and more popular in different applications and forms f.e. [12], [13]. The whole Alvis project is a part of that conception. It brings a modelling environment and, by generating LTS of the possible system states, it provides ability to verify formally given system with tools that operate on a such representation [6].

As it will be presented in section V, AVM instruction set almost completely reflects Alvis language statements. The main goal here is to provide identical execution paths as generated in LTS graph [4]. AVM can be considered as a high level virtual machine. It has a very complex commands and is more like BEAM Erlang VM than JavaVM [14], [15]. Also the architecture of the Alvis Virtual Machine can be classified as register VM. Most of the operations are performed "in place" with variable location treated as operational registers.

Like in most virtual machines, the common operation that precedes execution, is code loading. In AVM this procedure prepares binary code for execution.

The preliminary step is checking cryptographic keys. It is based on public-private key pair. A code is signed with private key. The machine has a public key of the software supplier.

If the cryptographic signature is included in code it will be checked before execution. If it passes, the next steps for preparing code for execution will be performed, otherwise the loading will be abandoned.

There are two strategies being considered for the loader. First one, is for devices with large RAM pool. In that case, the whole binary code is loaded to RAM. It is very simple yet effective. There is no need to copy initial values for variables. All offsets are relative to the beginning of a code or the beginning of a given block.

The second code loading strategy is for devices with limited RAM capacity and FLASH memory like embedded SoC or microcontrolers. In that case a code resides in program memory alongside with AVM itself and before execution an additional step has to be performed.

Data field of the variable is moved to the RAM and an additional record is created to translate original address (in ROM/EEPROM) to the actual location. Unfortunately, this slows the execution which is the price for decreasing RAM occupancy. During execution a function which fetches variable, its pointer is modified by additional address translation step. The direct and indirect variable fetching is presented on Fig. 3.

To support online code upgrade AVM may use double code buffer. If the device is powerful enough the new code may be loaded during executing the old one.

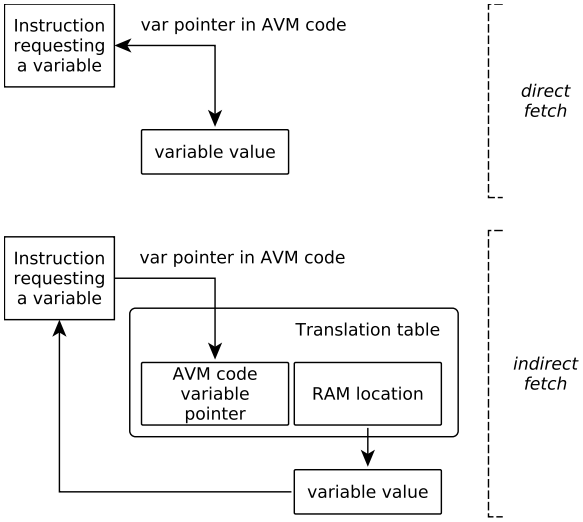


Figure 3. Direct vs indirect variable access

The machine performs all the necessary steps of preparing code for execution while executing the current one. It will be run as a background task. When new code will be ready, machine will restart and begin executing a new one.

Current AVM implementation is a bytecode interpreter. It fetches instruction from specified locations and runs them. Every AVM instruction has a function associated within virtual machine, responsible for its execution.

If a block is executed by VM, a *next instruction after block* (see block description in Section IV) is placed onto agent's *return stack*. When it reaches *block end*, this address is popped and execution continues from that location. In term of efficient virtual machine design and implementation it is suboptimal, but the main goal here is to reflect original model behaviour.

Dynamic memory management is minimized in AVM to simplify code execution and increase reliability. The only dynamic data structures required during execution are:

- stacks for agents for tracing block processing,
- data fields if code is run form ROM,
- address translation table if code is run from ROM,
- temporary storage for guard evaluation.

IV. Code format

In this section a detailed binary code organization for the Alvis Virtual Machine is presented. It is organized in top-down fashion, starting form overall description and ending with details about subsequent elements. The size information is presented as bytes.

Two byte word was chosen as a primary data size for AVM. Thus limits the maximal code size to 2^{16} bytes, including all extensions. As AVM is a high level virtual machine and its instruction are complex, it is enough for fitting quite large systems in its address space.

Table I presents general organization of the AVM code. First is the *header* block, then *functions* block. After it, agent's data are placed sequentially. The size of every block depends on the model.

Table I
General AVM binary code organization

Header	Functions	Agents	Crypto

The AVM *header* block is presented in Table II. It starts with a magic number, which is "AVM " in ASCII code. Next element is a version number. Virtual machine cannot load the code if magic word is incorrect or version number is higher than it can understand.

The function block offset represents displacement from beginning of the AVM binary code to the first function (see Sec. III). The last two elements describes agents in model. First one is an agent counter and the second is a table with offsets to a specific agent structures.

Table II
AVM Header

Name	Description	Size
MAGIC	Magic number 'AVM '	4
VERSION	Code version	4
FUN	Function block offsets table	2
ACNT	Size of the agents table	2
AGENTS	Agents block offsets table	2
SECURITY	Cryptographic extension block offset	2

The agent block is shown in Table III. It consists of: agent's name truncated to 12 characters, mainly for debug purposes, agent's code offset from beginning of AVM code, location of port and variable definitions. The overall single agent code organization is presented in Table V. The *STATE* field contains actual code pointer, execution block and agent state information (see [2]).

Table III
Active agent header

Name	Description	Size in bytes
NAME	Agent name	12
STATE	Agent state	8
CODE	Agent code block	6
PCNT	Ports count	2
PORTS	Ports table pointer	2
VCNT	Variables count	2
VARS	Variables table pointer	2

Table IV
Variable block

Name	Description	Size in bytes
NAME	Variable name	12
TYPE	Variable type	2
LOCATION	Value pointer	2

Table IV presents variable structure. The variable table consists of such elements. The whole variable block combines variable table and variable values. NAME and TYPE are left here mainly for debug reasons and their removal is considered to reduce code size.

Table V
AGENT BINARY CODE ORGANIZATION

Agent header	Ports	Variables	Code

The active port structure is presented in Table VI. This element is generated for every port's data type pair. It holds port's name type identifier for transferred data and pointer for the other side port structure.

Table VI
ACTIVE PORT BLOCK

Name	Description	Size in bytes
NAME	Port name	12
TYPE	Type of data	2
CPORT	Connected port	2

Passive agent definition is in Table VII. It is almost identical to active agent. The only difference is it holds passive ports definitions instead of active ones.

Table VII
PASSIVE AGENT BLOCK

Name	Description	Size in bytes
NAME	Agent name	12
STATE	Agent state	2
PCNT	Ports count	2
PPORTS	Ports table pointer	2
VCNT	Variables count	2
VARS	Variables table pointer	2
CODE	Code pointer	2

Table VIII
PASSIVE PORT BLOCK

Name	Description	Size in bytes
NAME	Port name	12
TYPE	Type of data	2
CODE	Port's code block	6

Table IX presents security extension block. It consists of: key type, key data length and key location as an offset from code beginning.

Key type specifies how the key data should be interpreted during code loading. Virtual machine which is not able to process a specific key type, should refuse to execute the code. Key type may by set to NULL value. In that case, no code verification is performed and key length and value pointer should also be set to NULL.

Table IX
SECURITY EXTENSION BLOCK

Name	Description	Size in bytes
TYPE	Key type	2
SIZE	Key length	2
LOCATION	Value pointer	2

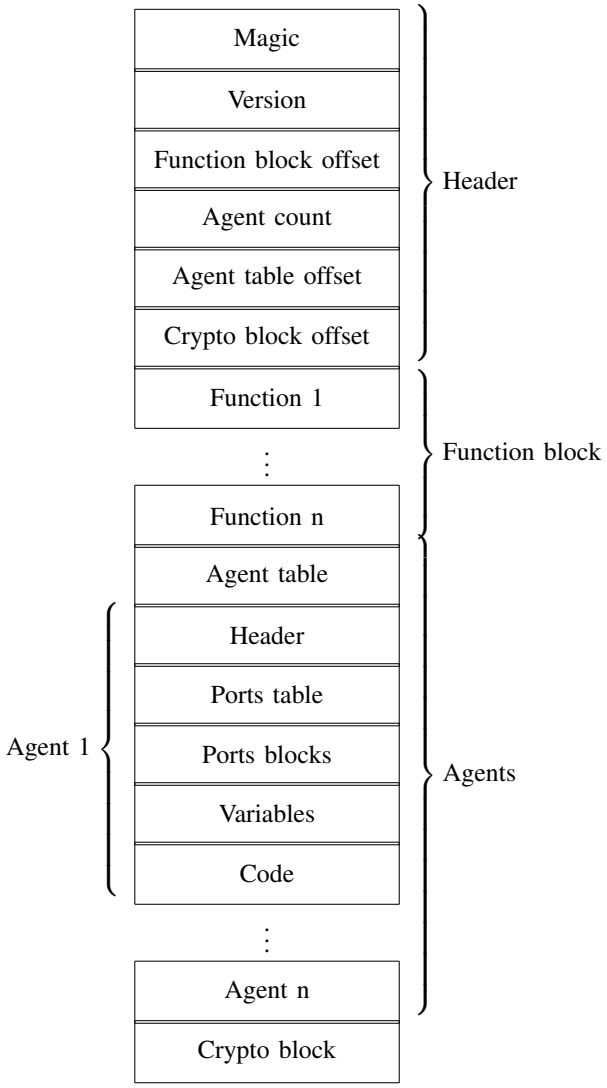


Figure 4. Example binary code map

Figure 4 shows an overview of the complete AVM binary code. The header is presented in detail. Then function block is placed. After it is an agent table. Then all the agents are located. The final block is occupied by cryptographic extension. This block location is not crucial for the AVM operation. It represents actual block placement but it can be reorganized. It is because most of the offsets are relative to the beginning of the AVM code (magic word location). This makes code generation more complicated because, in fact,

some linking operations are applied during that phase. An advantage of that approach is faster code loading and easy code execution. No linking is required during code loading. Also, multiple memory access is not required to access a specific structure nor additional memory to cache frequently used object locations.

V. INSTRUCTIONS

Below all the instructions supported by the Alvis Virtual machine are presented. All of them are shown as a small tables where there is instruction and type parameters (if any) in the first column, and short description in the second column. Binary format of the instruction is equal to the first column contents.

Instructions were divided into two blocks. The first one consists of instructions present in Alvis itself, whether the second one is composed of instructions added to allow Alvis models to be executed on small embedded systems with limited resources, in case it is impossible to fit native Haskell code.

Haskell is not a problem during generation of LTS graph because its already there. Executing even simplest functions as an arithmetic operations in Haskell is perfectly reasonable. But in embedded environment Haskell brings a huge overload to smaller systems. In such case, using it to do a few additions do not seem to be a good idea.

Moreover, during examination of already created test models during Alvis development, this subset allows for running quite large set of code without involving Haskell binaries.

The following types of arguments are used in AVM binary code:

- agP agent pointer - offset to specified agent structure;
- fP function pointer - offset to the start of specific function;
- vP variable pointer - offset to specific variable;
- pP port pointer - offset to specified port structure;
- ppP passive port pointer - offset to specified port structure;
- cP code pointer - offset to specified instruction;
- bs block specification - it consists of three code pointers (block start, block end, first instruction to execute after block);
- char 8 bit signed character;
- uint32 32 bit unsigned integer;
- int32 32 bit signed integer;
- double 64 bit double precision floating point number;
- list list of any of the simple data types (char, uint32, int32, double);

All offsets are relative to beginning of the AVM binary code. It speeds up code execution by allowing to fetch specific data without multiple memory access. Moreover, it simplifies a "virtual memory" (see Fig. 3) implementation for devices where code lays in read only area. In this case, some portions of the data have to be moved to random access memory which is required for execution. This design feature complicates a bit code generation, but it is a consequence of execution requirements for small devices.

NULL	do nothing

The *NULL* instruction does nothing. However, it increments agent's program counter and consumes some time during execution.

START	start agent
agP	agent pointer

The *START* instruction begins execution of agent pointed by the first argument. Its state is changed from *initialized* to *ready*.

EXIT	stop agent
agP	agent pointer

The *EXIT* instruction stops execution of agent pointed by the first argument. Its state is changed to *finished*.

EXEC	execute function
fp	function pointer

The *EXEC* instruction executes specified function. Function arguments are hard-coded inside so there is no need for passing them.

NEXEC	native exec
nfP	native function pointer
uint32	argument count
vP	result pointer
vP	first argument pointer
...	...
vP	n-th argument pointer

The *NEXEC* instruction is a wrapper for executing natively implemented functions. It requires pointers for all arguments and for result. Also types of AVM arguments have to match natively implemented function. Function code has to be compiled and linked with AVM.

IF	start agent
fP	guard pointer
bs	true block specification

The *IF* instruction is the simplest implementation of the *if* Alvis statement. It is the case when there is no *elseif* or *else* clause. If a guard function evaluates to *true*, it executes a code block, otherwise the next instruction specified by *bs* is selected.

IFE	if-else instruction
fP	guard pointer
bs	true block specification
bs	false block specification

The *IFE* instruction covers the case when there is an *else* statement in Alvis code. If its value is interpreted as true, its first block is executed, otherwise second block is executed.

IFEIFE	select instruction
uint32	branch count
fP	first guard pointer
...	...
fP	n-th guard pointer
bs	first branch block
...	...
bs	n-th branch block
bs	else branch block

The *IFEIFE* is the most complicated version of the *if* statement in Alvis. It is the case when there is *if-elseif-else* form of the statement. It consists of table of guards and a table of connected code blocks to execute. The last code block is an else branch one and it is executed when all guard functions evaluate to *false*. There is also the *IFEIF* virtual machine instruction. The only difference from *IFEIFE* is it has no else branch.

The *if* statement was split into several cases because during example code analysis, the most commonly used statement was *if* or *if-else* one. It was made to optimize VM instruction execution and to make implementation clearer.

LOOP	conditional loop
fP	guard pointer
bs	block specification

The *LOOP* instruction executes specified block as long as associated guard function evaluates to true. Otherwise, the next instruction denoted by *bs* is used.

LOOPE	timed loop
uint32	dalay value
bs	block specification

The *LOOPE* is a special loop instruction. It loops indefinitely, but each iteration should start every *delay value* milliseconds. To achieve the process a time stamp is taken before code block execution. After it, another time stamp is taken and the remaining time is calculated. If there is some time left, agent suspends its execution.

JUMP	unconditional jump
cP	code pointer

The *JUMP* instruction performs an unconditional jump to specified code location.

IN	input from port
vP	input variable pointer
pP	local port pointer
pP	remote point pointer

The *IN* instruction performs communication with other agent via port specified. It requires a variable structure pointer to store a new value, local port structure and remote port structure.

OUT	output to port
vP	output variable pointer
pP	local port pointer
pP	remote point pointer

The *OUT* instruction performs communication with other agent via port specified. It requires a variable structure pointer to send value, local port structure and remote port structure.

INP	input from passive port
vP	input variable pointer
pP	local port pointer
ppP	remote point pointer

The *INP* instruction performs passive agent call via specified port. The caller is an active agent. It requires a variable structure pointer to save new value, local port structure and remote passive port structure.

OUTP	output to passive port
vP	output variable pointer
pP	local port pointer
ppP	remote point pointer

The *OUTP* instruction performs passive agent call via port specified. The caller is an active agent. It requires a variable structure pointer to send value, local port structure and remote passive port structure.

INPP	input from passive port
vP	input variable pointer
ppP	local port pointer
ppP	remote point pointer

OUTpP	output to passive port
vP	output variable pointer
ppP	local port pointer
ppP	remote point pointer

The *INPP OUTPP* instructions perform passive agent call via port specified. The caller is a passive agent. They require a variable structure pointer to send or save to, local port structure and remote passive port structure.

SELECT	select instruction
uint32	branch count
fP	first guard pointer
...	...
fP	n-th guard pointer
bs	first branch pointer
...	...
bs	n-th branch pointer

The *SELECT* instruction reflects Alvis's *select* statement. It consists of a table of guard functions and a table of code blocks. Guards are evaluated sequentially. If guard evaluates to *true*, a corresponding code block is executed.

READY	check if port is ready for communication
pP	port pointer

PREADY	check if passive port is ready for communication
ppP	passive port pointer

The *READY* and *PREADY* instructions check if a specified port is ready for communication. It is required for guard functions mainly in select statement.

Table X
INTERNAL COMMANDS FOR FUNCTION EXECUTION

ADD	add operands
SUB	subtract operands
MUL	multiply operands
DIV	divide operands
HEAD	get head of a list
TAIL	get tail of a list
INS	insert element in list at the beginning
ADD	append element to list
AND	logical and
OR	logical or
NOT	logical not

Table X presents summarized list of AVM commands implemented for internal functions. Almost all of them take three arguments except for the last one which takes two. The binary layout of instructions is presented below.

{INSTR}	instruction code, see Table X
vP	result
vP	first operand
vP	second operand

All the operations are defined for appropriate datatypes. Arithmetical instructions are automatically applied to all simple types. Conversions are executed as in ISO-C standard [16].

Presented AVM instruction list should not be considered as closed. AVM is in active development phase and commands are added, removed and reorganized.

VI. Summary

The Alvis Virtual Machine was presented in the paper. The main goal of the project is to provide executable form of an Alvis models. There is an Alvis to Java conversion already done, which was a test drive for Alvis Compiler code generation facility. AVM is a second attempt for automatic Alvis code execution.

AVM was designed for executing a formally checked code in high availability environment. Thus a code signing and a double code buffer were introduced in it.

The first feature is crucial for upgrading AVM code in installations where the software provider has no full control over device and running unauthorized code is a security risk.

The second feature is required in situations when device has to work continuously while providing ability to hot code swapping.

AVM is currently under development and presented features may be a subject to a change.

References

[1] M. Szpyrka, P. Matyasik, and R. Mrówka, "Alvis – modelling language for concurrent systems," in *Intelligent Decision Systems in Large-Scale Distributed Environments*, ser. Studies in Computational Intelligence, P. Bouvry, H. Gonzalez-Velez, and J. Kołodziej, Eds. Springer-Verlag, 2011, vol. 362, ch. 15, pp. 315–341. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-21271-0_15

[2] M. Szpyrka, P. Matyasik, R. Mrówka, and L. Kotulski, "Formal description of Alvis language with α^0 system layer," *Fundamenta Informaticae*, vol. 129, no. 1-2, pp. 161–176, 2014.

[3] M. Szpyrka, P. Matyasik, and M. Wypych, "Alvis language with time dependence," in *Proceedings of the Federated Conference on Computer Science and Information Systems*, Krakow, Poland, 2013, pp. 1607–1612.

[4] ——, "Generation of labelled transition systems for alvis models using Haskell model representation," in *Proceedings of the 22nd International Workshop on Concurrency, Specification and Programming (CS&P 2013)*, vol. 1032. Warsaw, Poland: CEUR Workshop Proceedings, 2013, pp. 409–420.

[5] C. Baier and J.-P. Katoen, *Principles of Model Checking*. London, UK: The MIT Press, 2008.

[6] H. Garavel, F. Lang, R. Mateescu, and W. Serwe, "CADP 2006: A toolbox for the construction and analysis of distributed processes," in *Computer Aided Verification*, ser. LNCS, vol. 4590. Springer-Verlag, 2007, pp. 158–163. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-73368-3_18

[7] E. A. Emerson, "Model checking and the Mu-calculus," in *Descriptive Complexity and Finite Models*, ser. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, N. Immerman and P. G. Kolaitis, Eds. American Mathematical Society, 1997, vol. 31, pp. 185–214.

[8] R. Mateescu and M. Sighireanu, "Efficient on-the-fly model-checking for regular alternation-free μ-calculus," INRIA, Tech. Rep. 3899, 2000. [Online]. Available: http://dx.doi.org/10.1016/S0167-6423(02)00094-1

[9] R. Mateescu and H. Garavel, "Xtl: A meta-language and tool for temporal logic model-checking," 1998.

[10] B. O'Sullivan, J. Goerzen, and D. Stewart, *Real World Haskell*. Sebastopol, CA, USA: O'Reilly Media, 2008. [Online]. Available: http://dx.doi.org/10.1145/1668113.1668115

[11] "Glasgow Haskell Compiler documentation," http://www.haskell.org/haskellwiki/GHC.

[12] G. Cancro, W. Innanen, R. Turner, C. Monaco, and M. Trela, "Uploadable executable specification concept for spacecraft autonomy systems," in *Aerospace Conference, 2007 IEEE*, March 2007. doi: 10.1109/AERO.2007.352802. ISSN 1095-323X pp. 1–12. [Online]. Available: http://dx.doi.org/10.1109/AERO.2007.352802

[13] A. Khwaja and J. Urban, "Realspec: An executable specification language for modeling control systems," in *Object/Component/Service-Oriented Real-Time Distributed Computing, 2009. ISORC '09. IEEE International Symposium on*, March 2009. doi: 10.1109/ISORC.2009.36. ISSN 1555-0885 pp. 219–227. [Online]. Available: http://dx.doi.org/10.1109/ISORC.2009.36

[14] "Erlang documentation," http://www.erlang.org.

[15] "Oracle Java documentation," http://java.oracle.com.

[16] ISO, "Iso c standard 1999," Tech. Rep., 1999, iSO/IEC 9899:1999 draft. [Online]. Available: http://www.open-std.org/jtc1/sc22/wg14/www/docs/n1124.pdf

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1647–1656
DOI: 10.15439/2014F266
ACSIS, Vol. 2

Pragmatic Model-Driven Software Development from the Viewpoint of a Programmer: Teaching Experience

Jaroslav Porubän, Michaela Bačíková, Sergej Chodarev and Milan Nosáľ
Technical University of Košice, Department of Computers and Informatics
Letná 9, Košice, Slovak Republic
Email: {jaroslav.poruban, michaela.bacikova, sergej.chodarev}@tuke.sk, milan.nosal@gmail.com

***Abstract*—Model-driven software development is surrounded by numerous myths and misunderstandings that hamper its adoption. We have designed our course of model-driven development approach with the goal to introduce it from the viewpoint of a programmer as a pragmatic tool for solving concrete problems in development process. The course covers several techniques and principles of model-driven development instead of concentrating on a single tool. To explain these techniques we use a case-study that is iteratively developed by the students during the course. In the paper we explain the structure of our case study, contents of individual iterations, and our overall experience with this approach.**

I. Introduction

MODEL-DRIVEN software development approach (MDSD) promises increase of development speed and quality of resulting software by the use of formal model of the system as a basis for its implementation [1]. Understanding MDSD, however, suffers from several myths that hamper its adoption. This section discusses these myths to provide the motivational context to our work. In the rest of the paper we present our approach to MDSD teaching that is tailored to overcome these myths.

A. Myth 1: MDSD is a large-scale approach

MDSD is mostly viewed from the perspective of large-scale software architecture. A new system or a family of systems is supposed to be implemented by describing every significant part and aspect of the system using formal models (for example in [2]). In practice, however, it is not always the case. When a system development begins, it may not be known that a whole product family would be needed in the future. Therefore, it is not clear beforehand that model-driven development would be applicable and that investment in it would pay off.

Of course, in reality MDSD can be considered in a smaller scale, where only specific parts of the system are generated based on models. In this case, the knowledge of MDSD can be useful even for a single programmer (or a small team) working on a part of the system and introduction of MDSD may not require significant changes in the architecture of the system as a whole.

B. Myth 2: MDSD requires massive tool support

MDSD is often associated with integrated modelling tools or language workbenches. These tools cover development of meta-model, a domain-specific language used to express models and a generator that produces runnable code based on a model. Modelling tools may also provide environment for development of the model itself. These tools, however, are often complex and require high learning costs. What is more important, the use of such tools poses the risk of vendor lock-in.

Although integrated tools may be useful in a lot of situations, they are not necessarily required by the model-driven approach. It is possible to use a set of independent tools for separate parts of the model-driven development infrastructure (e.g., for language processing, for code generation, etc.). This approach allows looser coupling and greater flexibility in the choice of tools.

C. Myth 3: MDSD requires special software development process

It is considered that model-driven approach requires the use of a special software development process, where meta-model and modelling language must be completely specified and implemented before a model of a system can be developed. This opinion renders MDSD as very inflexible and incompatible with agile development processes that are currently favoured.

Modelling infrastructure, however, can be developed iteratively. Meta-model, language processor and generator can evolve together with the rest of the system. The use of small-scale MDSD and simpler tools as described in previous paragraphs greatly simplifies such iterative development process and allows using MDSD along with common agile methodologies.

D. Myth 4: MDSD is not widely used in practice

Without a deeper insight it seems that MDSD is not a widely used approach in practice. In reality, model-driven and generative approaches are indeed wide-spread and even considered a good practice for pragmatic programming [3]. Most of the examples, however, represent small-scale MDSD applications which include:

978-83-60810-58-3/$25.00

- Generators of database schema and object-relational mapping (ORM) code from the description of a data structure (used in various ORM tools).
- Generators of code for accessing web services based on WSDL description.
- Tools for graphical user interfaces design that generate code according to a graphical representation of the user interface.
- IDE plugins for specific technologies that are able to generate skeletons of repetitive artefacts (e.g., GWT plugin for IntelliJ Idea that can generate standard GWT RPC service artefacts).
- Spring Roo generative framework that allows to implement custom code generators for various repetitive code artefacts (currently published generators focus on web-based CRUD application domain).

Furthermore, the MDSD application is often hidden from a programmer by libraries and frameworks that allow to specify behaviour using a model without knowing details of model processing. In case of dynamic languages such as Ruby, internal domain-specific languages can be used for description of models and code generation can be replaced with run-time program modification using reflection. This approach makes the use of MDSD even less obvious.

II. Pragmatic Model-Driven Programming

We designed the MDSD course with these considerations in mind. The course is intended for graduate students that would mostly become software engineers in their future career. Because of this, we wanted to demonstrate the approach from the viewpoint of a programmer. This led us to the following goals:

1) *Keep it practical.* We wanted to maximize the possibility that our students would be able to use the learned skills and techniques in their future careers. This means that these techniques should be applicable in a wide range of situations.
2) *Teach principles using realistic examples.* Students should understand the basic principles of the topic as they have much higher level of applicability than any concrete tool. At the same time we should illustrate these principles using realistic examples, tools and approaches that can be directly used in practice.

For these reasons we need to teach MDSD in a way that challenges myths described in the previous section. First of all, we show the students that model-driven approach is not limited to large-scale solutions. Although we demonstrate development of a complete system using MDSD, parts of the system are modelled and generated separately showing different scales of modelling.

We also do not use any full-fledged MDSD tool for the whole development process. Instead, we concentrate on several development techniques and tools behind them from the perspective of the main components of the MDSD infrastructure. If we apply the language-oriented perspective, we can divide the components and techniques that we teach as follows:

1) *Abstract syntax* – meta-model that describes structure of models:
   - definition of meta-model using classes in object-oriented language,
   - composition and reuse of models.
2) *Concrete syntax* – domain-specific language used to specify a model:
   - implementation of delimiter-directed parser,
   - internal domain-specific languages,
   - the use of parser generator,
   - generic XML parser.
3) *Semantics* – generator used to produce code based on the model:
   - generation by direct transformation,
   - generation using templates, templates composition and reuse.

All of the listed technologies and approaches are used in the course. In the beginning of the course students are building simple version of MDSD tools themselves (e.g. parser, generator). Later on, when the complexity of tasks arises we are fluently switching to the well-known MDSD tools (e.g. parser generators, tempalting engines). In the end of the course, students are able to compare tools and approaches and choose the most effective one in a particular situation.

The techniques that we teach can be combined in various ways to power MDSD. At the same time, they can be used even separately for a wide range of programming tasks. From this aspect our approach is similar to the one used by Folwer in [4].

All taught techniques are demonstrated on a single case study that is developed in an iterative manner. Meta-model used in a case study is gradually extended showing evolution and composition of meta-models and languages during the system development. We also show that a meta-model can be reused even if the implementation of a language processor is replaced. Thus we demonstrate parallel evolution of different components of the MDSD infrastructure. It also shows the possibility to use usual agile development processes with connection to MDSD.

III. Case study

In this section we introduce the case study we use to teach MDSD in our course. *CrudComp* is a fictive company, which develops CRUD (create, retrieve, update delete) applications for different domains storing entities and their properties and relationships (e.g. employee, department). They have started with a single application but their success brought them new customers. *CrudComp* soon identified fundamental requirements shared across all the applications they developed for their customers. Each CRUD application has to provide means for:

- data entry,
- data validation,
- data persistence into an (external) memory, and
- data presentation in a user interface.

They also identified a few non-functional requirements for the applications concentrating on technology.

- *User interface technology* – the customers expect a migration to a new user interface type in the near future (e.g., mobile, web).
- *Storage technology* – currently the applications are supposed to work with a relational database, but some of the customers consider transition to a file system or a NoSQL database.
- *Service-oriented architecture* – *CrudComp* providentially expects that the customers will want to export the CRUD functionality as the web services for integration with other business applications.

These specifications led the *CrudComp* to design a multi-tier architecture for their CRUD applications consisting of 3 layers: user interface, service and data access. The architecture skeleton is depicted in Figure 1.

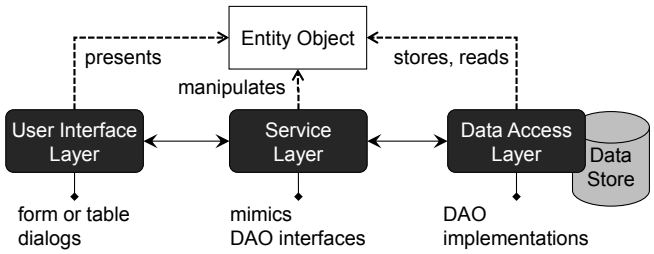


Fig. 1. Multi-tier architecture used by *CrudComp* CRUD applications

The architectural framework defined applications skeleton, but there would still be quite a lot of repetitive work. Luckily, one of the *CrudComp* employees was an enthusiast for MDSD and generative programming. The vision presented to the management was that the MDSD solution would not only speed up the development, but it would also make the communication with the customer more efficient. The domain-specific language used to specify the model of an application would be readable also by customers (in this context they were domain experts). Moreover, the customer would be even able to *write* the input file by himself. This way *CrudComp* would not only spare the time and money for the development, but it would also gain more effective ways to communicate with the customer. They have started with the application of the MDSD in a small scale (generating just small parts of the system) not to hamper the development of currently developed applications. And so he with the rest of the team decided to iteratively build a full MDSD solution with each iteration covering just a small portion of the whole CRUD application. Until the whole solution would be ready the programmers would have to implement the rest of the system manually to meet the deadlines.

The input of *CrudComp* MDSD solution was a simple model of the domain in terms of entities, their properties and relations. The output was supposed to be a part of a CRUD application. They decided to start with entity classes generation since the entity classes are the simplest artefacts of a CRUD application source code. Entity classes define the data structures manipulated in a CRUD application. Later they wanted to generate also data-access objects implementations, validators, etc. The only thing the programmers would need to do to get entity classes or later even the complete CRUD application is to write a simple input file that models the domain for a given customer. After running the generator they only needed to add a hand-written specific code to the generated source code to finish the application so that it would be fully functional and tailored for the specific customer. Thus they were able to reduce the amount of their repetitive work.

CrudComp presents a simple case of possible MDSD adoption candidate. As the reader can see we use the case study to motivate and explain the adoption of the MDSD technique. However, on the other hand we do not avoid development process issues. The case study context puts students in the role of *CrudComp* developers that need to iteratively and incrementally build an MDSD solution for a CRUD application family. Incremental MDSD adoption is a necessity to fit into the agile development processes.

We selected a CRUD application software family for our case study because it is the most common information system type in practice. It is also very similar to common project assignments at our university that our students are already used to. Moreover, examples in many tutorials (e.g. Spring, NetBeans JSF) are illustrated on a CRUD application.

During the course, each student works individually on his/her *CrudComp* project at home or in the class. Their progress is controlled in the class on a week basis to prevent procrastination and to help them with issues that raise during the development of the case study.

IV. Teaching MDSD Iteratively

The *CrudComp* case study is divided into 4 relatively self-contained incremental iterations. They all solve problems in the same problem domain – the implementation of CRUD applications. However, rather than solving the whole problem at once, each iteration solves a smaller sub-problem of the CRUD application generation. The first iteration starts with generating entity classes and DAO implementations, and each next iteration adds generation of some new functionality. This iterative approach allows us to teach 4 different approaches to MDSD while keeping the case study quite simple[1]. Thanks to the variety of the MDSD techniques used in the case study we can give the students a brief insight into the problems, advantages and disadvantages of these techniques and the students can compare them by themselves. Multiple approaches also enable us to introduce the problem of language composition.

While each iteration uses a different technique they all share the same tooling infrastructure used in MDSD (see Figure 2).

[1]Of course, this iterative approach is not only about showing multiple techniques. As we argued in Section II, its main objective is to explain (and illustrate, too) to the students that MDSD can be applied just to small portions of the whole system and one can even combine multiple approaches in context of the same system. Moreover, iterative approach nicely fits into agile development.

As the reader can see from the scheme in Figure 2 we accent the importance of the model that connects the problem domain with the implementation.

The case study is divided into 4 iterations that introduce the following techniques:

1) *Entities DSL* processed by a simple parser implemented in an ad hoc manner (delimiter-directed parser). This iteration solves the entities definition problem and enables the user of the generative system to generate the data tier for the CRUD application. Source code artefacts are generated by direct transformation and using templates.
2) *Constraints DSL* implemented as a Java-based internal language. It solves the problem of defining constraints upon entity properties and introduces the technique of language composition. Templates composition is introduced.
3) *Entities DSL with references* processed by a generated parser. The language adds a new functionality to the generated CRUD applications that allows to specify relations between entities.
4) And finally, *UI specification language* parsed by a standard XML parser. The UI specification language enables *CrudComp* to generate a standard user interface for a CRUD application. This iteration introduces templates reuse.

Each iteration ends with a full MDSD solution to a subproblem of the whole CRUD application generation. They all follow the whole scheme in Figure 2. The students can see that MDSD can be applied in a small scale and that it can be done relatively easily and quickly (they can see that even MDSD can be done in agile manner). The following sections discuss the iterations in detail and explain our motivation for each of the chosen approach.

A. Entities Language

The first iteration starts with a simple external DSL. From the viewpoint of the parsing approaches the objective of this iteration is to show the students that writing an ad hoc delimiter-directed parser for a very simple language can be the right choice – in some simple situations the "big guns" such as parser generators could just complicate the matter. To make the implementation of the entities language as simple as possible we exploit the file system for the concrete syntax (see Section IV-A2). On the other hand, the students can also see that if the language would get a little bit more complex, the parser implementation complexity could raise much more (thus we are preparing the ground for introducing other approaches).

Another reason why we start with an ad hoc parser is that students are often scared of parser generators. Usually they think parser generators are complicated and therefore can be used only by experts in language theory. We start with a simple ad hoc implementation to gain the students' attention and enthusiasm.

From the viewpoint of code generation we use both template-based generation and direct transformation. Again, we want the students to understand when a direct transformation generation is enough and when we can simplify generation with templates. To emphasize incorporating multiple approaches in one project we also generate multiple outputs from the same model.

1) Abstract Syntax: In the first iteration we start with a domain model that considers entities and their properties as the data structures handled by CRUD operations. Entity has a unique name and a set of its properties. Each property has a name too (that is unique in scope of one entity) and a type. For the purposes of the project string, integer and floating point number types suffice. The result of the first iteration is a language that covers the domain of CRUD entities.

The abstract form of a language sentence is represented by an in-memory object-oriented model – semantic model using Fowler's words [4]. For each entity there should be an in-memory object that would have a reference to a string with an entity name and a reference to a list of objects representing its properties, etc. The in-memory model is defined by GPL classes (Java classes in our case study). The language model for the problem domain of this iteration is shown in Figure 3.

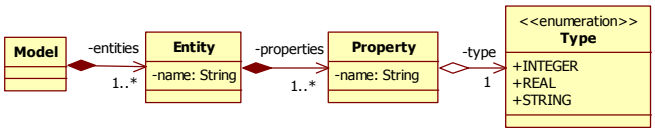


Fig. 3. Object-oriented language model of the entities language

2) Concrete Syntax: The first parsing approach we want the students to use is an ad hoc delimiter-directed parser. To keep the implementation effort manageable we chose a simple pragmatic syntax. The whole model is defined in a single directory in a file system. In this model directory there is a set of files that define entities. One file specifies one entity. The name of the entity is derived from the name of the file. This way we mimic the Java programming language that requires the name of the file to be the same as the name of the class specified by that file. Since we use standard file system to specify the set of entities in the domain model we can keep the internal structure of the entity files simple. The internal structure of the entity files is used to specify entity properties, each on a single line. An example of a language sentence is shown in Listing 1.

Listing 1. Two entities specified in a file-based entities language

```
<model>
  |———— <Department>
  |            name : string
  |            code : string
  |———— <Employee>
               name : string
               age : integer
```

The students have to implement a simple parser `LineParser` that scans a given directory for files and parses them to create in-memory objects representing entities. File system scanning is done using standard Java File API. Files are

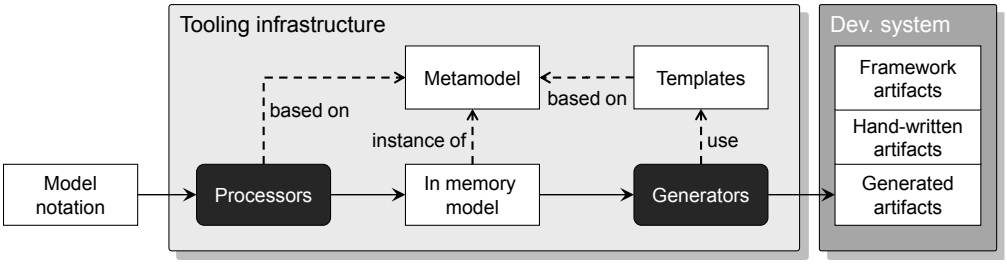


Fig. 2. Tooling infrastructure of our MDSD case study solution

text based and are parsed using just the `String` class and its methods and regular expressions. The syntax of an entity file is very simple. Each line in the file specifies a single property of the entity. The property definition starts with property name and ends with its type that is separated from the property name by a ':' character. Entity files support single line comments starting with '#'.

3) Semantics: Once the students have the in-memory model they use it to generate source code artefacts. In the first iteration they generate just a part of the whole system – the data tier of the CRUD application (simulating small-scale MDSD). For example, for the Employee entity from Listing 1 they are supposed to generate a Java entity class and a data-access object with appropriate CRUD operations. In the standard line of the case study we use JDBC to prepare SQL statements and run them on a database, but students are encouraged to use other technologies (such as Hibernate) if they have experience with them. To show to the students that we can generate multiple output artefacts from the same model the case study requires that the students would also generate a database schema creation script for a specific database (e.g. Java Derby).

We chose templates as the basic generative approach, in particular the Velocity templating engine. Both entity and DAO classes are written as templates that are instantiated using the in-memory model obtained by parsing the DSL. However, we require the database schema script to be generated using the direct transformation approach. We want the students to realize that sometimes (for very simple output artefacts or when the static portion of the generated source code is relatively small) the template-based approach is unnecessary complex and it is more appropriate to use program transformation.

B. Constraints Language

The second iteration introduces another pragmatic solution – internal language based on a host GPL. We want to show to the students that if syntactic restrictions posed by the host GPL are not a problem, an internal DSL can significantly decrease parser implementation costs.

In this iteration the students define a new language that have to be composed with the entities language implemented in the previous iteration. Since the new language models an extension of the entities domain the two languages need to be composed. Thus the students are introduced to language composition on models. And finally, in the process of code generation we show them that templates can be composed, too. Template composition can be used to modularize and simplify the templates.

1) Abstract Syntax: The second iteration extends the problem domain with property constraints. In addition to property name and type, we want to be able to specify constraints about properties. For example, the value of a particular property might be required, it might have restrictions on range or length, etc. Students will use Java-based internal domain-specific language for constraints specification. Instead of extending the LineParser they will implement a new constraints language that will be composed with the entities language.

Since the constraints internal language is a new language, it has its own model. The model of the constraints language is shown in Figure 4. Classes highlighted in red represent references to the entities language. `EntityRef` and `PropertyRef` classes have both a `name` attribute that refers to an Entity concept and to a Property concept of the original entities language, respectively. For the purpose of the composition the entities language model has to be extended, too. The `Property` class representing the Property concept of the language gets a new attribute with a list of its constraints (similarly as the `PropertyRef` class).

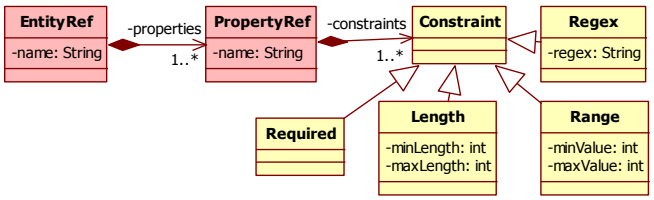


Fig. 4. The constraints language model

In addition to parsing, the constraints language have to be composed with the entities language. Students have to implement a method that validates references from the constraints language so that they refer to existing entities from the entities language model. After the validation they have to compose both models. That basically means that they have to assign constraints from the constraints language model to the appropriate properties objects from the entities language model.

2) Concrete Syntax: The second iteration of the teaching process is aimed at internal languages. Instead of imple-

menting their own parser, the students are shown how to reuse the compiler of the host general purpose language. This pragmatic solution is a façade to the language model that can be used to build constraints language expressions using domain-specific concrete syntax. In Listing 2 there is an example of a sentence specifying constraints on the name property of the Employee entity from Listing 1. The example specifies that every Employee must have a name and it cannot exceed 30 characters.

Listing 2. Constraints for the Employee entity in the constraints language

```
public class Constraints extends ConstraintBuilder {
  protected void define() {
    entity_ref("Employee",
      property_ref("name",
        required(),
        max_length(30)));
  }
}
```

The façade to the language model is called expression builder. Expression builder in the *CrudComp* case study is implemented as a Java class that provides creation methods with names from the problem domain. Part of the implementation for constraints expression builder is shown in Listing 3. It is implemented using the *nested methods* design pattern of internal languages inspired by Fowler [4].

Listing 3. Expression builder for the constraints language

```
public abstract class ConstraintBuilder {
  private List<EntityRef> entities = new ArrayList<EntityRef>();
  private Model model;

  protected abstract void define();

  protected void entity_ref
      (String name, PropertyRef... properties) {
    entities.add(new EntityRef(name, properties));
  }

  protected PropertyRef property_ref
      (String name, Constraint... constraints) {
    return new PropertyRef(name, constraints);
  }

  protected Required required() {
    return new Required();
  }
  :
}
```

3) Semantics: From the viewpoint of the constraints language semantics the students have to extend the DAO implementation to add a test method that validates objects of entity classes to match the specified constraints. Here we introduce another concept – template composition. For each constraint there should be a template just with the corresponding test. For example, in Listing 4 there is a Velocity template for the *required* constraint that tests whether an attribute of an entity class is not null (the `toUCIdent` method transforms the name into upper case identifier). In the test method of the DAO template there is a loop that goes through all the constraints assigned to the properties of the current entity class and instantiates and includes the appropriate template for each found constraint. This way we can avoid multiple 'if-else' conditional branches in the DAO template.

Listing 4. Test template for the required constraint

```
if(object.get${generator.toUCIdent($property.name)}() == null) {
  throw new ValidatorException("Property '$property.getName()'"
      + " of entity '$entity.getName()' is required.");
}
```

C. Entities Language with References

The third iteration moves the focus to the traditional MDSD tools. Now a new parser is not implemented in an ad hoc manner, but the students are supposed to work with a parser generator. The previously used approaches were supposed to show to the students that a simple DSL can be easily built without a lot of knowledge about the language theory. This iteration is used to show them that with modern approaches to parser generation, generating a parser is not difficult and for a non-trivial language it is much more effective than writing an own implementation.

To keep the course pragmatic we favoured model-based approach to parser generation (introduced in [5]). Students use the YAJCo [5] model-based parser generator that considers the object-oriented model of the entities language to be the specification of the language abstract syntax. Thus the students do not have to explicitly worry about the language grammar (although we show them the correspondence between the EBNF-based and model-based grammar specification). Using model-based parser generation does not necessarily require extensive knowledge of language and grammar theory.

1) Abstract Syntax: In the third step the problem domain is extended with relationships between entities. An entity uses a reference to other entity to express a relationship. For example, an employee works in a specific department. This relationship will be expressed by a reference from the Employee entity to the Department entity.

In this iteration we create a new parser for an external DSL that supports entities, constraints and references. The language model from previous iterations is reused thus mimicking agile evolution of the MDSD solution. Prototype parsers implemented in previous iterations are discarded, but the model and the generators are still used. A new parser populates the same `Model`, `Entity`, `Property`, and `Constraint` classes that were previously instantiated by the `LineParser` and the `ConstraintBuilder`. From the viewpoint of the language model, only the `Reference` class is an addition. Figure 5 shows the new `Reference` class with relations to the existing `Entity` class.

2) Concrete Syntax: The new entities language with references is an external language that is parsed by a generated parser. YAJCo parser generator uses the language model expressed by Java classes as a definition of the language abstract syntax. In addition to the model, students have to specify the concrete syntax of the language so that sentences like the one in Listing 5 can be processed.

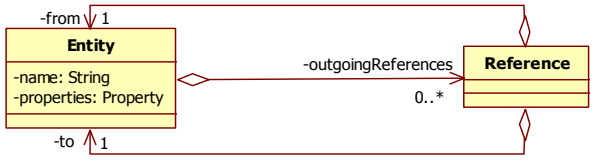


Fig. 5. `Reference` class and its relationship with the `Entity` class

Listing 5. Sentence in the entities language with references

```
entity Department {
  name : string required, length 5 30
  code : string required, length 1 4
}
entity Employee {
  name : string required, length 2 30
  age : integer
}
reference from Employee to Department
```

YAJCo uses Java annotations to associate concrete syntax patterns with the language abstract syntax expressed by Java classes. Each constructor of a language model class is considered an alternative in the grammar rule for expanding the language concept represented by that class. A constructor is a mean for creating an object from formally defined input. Annotations are used to associate concrete syntax to these rules, e.g., keywords with `@Before`/`@After`, number of occurrences with `@Range`, etc. Listing 6 illustrates modelling an expansion rule for the Entity concept of the class by annotating parameters of the constructor (for illustration of the duality there is an EBNF-based version of the rule).

Listing 6. Entity concept concrete syntax expressed by YAJCo annotations

```
public class Entity implements Named {
  :
  // Entity –> 'entity' NAME '{' Property+ '}'
  public Entity(@Before("entity") String name,
                @Before("{") @After("}") @Range(minOccurs=1)
                Property[] properties) {
    this.name = name;
    this.properties = properties;
  }
  :
}
```

In this point the students already have the abstract syntax of the language – they have the language model. To generate a parser they only need to properly annotate constructors of the model classes to specify the concrete syntax of the language.

3) Semantics: This iteration introduces only small additions to the generated artefacts. The students have to alter the database schema script generator, entity class template and DAO template to support references between entities.

D. User Interface Specification Language

The last iteration we use to teach the concept of generic languages (called Commercial-Off-The-Shelf by Kosar et al. in [6]). If a language designer keeps the syntactic restrictions defined by a generic language he/she can then reuse its generic parser. Generic languages (XML, YAML, .properties, etc.) are currently very popular in industry, especially for configuration languages [7]. This popularity is the reason why we believe that generic languages should be a part of an MDSD course. From the viewpoint of code generation and templates we use this iteration to show how the templates can be reused.

1) Abstract Syntax: In the last iteration the problem domain is extended to consider the user interface of the CRUD application. Entity objects are presented to application users in tables, each of which has a set of columns corresponding to entity properties. Not all the properties of a particular entity have to be presented and therefore there does not necessarily have to be a column for each property. To support creating and editing entity instances, a form has to be specified. Again, for each property, a field in the form can be defined.

The language model for the UI specification DSL includes new classes that describe concepts of a CRUD user interface. In Figure 6 there is a class diagram showing the language model of the UI specification language. A user interface consists of tables and forms for the entities. Tables and forms are special cases of a dialog. Each dialog has its components; in case of tables those are columns and in case of forms the components are fields. The `Dialog` and the `Component` classes have attributes prepared for composition with the entities language.

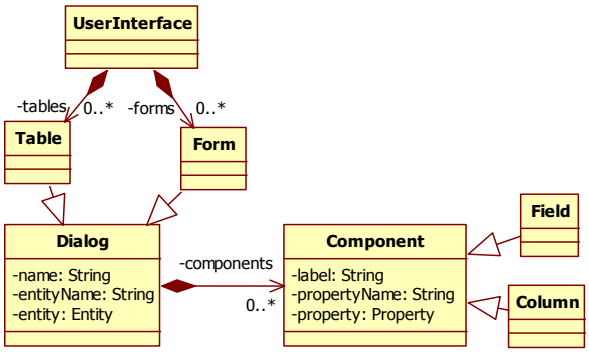


Fig. 6. User interface specification language model

2) Concrete Syntax: The concrete syntax of the UI specification language is XML-based. For parsing the language we chose the Java Architecture for XML Binding (JAXB). JAXB enables to marshal (serialize) a Java object tree into a corresponding XML document and to unmarshal (deserialize) an XML document into its in-memory object oriented representation. The classes that JAXB works with are indeed the model of the XML-based language that specifies its abstract syntax. JAXB uses *convention over configuration* design pattern to assume a concrete XML-based syntax of the language. For example, by default, JAXB maps class attributes to XML elements with the same name (mapping options between XML format and object trees are discussed in detail in [8]). The JAXB schemagen tool that can be used to generate XML Schema Definition for the XML language.

Listing 7 shows a simple user interface specification for the CRUD application with Employee entities using an XML-

based notation. Again the students' task is to annotate the language model (Figure 6) so that the JAXB would be able to marshal and unmarshal instances of the model to XML documents with the format shown in Listing 7.

Listing 7. XML-based user interface specification language

```
<ui>
  <form name="EmployeeForm" entity="Employee"
        label="Employee">
    <field property="name"/>
    <field property="age"/>
  </form>
  <table name="EmployeeTable" entity="Employee"
        label="Employee" editFormDialog="EmployeeForm">
    <column property="name"/>
    <column property="age"/>
  </table>
</ui>
```

JAXB annotations are used to specify deviations from the default mapping. An excerpt from the `Dialog` class with JAXB annotations in Listing 8 represents an example of a mapping definition. `@XMLTransient` annotations exclude program elements from mapping. E.g. the Dialog class itself is excluded since its descendants will suffice. `@XmlID` annotation specifies that the `name` attribute is an identifier of `Dialog` (or its subclasses) objects. `@XmlAttribute` overrides mapping to XML element and specifies that the `name` attribute of the `Dialog` class will be mapped rather to XML attribute.

Listing 8. Dialog class annotated with JAXB annotations

```
@XmlTransient
public abstract class Dialog implements Named {
  @XmlID @XmlAttribute(required=true)
  private String name;
  @XmlAttribute(name="entity", required=true)
  private String entityName;
  @XmlTransient
  private Entity entity;
  @XmlAttribute(required=true)
  private String label;
  @XmlTransient
  private Component[] components;
  :
}
```

3) Semantics: The last iteration finishes the CRUD application generator. The students will implement generators that will provide a user interface for the data tier generated by the generator implemented in the previous iterations. Currently as the standard line in the case study we use a console-based user interface so that the project would be simpler and at least partially manageable even for under-average students. The standard console-based UI solution requires providing templates for forms and tables. In addition, the students have to write a simple template for the main class of the application that provides the main menu for using it. Of course, we encourage the students to rewrite the project to support other types of user interfaces – we have seen multiple web-based UIs (e.g., HTML+JavaScript, Java Server Faces), desktop UIs based on Swing and also mobile clients (e.g., Windows Phone 8 communicating with server through web services, Blackberry) developed by our students.

From the viewpoint of generation techniques we ask the students to reuse the templates. The UI has to validate users input to avoid violating constraints on entity properties[2]. Here they have to reuse the constraints validation templates written in the second iteration (e.g. the template shown in Listing 4). This way they can see that a good decomposition of templates can also support template reusability.

V. Evaluation

To determine the impact of using MDSD in our course, we administered a survey to the students in our classes. 58 students responded the survey. Following questions were used in the questionnaire

A Single choice questions:

1. What were your experiences with model-driven software development (MDSD) before this course? *(a) I have not heard of it before, (b) I have heard of it before but I have never used it, (c) I have already used this approach before this course.*
2. Do you think you understood MDSD? *1 - Strongly agree, 2 - Agree, 3 - Disagree, 4 - Strongly disagree.*
3. Would you use the techniques learned in this course in practice? *1 - Strongly agree, ..., 4 - Strongly disagree.*
4. Were you satisfied with the iterative way of development used in the course? *1 - Strongly agree, ..., 4 - Strongly disagree.*
5. Rate the amount of work needed to complete the project solved in the course. *(a) Significantly more than in other courses, (b) More than in other courses, (c) Less than in other courses, (d) Significantly less than in other courses.*
6. The course belongs to your: *(a) favourite subjects, (b) rather favourite subjects, (c) rather not favourite subjects, (d) not favourite subjects.*

B Open text questions:

7. What did you like about the course?
8. What is the biggest problem you had during the course?
9. What would you change about the course?
10. Which of the learned techniques would you use and in what situations/projects/platforms?

As the reader can see, the first two questions of the questionnaire are oriented to students' knowledge about MDSD before and after the course. The 2nd and 10th question are targeted to practical usage of the learned knowledge. The rest

[2]In this simple console-based application the validation both in data tier and in UI is redundant, but we want the students to have an opportunity to reuse the templates written for constraints. In the real world, most of the common CRUD applications are web based. In web input, validation in UI forms is important for user experience. And duplicate validation on server is necessary if the server exports services that can be used to create or update entities.

of the survey addressed the course, it's form and the problems that the students might have had during the course.

Results

The results we obtained from the first two questions revealed that most of the students (57%) have never heard of MDSD and only 5% have used an MDSD approach before the course. After finishing the course, almost 86% of the students think they understand MDSD and only one student feels s/he does not understand MDSD at all.

More than a half of the students (almost 58%) think that they will use the MDSD techniques in practice. Here we have to note, that not all of our students are programmers and many students are focused on computer networks for example. According to the answers of the 10th question, more than a half of all students (51%) specified also relevant examples of using specific techniques in practice. This fact implies that *more than a half of the students sufficiently understood MDSD principles and techniques, they can distinguish between them and know how to use them in practice*.

The results of the 4th question shows that majority of the students (93%) liked the iterative approach used in our course.

It was surprising and gratifying for us to learn that although 88% of students thought the course puts an excessive amount of work on them, however 70% marked the subject as their favourite or rather favourite.

The problems that our students encountered most frequently were mainly misunderstanding of several tasks in the course materials (30%), technological issues (IDE, operating system compatibility, etc.) or Java (25%). *17% of students had no problem with the course*. Only a little number of students had problem with the techniques used - YAJCo (3 students), annotations (2 students) or velocity (7 students).

Although the students had issues, the results of the 9th question show that *more than a half* of them (52%) felt that they *would not change the course materials at all*.

The results obtained in the open text section showed that the students liked the iterative approach very much and they are satisfied with the consultation during exercises. Many of them marked the exercises as useful and they liked the implementation. Some of them like the various techniques used and favour the possibility of using the learned techniques in practice.

We can conclude that the course is successful and orientation to the practice had motivating impact on the students. Problems lied mainly in the formulation of several tasks, which were hard to interpret for weaker programmers. For this reason, we introduced discussions and evaluations of concrete tasks into our course materials, to be able to obtain task-specific feedback and improve the course materials in the future.

VI. Related Work

The motivation for teaching MDSD at our university is based on its promises of narrowing the semantic gap between problem and solution domains. Selic [9] argues that these benefits of using models are even greater in software than in other engineering disciplines (due to less diversity in skills needed for the complete MDSD implementation). Introduction of the MDSD course at our university was a response to the studies and works that on the one hand proclaim the benefits of MDSD, but on the other hand state that MDSD is given little attention in education (e.g., an early work by Cowling [10]). The main problem with MDSD teaching at our university is that myths discussed in Section I were and still are strongly rooted among our students. Although there are numerous articles describing research challenges in MDSD (e.g., work by France et al. [11]) we faced the problem of MDSD unpopularity among the students. And our students considered most of the scientific papers on the topic as just proofs of those myths (they usually deal with the highly specific problems). In our teaching approach we tried to extract the fundamental MDSD principles and show them to the students on simple pragmatic examples. The principles had to be directly applicable in practice (considering the small scale application, application in the agile methodologies, etc.).

Considering taught principles we explain the MDSD topic from the viewpoint of the language-oriented programming (see Section II). Although this viewpoint covers basically the same challenges and benefits as rather "classic" MDSD teaching approaches such as the one by Clarke et al. [12], our approach is more language-centric. We decided to extensively cover also the topic of formal languages, since currently in the industry there is ubiquitous need for developing and working with little languages (especially configuration languages [7]). Not only this course teaches the students useful knowledge, but it also serves as a motivational factor; the most important attribute of a course for students is whether they will be able to apply the learned topics in their future career.

Problem with used tools in MDSD teaching was discussed in multiple scientific works. There are cases in which teachers chose complex MDSD tools and they do not report any significant problems with students using complex solutions. For example, Tekinerdogan [13] and Clarke et al. [12] used Eclipse Modeling Project (EMP) tools, Pareto [14] used Microsoft DSL toolkit. However, there are reports of students having problems with working with such complex tools. For example, Batory et al. [15] tried to use the Eclipse Modeling Framework (EMF), but their students were overwhelmed by the technology. The failure to successfully understand and work with the EMF resulted in, using words of Batory et al., "a bitter taste" for them, and worse, even their students. We did not use a single complex tool to defy the myths about the solely large scale MDSD application and the need of massive tool support.

Schmidt et al. [16] identified three approaches to MDSD teaching:

- *Purely theoretical approach* that focuses on theoretical knowledge and neglects the practical exercise of the MDSD principles by students themselves.
- *Tool-supported approach* is a teaching approach that uses a single complex MDSD tool (e.g., EMF in [12]).

- *Practical approach* that focuses on underlying concepts rather than the use of a concrete tool.

Schmidt et al. use the practical approach in which they ask students to implement the generator tool themselves. They favoured this approach to the tool-supported approach since with the practical approach students have to directly apply the MDSD elementary principles themselves. Using a complex MDSD framework risks that some of the basic principles might be encapsulated by the framework and thus hidden from the students. Although this motivation differs from ours (we did not use a complex tool to show that MDSD can be applied without a massive tooling support) we ended with the very similar approach focused rather on principles than on tools.

Considering the domain of the course project, the used domain usually differs from work to work. For example, Mosterman [17] uses the domain of embedded systems, Clarke et al. [12] use the domain of communication services, or Batory et al. [15] let the students choose a domain of their interest. For the case study in our course we selected the domain of CRUD applications for two reasons, (1) it is a well-known domain for our students, and (2) it is widespread in practice (considering frameworks such as Spring Roo or Ruby on Rails). While the usage of a well-known domain does not bother the students with unnecessary learning load, the fact that the domain is widespread in the industry serves as a motivational factor.

From the viewpoint of the teaching approach, most of the articles report using classic development with a single iteration (e.g., Clarke et al. [12]). We use iterative approach to show the options in using MDSD for incremental, agile development and also to reduce the focus on a complex MDSD tool and to rather move it to MDSD principles. Iterative teaching approach is also used by Schmidt et al. [16]. They use the iterative approach for the same reason as we do; they want to focus on MDSD principles rather than on tools. In the first iteration their students implement their own generator tool, in the second iteration they extend the tool, and only in the last iteration they implement a language using a complex MDSD tool.

VII. Conclusion

In this paper we have presented our approach to teaching model-driven software development. The goal of our course is to explain the basic principles and concepts of model-driven and generative development. These concepts are illustrated using several different practical tools and techniques that can be used in different combinations and in projects of different scale. The presented approach could be also inspirative when adapting the model-driven approach in a software project. Thanks to the case study students can acquire practical experience with each of presented techniques during the course. Iterative character of development also provides insight into the use of MDSD as a part of the development process.

Acknowledgment

This work was supported by project KEGA No. 019TUKE-4/2014 Integration of the Basic Theories of Software Engineering into Courses for Informatics Master Study Programmes at Technical Universities - Proposal and Implementation.

References

[1] T. Stahl, M. Voelter, and K. Czarnecki, *Model-Driven Software Development: Technology, Engineering, Management*. John Wiley & Sons, 2006. ISBN 0470025700

[2] A. Demir, "Comparison of Model-Driven Architecture and Software Factories in the Context of Model-Driven Development," in *Proceedings of the Fourth Workshop on Model-Based Development of Computer-Based Systems and Third International Workshop on Model-Based Methodologies for Pervasive and Embedded Software*, ser. MBD-MOMPES '06. Washington, DC, USA: IEEE Computer Society, 2006. doi: 10.1109/MBD-MOMPES.2006.5. ISBN 0-7695-2538-5 pp. 75–83.

[3] A. Hunt and D. Thomas, *The pragmatic programmer: from journeyman to master*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999. ISBN 0-201-61622-X

[4] M. Fowler, *Domain Specific Languages*. Addison-Wesley Professional, 2010. ISBN 0-321-71294-3

[5] J. Porubän, M. Forgáč, M. Sabo, and M. Běhálek, "Annotation Based Parser Generator," *Computer Science and Information Systems*, vol. 7, no. 2, pp. 291–307, Apr. 2010. doi: 10.2298/CSIS1002291P. [Online]. Available: http://www.comsis.org/archive.php?show=ppr230-0911

[6] T. Kosar, P. E. Martínez López, P. A. Barrientos, and M. Mernik, "A preliminary study on various implementation approaches of domain-specific language," *Inf. Softw. Technol.*, vol. 50, no. 5, pp. 390–405, Apr. 2008. doi: 10.1016/j.infsof.2007.04.002

[7] M. Nosáľ and J. Porubän, "XML to Annotations Mapping Patterns," in *2nd Symposium on Languages, Applications and Technologies*, ser. OpenAccess Series in Informatics (OASIcs), J. P. Leal, R. Rocha, and A. Simões, Eds., vol. 29, 2013. doi: 10.4230/OASIcs.SLATE.2013.97. ISBN 978-3-939897-52-1. ISSN 2190-6807 pp. 97–113.

[8] R. Lämmel and E. Meijer, "Revealing the X/O impedance mismatch: changing lead into gold," in *Proceedings of the 2006 international conference on Datatype-generic programming*, ser. SSDGP'06. Berlin, Heidelberg: Springer-Verlag, 2007. doi: 10.1007/978-3-540-76786-2_6. ISBN 3-540-76785-1, 978-3-540-76785-5 pp. 285–367.

[9] B. Selic, "The Pragmatics of Model-Driven Development," *IEEE Softw.*, vol. 20, no. 5, pp. 19–25, Sep. 2003. doi: 10.1109/MS.2003.1231146

[10] A. J. Cowling, "Modelling: a neglected feature in the software engineering curriculum," in *Proceedings of the 16th Conference on Software Engineering Education and Training 2003*, ser. CSEE T 2003, March 2003. doi: 10.1109/CSEE.2003.1191378. ISSN 1093-0175 pp. 206–215.

[11] R. France and B. Rumpe, "Model-driven Development of Complex Software: A Research Roadmap," in *2007 Future of Software Engineering*, ser. FOSE '07. Washington, DC, USA: IEEE Computer Society, 2007. doi: 10.1109/FOSE.2007.14. ISBN 0-7695-2829-5 pp. 37–54.

[12] P. J. Clarke, Y. Wu, A. A. Allen, and T. M. King, "Experiences of Teaching Model-Driven Engineering in a Software Design Course," in *ACM/IEEE 12th International Conference on Model Driven Engineering Languages and Systems*, ser. MODELS'09. Denver, Coloardo, USA: IEEE Computer Society, 2009.

[13] B. Tekinerdogan, "Experiences in teaching a graduate course on model-driven software development," *Computer Science Education*, vol. 21, no. 4, pp. 363–387, 2011. doi: 10.1080/08993408.2011.630129

[14] L. Pareto, "Teaching Domain Specific Modeling," *Symposium at MODELS 2007*, p. 7, 2007.

[15] D. S. Batory, E. Latimer, and M. Azanza, "Teaching Model Driven Engineering from a Relational Database Perspective," in *MoDELS*, ser. Lecture Notes in Computer Science, A. Moreira, B. Schätz, J. Gray, A. Vallecillo, and P. J. Clarke, Eds., vol. 8107. Springer, 2013. doi: 10.1007/978-3-642-41533-3_8. ISBN 978-3-642-41532-6 pp. 121–137.

[16] A. Schmidt, D. Kimmig, K. Bittner, and M. Dickerhof, "Teaching Model-Driven Software Development: Revealing the "Great Miracle" of Code Generation to Students," in *Sixteenth Australasian Computing Education Conference (ACE2014)*, ser. CRPIT, J. Whalley and D. D'Souza, Eds., vol. 148. Auckland, New Zealand: ACS, 2014, pp. 97–104. [Online]. Available: http://crpit.com/confpapers/CRPITV148Schmidt.pdf

[17] P. Mosterman, "Automatic Code Generation: Facilitating New Teaching Opportunities in Engineering Education," in *36th Annual Frontiers in Education Conference*, Oct 2006. doi: 10.1109/FIE.2006.322699. ISSN 0190-5848 pp. 1–6.

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1657–1665
DOI: 10.15439/2014F131
ACSIS, Vol. 2

MuSCa: A Multiscale Characterization Framework for Complex Distributed Systems

Sam Rottenberg*, Sébastien Leriche†, Chantal Taconet*, Claire Lecocq* and Thierry Desprats‡
* Institut Mines Télécom/Télécom SudParis, CNRS UMR 5157 SAMOVAR
Email: firstname.lastname@telecom-sudparis.eu
† Toulouse University, ENAC, Email: firstname.lastname@enac.fr
‡ Toulouse University, CNRS UMR 5505 IRIT, Email: firstname.lastname@irit.fr

***Abstract*—Nowadays, complex systems are distributed over several levels of Information and Communications Technology (ICT) infrastructures. They may involve very small devices such as sensors and RFID, but also powerful systems such as Cloud computers and knowledge bases, as well as intermediate devices such as smartphones and personal computers. These systems are sometimes referred to as multiscale systems. The word "multiscale" may qualify various distributed systems according to different viewpoints such as their geographic dispersion, the networks they are deployed on, or their users' organizations. For one entity of the multiscale system, communication technologies, non-functional properties (for persistence or security purpose) or architectures to be favored may vary from one scale to another. Moreover, ad hoc architecture of such complex systems are costly and non-sustainable. In this paper, we propose a scale-awareness framework, called MuSCa. This framework includes a characterization process based on the concepts of viewpoints, dimensions and scales. These concepts constitute the core of a dedicated metamodel. The proposed framework allows multiscale software designers to share a taxonomy for qualifying their own system. At system design time, the result of such a qualification is a model from which the framework produces scale-awareness artifacts. As an illustration of this model-driven approach, we show how multiscale probes are generated to provide multiscale components with an embedded scale-awareness ability.**

***Index Terms*—Multiscale Distributed systems, Model Driven Engineering**

I. Introduction

SEVERAL recent research works [1], [2], [3] consider complex distributed systems that include both very small systems such as objects from the Internet of Things (IoT) paradigm, and powerful systems such as those found in the Cloud. This collaboration enables each system to benefit from the capabilities of the others. Some of these systems also involve intermediate computers such as mobile devices or proximity servers. Those complex systems could also be viewed as *multiscale distributed systems*.

As stated in [4], a "complex system" is any system comprised of a great number of heterogeneous entities, where local interactions among entities create multiple levels of collective structure and organization. Identifying underlying superstructures of complex systems is a challenge. A multiscale analysis of complex systems provides reduced views of those systems with simplified structures, such as presented in [5].

Multiscale distribution is a different concept from large-scale distribution. Large scale has a quantitative meaning, whereas multiscale has an heterogeneity meaning [6], [7]. The system heterogeneity may come from differences of latency or protocols of involved networks, from differences of storage capacity or nature of devices, or from dispersion variations between entities. We propose to study the multiscale nature of a complex system at design time. Approaches that allow developers to work at a high level of abstraction are needed. Model-Driven Engineering (MDE) approaches may help to describe complex systems at different levels of abstraction and from a variety of perspectives [8].

The contribution of this paper is a multiscale characterization framework, called MuSCa (***Mu**lti**S**cale **Ch**a**r**acterization framework*). This framework provides a multiscale taxonomy to describe at design time the multiscale nature of complex distributed systems. The first contribution is a multiscale characterization process. It is based on the concepts of viewpoints, dimensions and scales. We follow a model-driven approach to produce a multiscale characterization editor to qualify complex distributed systems. As a second contribution, multiscale probe artifacts are generated for runtime scale-awareness purpose. With those artifacts, system entities become aware of their place in the organization of the system. In the future, multiscale characterization, and multiscale probes may enable software stakeholders to build, deploy, and manage complex distributed systems.

This paper is organized as follows. Section II presents the motivations for a multiscale characterization framework. Then, Section III proposes a generic characterization process for multiscale systems. The MuSCa framework is presented in Section IV. Finally, Section V presents related works, and Section VI concludes the paper.

II. Motivations for a multiscale characterization framework

This Section presents the motivations for a multiscale characterization framework. Section II-A discusses the heterogeneity of complex distributed systems from the ICT infrastructure viewpoint. Afterwards, Section II-B details some motivating examples for multiscale characterization. Finally, based on these motivations, Section II-C outlines our contribution.

978-83-60810-58-3/$25.00

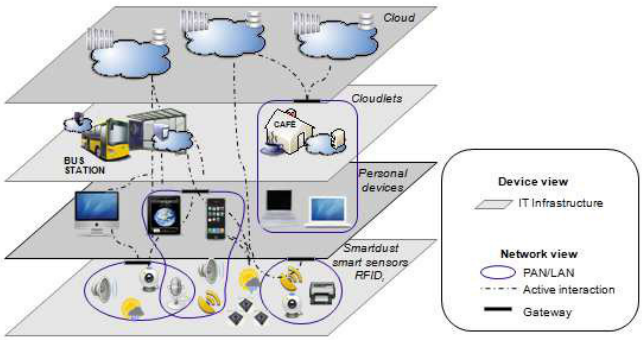


Fig. 1. ICT infrastructure levels of a multiscale distributed system

A. Complex distributed systems and multiscale distributed systems

Fig. 1 depicts our vision of various ICT infrastructure levels that compose complex distributed systems. This figure considers four ICT levels: at the bottom, the smartdust, Radio Frequency Identification (RFID), smart sensors level; at the top, the Cloud computing level; and two intermediate infrastructure levels: firstly personal devices and secondly cloudlets [9] in cafes and bus shelters.

A system, such as the one depicted in Fig. 1 has the topology of a complex system. The interactions between the system entities are decentralized. Many interactions are local but some of them take place between levels. Therefore, the interaction graph between the entities is non trivial, and is difficult to simplify. Moreover this kind of complex systems is composed of a great number of heterogeneous entities, collaborating through different networks, protocols, rules, depending on their respective organizations, geographical distance for instance.

In order to better understand and master the inner complexity of such highly heterogeneous systems, the study of the multiscale nature of a complex system may help to obtain meaningful organizations of a complex system.

B. Motivating scenario

This section presents a motivating scenario that is used throughout the paper in order to illustrate and to evaluate our contribution. This scenario involves a context management system, which is a complex distributed system. This system is deployed through many entities in the city of Toulouse in France. This system aims at enabling a large number of end-users and connected objects to share their context information, such as their location. This scenario is studied through two different aspects: the deployment aspect and the context data filtering aspect.

Considering the deployment aspect, in this scenario it is required that for each local network (LAN) containing at least one context-aware object, one software component dedicated to filter context management data must be installed in the same LAN, on any device having a bandwidth greater than 50 MB/s. Moreover, to get a scalable architecture, another software component dedicated to route context management data must be installed in a hierarchical way, one for each geographical dimension of the city of Toulouse (one for each building, one for each district and one for the city). At last, the routing component for the city of Toulouse must be installed in a cloud and end-user context-aware components must be installed on smartphones.

The second aspect is the context data filtering aspect. In the studied scenario, the user should be able to express constraints about the information he or she wants to receive and about the users that are allowed to receive his or her context information. The two following use cases are studied in this paper. In the first use case, a user, called Sophie, is going to the theater and wants to find a place to park her car. She wants to use the context management system with her smartphone to see the parking places available at foot distance from the theater, where she has just arrived. In the second use case, Sophie wants to share her location, but only with her friends located in her neighborhood.

In this scenario, it is required to express scales of distances (e.g., foot distance, same neighborhood) between system entities—i.e., users and parking places—but also scales of devices (e.g., smartphone, super-computer), network topology, or even geographical administration. These scales constraints drive interactions between entities.

C. From motivations to our contribution

All these use cases motivate the need for a multiscale vision of complex distributed systems. This vision enables a system designer or a user to express constraints concerning different points of view (e.g, geographic dispersion of system entities, network organization, social organization, devices). The solution proposed in this paper is a multiscale vocabulary on which is based a multiscale characterization process for complex distributed systems. This process is then formalized and applied through a MDE approach. The main results of this approach are, firstly, a shared extendable multiscale taxonomy, which can be used to characterize a system or express configuration and behavior constraints, and secondly, generated artifacts, which enable to enforce the constraints at runtime.

III. Multiscale characterization process of complex distributed systems

The MuSCa approach is presented in Section III-A. Then, Section III-B defines a multiscale vocabulary. This vocabulary is illustrated for the geography viewpoint in Section III-C. Other multiscale viewpoints are discussed in Section III-D. Finally, Section III-E presents the full characterization process.

A. MuSCa approach

Fig. 2 depicts the general approach followed by the MuSCa framework. The design process, which is detailed with a SPEM [10] diagram, is composed of two main activities. The first activity is the multiscale characterization to render a multiscale analysis of a complex distributed system. In order

to guide the system designer and to capitalize on previous characterizations, this activity takes the MuSCa taxonomy as an input. This taxonomy contains all the multiscale characterization terms that have already been used in previous characterizations. The result of this activity is a MuSCa model, which is a restriction of the MuSCa taxonomy. If a system designer wants to use new multiscale terms when describing the model, these terms are added to the taxonomy and will be available for the next characterization. In the second activity, multiscale probe artifacts are generated from a MuSCa model. Those probes consolidate data provided by lower level probes, called basic probes. They enable to identify the scales of the system entities at runtime.

B. Multiscale vocabulary

The multiscale characterization process must be based on a precise vocabulary. In the following the concepts of viewpoint, dimension, measure, scale, scale set are defined. These concepts will constitute the structure of the MuSCa metamodel presented in Section IV-B.

The architecture of a system is obtained by studying this system from different *viewpoints*, each viewpoint leads to a view of the system [11].

Each viewpoint is studied through *dimensions*. A multiscale dimension is a measurement of a particular characteristic of a system view for a particular viewpoint.

A dimension is associated to a *measure*, which can be either numeric or semantic.

Using a measure, a dimension can be divided into *scales*. A scale matches, respectively, orders of magnitude for numeric measures, or sets of elements that share common semantic characteristics for semantic measures.

A *scale set* is the set of scales chosen for a given dimension and measure couple.

C. Geography viewpoint

To illustrate the MuSCa vocabulary, Fig. 3 presents possible dimensions and scales for the geography viewpoint. The study of other viewpoints is available for download[1].

For the geography viewpoint, we have chosen to study the multiscale nature of a system through two dimensions, respectively associated with one numeric measure and one semantic measure. The *distance* dimension measures in meter the maximum distance between a set of entities. A set of four scales has been selected for this numeric measure: *local* under $10m$, *footdistance* between $10m$ and 10^3m, *cardistance* between 10^3m and 10^5m, and *planedistance* above 10^5m. For this dimension, the center of each scale is distant from several orders of magnitude from the center of the other scales. The *administrative division* dimension is associated to a semantic measure. It is also applied to a set of entities. It measures their smallest common division. For this analysis, we have selected a set of six scales: *Building*, *District*, *City*, *Region*, *Country* and *World*. One can notice that according to the chosen dimension and scales to study a geography view of the system, the possible organizations and interactions between entities of the system differ.

[1]http://anr-income.fr/uploads/MultiscaleViewpoints.pdf

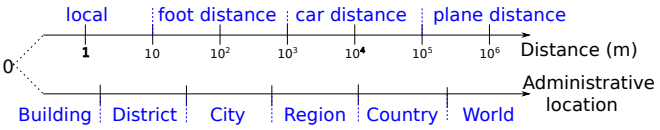


Fig. 3. Dimensions and scales for the geography viewpoint

D. Other multiscale viewpoints

The geography viewpoint is not the only viewpoint to analyze during the multiscale characterization of a distributed system. There are other fundamental viewpoints to consider such as the study of the devices of the system (device viewpoint), the social organization (user viewpoint), the administration organization (administration viewpoint), the network connections between the system entities (network viewpoint). Indeed, they are related to the main issues at stake in the design, implementation and deployment of such complex systems: the need of computing power or storage capacity (device viewpoint), of interaction between distant entities (network viewpoint), and of social organizations (user viewpoint).

However, the above viewpoint list is not exhaustive and other viewpoints, such as data, or time, could also be considered. Depending on the properties to be highlighted for the systems, one may choose to study different viewpoints and dimensions; this is the reason why we propose an open multiscale characterization process.

E. Multiscale characterization process

The multiscale nature of a distributed system should be studied independently from each considered viewpoint. For each viewpoint, a restricted view of the system is considered. Then, for this view, one or several dimensions are chosen. To identify scales for a given dimension, each dimension is associated with a numeric or a semantic measure. Depending on the type of the measure, the resulting scales match, respectively, orders of magnitude for numeric measures, or sets of elements that share common semantic characteristics for semantic measures. The choice of the viewpoints, dimensions/measures and of the scales relevant for a multiscale characterization, is left open depending on the properties of the system one wants to highlight. The objective of these choices is to put to the fore specific characteristics to deal with during the system design, or the system runtime.

The multiscale nature of a system is relative to a multiscale characterization; it is studied independently from each viewpoint. For a given viewpoint, and for a couple dimension/measure, each element of the restricted view of the system is associated with a scale. For a given characterization, and a given viewpoint, a distributed system is qualified as multiscale when, for at least one dimension, the elements of its view are associated with different scales.

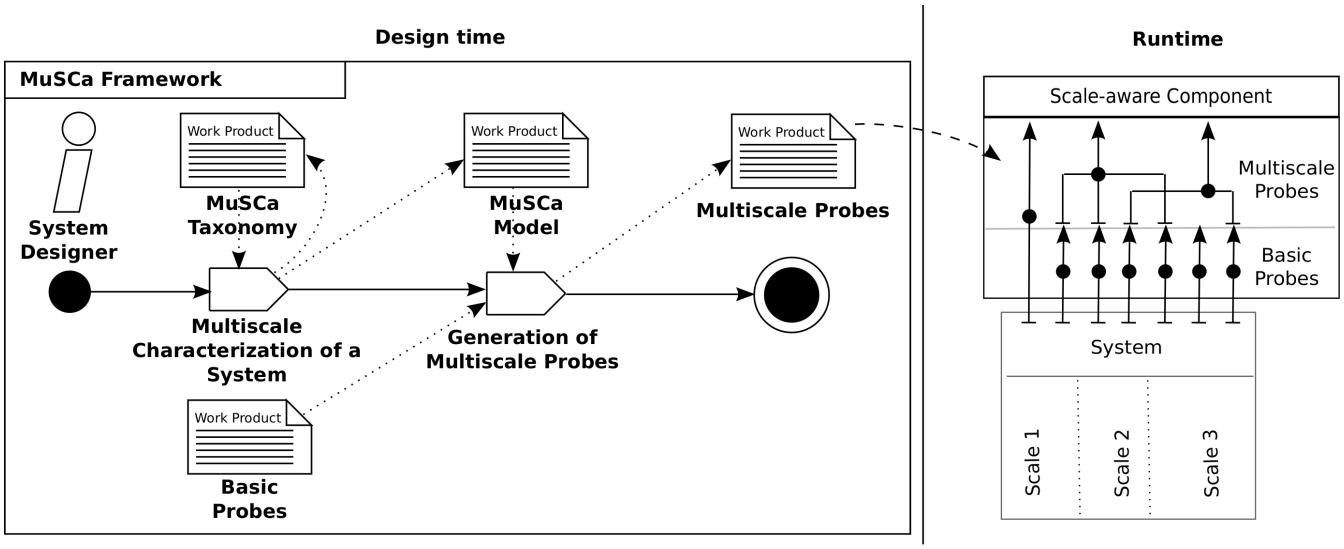


Fig. 2. MuSCa framework approach

IV. MuSCa framework

This section presents the MuSCa framework. Firstly, Section IV-A presents the model-driven approach. Then, Section IV-B formalizes the characterization process with the MuSCa metamodel, and Section IV-C gives an example of a MuSCa model. Thereafter, Section IV-D describes the generated artifacts —i.e., multiscale probes. Finally, Section IV-E presents some MuSCa implementation details and different utilizations of MuSCa are given in Section IV-F.

A. Model-driven approach

In order to formalize the multiscale characterization process, and to use it in the design and deployment of scale-aware distributed systems, we have chosen to follow a model-driven approach (using the four OMG meta-modeling layers [12]). Fig. 4 shows the mapping between the model-driven architecture levels and the MuSCa levels. The MuSCa metamodel (M2 level) is defined with the Ecore meta-metamodel (M3 level). The classes of the MuSCa metamodel represent multiscale concepts. This metamodel is used to define characterization models (M1 level). This characterization may be used for one or several real world systems (M0 level). We also follow the model-driven approach in order to automatically produce artifacts, for instance, producing probe artifacts for scale-awareness.

B. MuSCa metamodel

The MuSCa metamodel is shown in Fig. 5. This metamodel is based on the vocabulary used in the multiscale characterization process.

An instance of MSCharacterization is the result of a characterization process. A characterization considers several Viewpoints —e.g., Geography, User, Device and Network viewpoints (M1 level classes). Each viewpoint determines a restricted view of the system that is studied independently. A

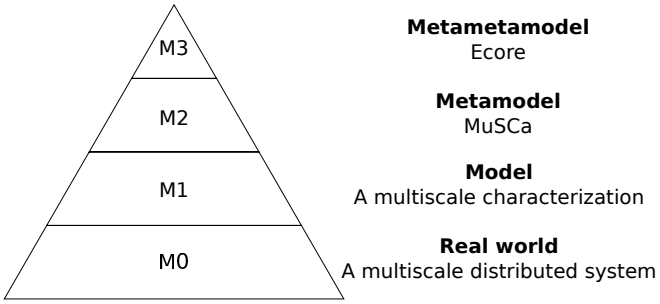


Fig. 4. MuSCa: Model-driven architecture levels

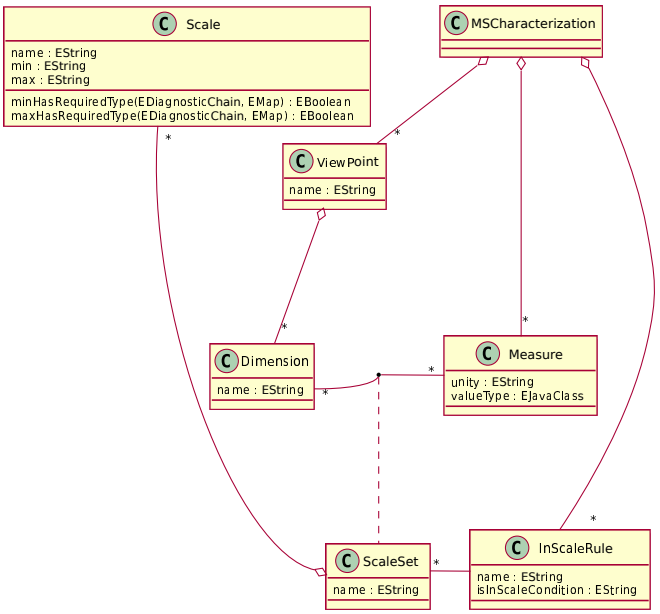


Fig. 5. MuSCa: Multiscale characterization metamodel

view of the system from a given viewpoint is studied through several Dimensions, which are measureable caracteristics of the elements of the view. For example, the Device viewpoint can be analyzed through the StorageCapacity (M1 level) dimension of the system devices. As previously mentioned, a Dimension is measurable, meaning that it can be associated with one or several Measures. For example, at M1 level, the StorageCapacity dimension may be measured with the Bytes measure or the KiloBytes measure. For the association of one dimension with one measure, the designer defines a ScaleSet, which is an ordered set of scales relevant for the system. Each scale set is associated to an InScaleRule, which determines the condition that the measured dimension of an entity, for this scale set, must satisfy so that the entity is associated to a scale. For example, for numeric measures, a Scale is defined by its min and max bounds and a rule can express that an entity is associated to a scale because its measured value is strictly between min and max. Finally, for some viewpoints, the system may present several instances of one scale. For example, if we take the Geography viewpoint, in the AdministrativeLocation dimension, the City scale (M1 level) often has several instances (M0 level) —i.e., the different cities where entities of the system are present.

C. MuSCa model as a multiscale characterization

Fig. 6 illustrates an extract of a MuSCa model that is the result of a characterization process applied to the scenario presented in Section II-B. As mentioned in Section III-A, this model is a restriction of the MuSCa taxonomy. Four viewpoints have been selected: device, network, geography and user viewpoints (the figure only shows the geography viewpoint scales).

We study the geography viewpoint through two dimensions. The first dimension, which we call the "smallest common location" dimension, measures the distribution of the system by studying the smallest common administrative location of a set of scale-aware entities. For this dimension, measured in what we call the "smallest common location measure" (semantic measure), we identify the following scales: building, district, city, region, country, and world. The second dimension, called the "smallest common distance" dimension, studies the distribution of system entities in terms of distances between each other. For this dimension, measured in meters (numeric measure), we identify the following scales: local, foot distance, car distance, and plane distance. Scales are characterized by the min and max bounds in meters. An illustration of these two geography dimensions can be found in Section IV-F, Fig. 7.

D. From MuSCa model to multiscale probes

From a MuSCa model, multiscale probe artifacts are automatically produced. These probes are monitoring programs that are to be deployed on each entity of a multiscale system. One probe is generated for each viewpoint. Each probe exposes at least one method by dimension. This mandatory method returns a scale for a set of entities (e.g., the smallest

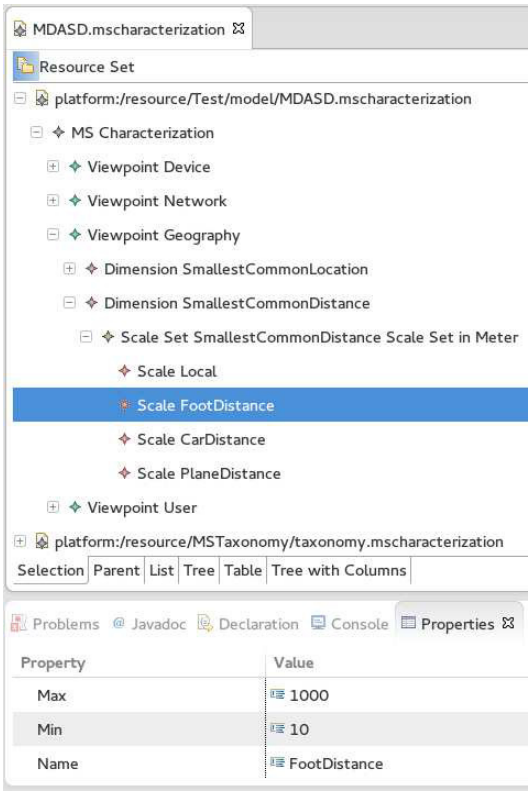


Fig. 6. Example of a MuSCa model

common location measure may return the City scale). For some dimensions, MuSCa also generates one method, which returns the scale instance for a set of entities (e.g., the name of the city). For numeric measures, the methods are automatically generated. The generated methods can be completed to implement a specific logic, in particular to call basic probes, as shown in Fig. 2, or to implement specific semantic measures.

An extract of an automatically generated probe is presented in Listing 1. This probe exposes the getScale method line 5, which uses the getValue method line 3 to call the basic probe. The getScale method returns one of the scales listed between lines 2 and 5 in Listing 2. This last listing contains two generated methods: isInScale method at line 26, which tests if a given measure value is between min and max bounds of a given scale, and getScale method at line 15, which returns the first scale corresponding to a measure value.

```
1  public interface IGeographyMeasurable {

3    public Meter_Value
4      getValue_SCD_In_Meter(List<IGeographyMeasurable> args);
5    public SCD_In_Meter_Scale
6      getScale_SCD_In_Meter(List<IGeographyMeasurable> args);

8    // same methods for SmallestCommonLocation dimension

10 }
```

Listing 1. Extract of generated IGeographyMeasureable interface

```
1  public enum SCD_In_Meter_Scale {
2    LOCAL("0", "10"),
3    FOOTDISTANCE("10", "1000"),
```

```
4     CARDISTANCE("1000", "100000"),
5     PLANEDISTANCE("100000", "Infinity");

7     private final Meter_Value min;
8     private final Meter_Value max;

10    SCD_In_Meter_Scale(String min, String max) {
11      this.min = new Meter_Value(min);
12      this.max = new Meter_Value(max);
13    }

15    public static SCD_In_Meter_Scale
16      getScaleFromMeasureValue(Meter_Value value) {
17        for (SCD_In_Meter_Scale scale :
18          SCD_In_Meter_Scale.values()) {
19            if (scale.isInScale(value)) {
20              return scale;
21            }
22        }
23        return null;
24    }

26    private boolean isInScale(Meter_Value value) {
27      return (value.compareTo(this.min) >= 0)
28        && (value.compareTo(this.max) <= 0);
29    }
30  }
```

Listing 2. Extract of generated SmallestCommonDistance_In_Meter_Scale enumeration

This probe has been generated with the Acceleo[2] code generator. As an example, Listing 3 shows an extract of an Acceleo template. When this template is applied to the MuSCa model presented in Fig. 6, it generates the scales listed between lines 2 and 5 of Listing 2.

```
1 [for (scale : Scale | aScaleSet.scales) separator(',\n')]
2   [scale.name.toUpper()/]("[scale.min/]", "[scale.max/]")
3 [/for];
```

Listing 3. Extract of a MuSCa Acceleo template

E. MuSCa implementation

We have implemented MuSCa with the Eclipse Modeling Framework Project[3] (EMF). The MuSCa metamodel is defined as an instance of the Ecore metametamodel. EMF generates a specialized model editor. We have extended the editor for validation purpose. This editor has been used to define the MuSCa model presented in Fig. 6. Then, the Acceleo code generator is used to produce multiscale probes implemented in Java.

For illustration purpose, we have detailed in this paper the multiscale probe generated for the geography/smallestCommonLocation viewpoint/dimension couple. The probes corresponding to the device/storageCapacity and geography/smallestCommonDistance couples have also been implemented. Moreover, other multiscale probes can be generated through the same process, for any given viewpoint/dimension couple, provided there is an available basic probe for the corresponding dimension.

F. MuSCa in action

So far, this work is used through three aspects. First, MuSCa has been used to study different IoT scenarios in the INCOME[4] project. Then, MuSCa is used to specify and implement multiscale deployment requirements in a multiscale software deployment tool. At last, MuSCa is used to add multiscale requirements between data producers and consumers and generate the multiscale probes used in the implementation. This section presents a synthesis of these three aspects.

1) Multiscale IoT Scenarios analysis: In the INCOME project, in order to characterize the multiscale nature of several IoT scenarios, the MuSCa vocabulary has been used as a reading grid. The scenarios have been analyzed through different viewpoints and dimensions to highlight their revelant scales. This approach enabled the INCOME project members to compare a great variety of scenarios. This work helped to build the MuSCa taxonomy, which was based on the viewpoints, dimensions, measures and scales identified during the study of the scenarios.

2) Multiscale deployment: Deployment of software entities on devices in a multiscale system is another concern of the INCOME project. For this purpose, MuSCa has been used to define MuScADeL [13], a domain-specific language (DSL) dedicated to multiscale and autonomic software deployment. MuScADeL allows deployment designers to abstractly define deployment properties without exact knowledge of the devices and networks the system will be deployed on. MuSCa helps deployment designers to characterize the multiscale nature of a system from several viewpoints such as device, network, administration and geography. An example of the language (deployment of the motivating scenario) is presented in the Listing 4.

```
2   // Definition of probes
3   Probe Network {...}
4   MultiScaleProbe MSNetwork {..}
5   MultiScaleProbe Geography {..}
6   MultiScaleProbe Device {..}
7   // Definition of a criterion
8   BCriterion Crit50MB { Network.bandwidth > 50; }
9   // Definition of deployment requirements
10  Deployment {
11    // deployment of filter components
12    F @ Crit50Mb, Each MSNetwork.Type.LAN;
13    // deployment of hierarchical routing components
14    R @ Each Geography.Location.Building,
15        Geography.Location.City("Toulouse");
16    R @ Each Geography.Location.District,
17        Geography.Location.City("Toulouse");
18    R @ Geography.Location.City("Toulouse"),
19        Device.PowerProcessing.Cloud;
20    // deployment of end-user context-aware components
21    C @ All Device.Type.Smartphone,
22        Geography.Location.City("Toulouse");
23  }
```

Listing 4. Multiscale deployment constraints

As a MuScADeL code is linked to a MuSCa specific model, the MuScADeL editor can check that dimensions and scales conform to the ones defined in the MuSCa model associated with it. In addition, multiscale requirements are verified at runtime by the multiscale probes generated for this MuSCa model.

MuScADeL runs alongside MuSCa, as an Eclipse plugin, allowing the deployment designer to be able within the same

[2]http://www.eclipse.org/acceleo/

[3]http://www.eclipse.org/modeling/emf/

[4]http://anr-income.fr

engineering tool (Eclipse) to define new multiscale viewpoints, dimensions or scales, before using them in the deployment DSL.

3) Multiscale context data filtering: Concerning context data filtering, there are ongoing works in the INCOME project to extend context data routing and filtering with expressions that use the multiscale vocabulary. These scale-aware routing requirements can be added to context data producer and consumer contracts [14]. The aim of these works is to express privacy and quality of context constraint based on multiscale concerns —e.g., to get context data from parking places located at foot distance, or to share context data with user located in the same neighborhood. These constraints are defined from a model of the multiscale characterization of the deployed system, which gives to the contract designers a shared vocabulary restricted to the existing scales in the system.

Once the constraints are defined, they can be enforced thanks to the multiscale probes generated with the model-driven approach. These probes, which are generated from the same MuSCa model as the one used to define the constraints, can characterize a system entity at runtime in order to decide if this entity matches a constraint or not. For example, Fig. 7 illustrates the areas that match the context data filtering constraints presented in Section II-B —i.e., Sophie's neighborhood and foot distance scale instances. This figure was produced with the help of a generated geography multiscale probe, which has been implemented by calling the reverse geocoding API of the OpenStreetMap[5] Nominatim[6] service, and the map has been produced using the JMapViewer[7] API (also part of the OpenStreetMap project).

This example illustrates the interest of the MuSCa metamodel. The multiscale characterization of the system, which contains rules to associate system entities to scales, is expressed in a MuSCa model with a shared vocabulary, in a declarative way. The MDE approach allows to automatically generate the appropriate probes depending on the rules and scales declared in the MuSCa model. As the same MuSCa model is used to express context data routing constraints, the probes can be used to identify the areas that match these routing constraints.

V. Related works

We have noticed the rising presence of complex distributed systems in several recent research studies [15], [1], [2], [3], [16]. Some of these systems are explicitly described by their authors as multiscale [15], [3], [16]. However, we did not find any definition of the *multiscale* vision of distributed systems and when it should be applied. We propose a framework to characterize the multiscale nature of distributed systems.

To obtain a multiscale analysis of a complex distributed system, two methods may be applied. The first one is a bottom up approach, it studies at runtime real complex distributed system. Some recent works propose algorithms to study collaboration patterns of real complex systems. For example, in [5], the authors propose a community detection algorithm to find structures of communities to partition a complex system. These families of algorithms study the hierarchical structure of real complex systems. Due to the dynamic nature of complex systems, multiscale runtime analysis is unachievable. The second one is a top-down approach, it studies complex distributed systems at design time. New generations of approaches that allow developers of complex systems to work at a higher level of abstraction are needed [8]. Developing complex systems without such approaches increases the use of code-centric technologies, hinders developers from focusing on functional, non-functional and architectural needs, and requires herculean efforts. It produces hand-crafted systems, strongly coupled with technologies, that are neither maintainable nor upgradable. The use of MDE may help to describe systems at multiple levels of abstraction and from a variety of perspectives [17], [8]. MDE facilitates developers' work at design time, by providing specialized modeling languages, metamodels and code generation tools. Moreover, MDE can also be used at runtime to maintain a model of the running system. The interest of MDE is demonstrated in several research areas, as for context-aware pervasive systems [18], [19], runtime adaptation [20], or multi-cloud systems management [21]. Concerning the multiscale vision of complex systems, a multiscale UML profile for the deployment of complex systems is proposed in [22]. The model proposes three fixed scales: infrastructure, communication and deployed entities. In this approach, scales are rather views of a system. There is no distinction of levels in a given view —i.e., scales.

This study highlights the need for a top-down approach to study the multiscale nature of complex distributed systems. Model-driven approaches are interesting as they propose different levels of abstraction, and model transformations. It also identifies the lack of a shared multiscale taxonomy and scale-aware framework. We believe that the MuSCa framework fills these needs by helping system designers to build a multiscale vision of their complex systems. This framework is open : one

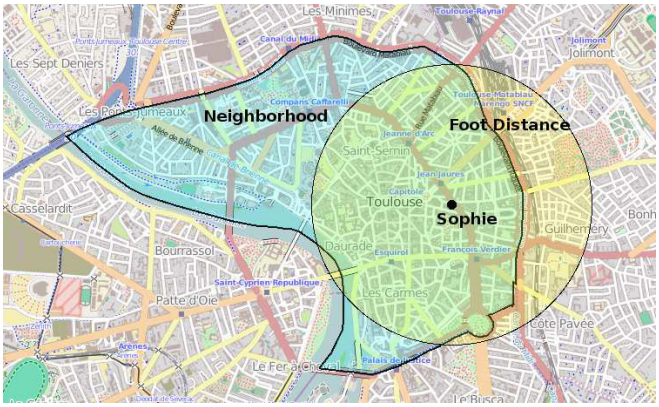


Fig. 7. Sophie's neighborhood and foot distance scale instances

[5]http://www.openstreetmap.org
[6]http://wiki.openstreetmap.org/wiki/Nominatim
[7]http://wiki.openstreetmap.org/wiki/JMapViewer

can add new viewpoints and associated dimensions, measures, scale sets and scales as needed.

VI. Conclusion

Multiscale distributed systems raise new kind of issues such as heterogeneity management, granularity variations, and distribution over the scales. This paper presented MuSCa, a framework to study multiscale nature of complex distributed systems and to provide them with scale-awareness capability.

The framework is based on a characterization process, which helps designers to study the multiscale nature of a given system. We have analyzed many scenarios and use cases for multiscale systems for the INCOME project. The characterization process enables the designer either to select—among existing ones—the viewpoints, dimensions and scales relevant for a given system, or to define new ones. For one characterization, MuSCa generates probes for runtime scale-awareness. Each characterization enables the framework to extend its multiscale taxonomy and its multiscale probes. Thus the framework learns and memorizes new viewpoints, dimensions and scales to be proposed for a later characterization.

MuSCa is helpful to build multiscale distributed systems that cope with some of the previously mentioned challenges. A good knowledge of the multiscale nature of a system contributes to choosing appropriate architectural patterns for each multiscale distributed system.

Currently, MuSCa is used the INCOME project. A DSL for software deployment was successfully designed using MuSCa. We also plan to use the scale-awareness capability to filter the distribution of the events in a distributed event based system.

Acknowledgments

This work is partly funded by the INCOME ANR project (ANR-11-INFR-009, 2012–2015) in which the following French partners are taking part: IRIT (Institut de Recherche en Informatique de Toulouse), Télécom SudParis, and ARTAL Technologies. The authors thank all the members of this project.

References

[1] M. Satyanarayanan, "Mobile computing: the next decade," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 15, no. 2, pp. 2–10, Aug. 2011. doi: 10.1145/2016598.2016600. [Online]. Available: http://dx.doi.org/10.1145/2016598.2016600

[2] M. van Steen, G. Pierre, and S. Voulgaris, "Challenges in very large distributed systems," *Journal of Internet Services and Applications*, vol. 3, no. 1, pp. 59–66, 2012. doi: 10.1007/s13174-011-0043-x. [Online]. Available: http://dx.doi.org/10.1007/s13174-011-0043-x

[3] G. Blair and P. Grace, "Emergent middleware: Tackling the interoperability problem," *Internet Computing, IEEE*, vol. 16, no. 1, pp. 78–82, Jan. 2012. doi: 10.1109/MIC.2012.7. [Online]. Available: http://dx.doi.org/10.1109/MIC.2012.7

[4] D. Chavalarias *et al.*, "French roadmap for complex systems 2008-2009," Mar. 2009. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00392486

[5] P. Pons and M. Latapy, "Post-processing hierarchical community structures: Quality improvements and multi-scale view," *Theoretical Computer Science*, vol. 412, no. 8- 10, pp. 892–900, Mar. 2011. doi: 10.1016/j.tcs.2010.11.041. [Online]. Available: http://dx.doi.org/10.1016/j.tcs.2010.11.041

[6] M. Satyanarayanan, "Scalable, secure, and highly available distributed file access," *IEEE Computer*, vol. 23, no. 5, pp. 9–18, May 1990. doi: 10.1109/2.53351. [Online]. Available: http://dx.doi.org/10.1109/2.53351

[7] H. Sandhu and S. Zhou, "Cluster-based file replication in large-scale distributed systems," in *Proceedings of the ACM Sigmetrics Performance '92 Conference*, ser. SIGMETRICS '92/PERFORMANCE '92. New York, NY, USA: ACM, May 1992. doi: 10.1145/133057.133092. ISBN 0-89791-507-0 pp. 91–102. [Online]. Available: http://dx.doi.org/10.1145/133057.133092

[8] R. France and B. Rumpe, "Model-driven development of complex software: A research roadmap," in *2007 Future of Software Engineering*, ser. FOSE'07. Washington, DC, USA: IEEE Computer Society, 2007. doi: 10.1109/FOSE.2007.14. ISBN 0-7695-2829-5 pp. 37–54. [Online]. Available: http://dx.doi.org/10.1109/FOSE.2007.14

[9] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for VM-based cloudlets in mobile computing," *IEEE Pervasive Computing*, vol. 8, pp. 14–23, Oct. 2009. doi: 10.1109/MPRV.2009.82. [Online]. Available: http://dx.doi.org/10.1109/MPRV.2009.82

[10] Object-Management-Group, "Software & Systems Process Engineering Metamodel (SPEM) v2.0," formal/2008-04-01, Apr. 2008.

[11] ISO/IEC/IEEE, "Systems and software engineering — architecture description," ISO/IEC/IEEE Joint Technical Committee, International Standard ISO/IEC/IEEE-42010:2011, Dec. 2011.

[12] J. Bézivin and O. Gerbé, "Towards a precise definition of the OMG/MDA framework," in *Proceedings. 16th Annual International Conference on Automated Software Engineering, 2001, (ASE 2001)*, Nov. 2001. doi: 10.1109/ASE.2001.989813. ISSN 1938-4300 pp. 273–280. [Online]. Available: http://dx.doi.org/10.1109/ASE.2001.989813

[13] R. Boujbel, S. Leriche, and J.-P. Arcangeli, "A DSL for multi-scale and autonomic software deployment," in *ICSEA 2013, The Eighth International Conference on Software Engineering Advances*, Oct. 2013, pp. 291–296.

[14] S. Machara Marquez, S. Chabridon, and C. Taconet, "Trust-based context contract models for the internet of things," in *Ubiquitous Intelligence and Computing, 2013 IEEE 10th International Conference on and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC)*, Vietri Sul Mare, Italy, Dec. 2013. doi: 10.1109/UIC-ATC.2013.73 pp. 557–562. [Online]. Available: http://dx.doi.org/10.1109/UIC-ATC.2013.73

[15] M. Kessis, C. Roncancio, and A. Lefebvre, "DASIMA: A flexible management middleware in multi-scale contexts," in *Proc. Sixth International Conference on Information Technology: New Generations, 2009. ITNG '09*, Apr. 2009. doi: 10.1109/ITNG.2009.338 pp. 1390–1396. [Online]. Available: http://dx.doi.org/10.1109/ITNG.2009.338

[16] J. P. Arcangeli *et al.*, "INCOME a multi-scale context management for the internet of things," in *Ambient Intelligence*, ser. Lecture Notes in Computer Science, F. Patern, B. Ruyter, P. Markopoulos, C. Santoro, E. Loenen, and K. Luyten, Eds. Springer Berlin Heidelberg, 2012, vol. 7683, pp. 338–347. ISBN 978-3-642-34897-6. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-34898-3_25

[17] D. C. Schmidt, "Guest editor's introduction: Model-driven engineering," *IEEE Computer*, vol. 39, no. 2, pp. 25–31, Feb. 2006. doi: 10.1109/MC.2006.58. [Online]. Available: http://dx.doi.org/10.1109/MC.2006.58

[18] E. Serral, P. Valderas, and V. Pelechano, "Towards the model driven development of context-aware pervasive systems," *Pervasive and Mobile Computing*, vol. 6, no. 2, pp. 254–280, Apr. 2010. doi: 10.1016/j.pmcj.2009.07.006. [Online]. Available: http://dx.doi.org/10.1016/j.pmcj.2009.07.006

[19] S. Chabridon, D. Conan, Z. Abid, and C. Taconet, "Building ubiquitous QoC-aware applications through model-driven software engineering," *Science of Computer Programming*, vol. 78, no. 10, pp. 1912–1929, Oct. 2013. doi: 10.1016/j.scico.2012.07.019. [Online]. Available: http://dx.doi.org/10.1016/j.scico.2012.07.019

[20] G. Blair, N. Bencomo, and R. France, "Models@ run.time," *Computer*, vol. 42, no. 10, pp. 22–27, Oct. 2009. doi: 10.1109/MC.2009.326. [Online]. Available: http://dx.doi.org/10.1109/MC.2009.326

[21] N. Ferry, A. Rossini, F. Chauvel, B. Morin, and A. Solberg, "Towards model-driven provisioning, deployment, monitoring, and adaptation of multi-cloud systems," in *Proceedings of the 2013 IEEE Sixth International Conference on Cloud Computing*, ser. CLOUD '13. Washington, DC, USA: IEEE Computer Society, Jun. 2013. doi: 10.1109/CLOUD.2013.133. ISBN 978-0-7695-5028-2 pp. 887–894. [Online]. Available: http://dx.doi.org/10.1109/CLOUD.2013.133

[22] A. Gassara, I. Bouassida Rodriguez, and M. Jmaiel, "Towards a multi-scale modeling for architectural deployment based on bigraphs," in *Software Architecture*, ser. Lecture Notes in Computer Science, K. Drira, Ed. Springer Berlin Heidelberg, 2013, vol. 7957, pp. 122–129. ISBN 978-3-642-39030-2. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-39031-9_11

Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 1667–1676
DOI: 10.15439/2014F244
ACSIS, Vol. 2

Efficient Description and Cache Performance in Aspect-Oriented User Interface Design

Tomas Cerny*, Miroslav Macik†, Michael J. Donahoo‡ and Jan Janousek§
*Computer Science, †Graphics and Interaction at Czech Technical University,
Charles square 13, Prague 2, Czech Republic, Email: {tomas.cerny}*, {macikmir}†@fel.cvut.cz
‡Computer Science at Baylor University, Waco, TX
One Bear Place #97356, , 76798-7356, USA, Email: jeff_donahoo@baylor.edu
§Theoretical Computer Science at Czech Technical University
Thakurova 9, Prague 6, Czech Republic, Email: jan.janousek@fit.cvut.cz

***Abstract*—Increasing demands on web user interface (UI) usability, adaptability, and dynamic behavior drives ever growing development and maintenance complexity. Conventional design approaches scale poorly with such rising complexity, resulting in rapidly increasing costs. Much of the complexity centers around data presentation and processing. Recent work greatly reduces such data complexity through the application of Aspect-Oriented UI (AOUI) design, which separates various UI concerns; however, rendering in conventional and even AOUI approaches fails to maintain this separation, often resulting in high repetitions of concern fragments due to tangling. Even worse, mixing of dynamic and immutable components greatly limits caching efficacy as each have differing lifetimes. We extend AOUI design to push down concern separation to rendering, which reduces description size, through repetition reduction, and enables separate caching of individual concerns. Our results show considerable size reduction of UI descriptions for data presentations, faster load times and extended caching capabilities.**

I. Introduction

ENTERPRISE web applications have became common for domestic and international business in the last two decades. Users expect that enterprise applications support various browsing devices and provide attractive, usable, and fast UIs. Usability and speed often work in contradiction for web applications. Features to enhance usability often increase application size, which slows responsiveness, particularly in low bandwidth network such as those used for mobile access.

Despite clear expectations from UIs of enterprise web applications, conventional design approaches struggle from multiple deficiencies. For example, the UI descriptions for data presentations must restate information [18] from lower layers of the application, in order to extend them. This brings the risk of mistype errors caused by inconsistencies. Furthermore, modifications to the application data definitions require manual changes to the UI. The complexity is mostly evident when the UI description uses Domain Specific Languages (DSLs) [26] with limited type-safety. Furthermore, conventional approaches realize multiple UI concerns tangled together in a single component [5]. This not only limits component readability, but mostly limits its reuse, since such the component is strongly specialized. This "multi-concern component solution" results from the inability of conventional approaches to capture different concerns separately [33]. Such inability is also evident in object-oriented (OO) design [20]. Providing UI for given data in two slightly different situations (e.g., normal and mobile version of a website) may require to implement two similar UI components that differ only in details [5], [24], but requiring their separate maintenance. The UI development efforts are apparent from research [5], [18], which shows that approximately 48% of application code and 50% of development time is devoted to implementing UIs. This percentage grows with UI abilities and context-specificity.

Next, consider UI delivery to remote clients. In most cases, the UI is expressed as HTML and streamed to clients over the Internet. Although, supplemented with immutable resources such as images, stylesheets or JavaScripts the description itself is provided as a single piece of information. Such a single block of information has limited caching options and its size might be extensive. As stated above, UI components that present application data tangle multiple concerns together. This concern mix is also evident in the HTML streamed to a client. For example, an HTML data form mixes together field presentation, form layout, data binding, field validation, etc. Client web browser interprets the delivered HTML to present the UI description to the user.

Aspect-oriented design for UIs [5] reduces information restatement and supports separation of UI concerns for components presenting data [24]. The reduction of restatement is addressed through automated code-inspection [19] that supplies information for transformation to the UI. The aspect-oriented transformation involves integration of various UI concerns. This approach works at runtime and thus considers both static and contextual information. Each data presentation is assembled on demand, based on a given data instance. Concerns are captured individually and integrated based on contextual conditions. Individual concerns can thus be reused across different presentations and data. The outcome is significant UI code reduction and an assurance of correlation between the data definition and the UI presentation, which eliminates errors introduced by human factor [5], [9].

This research was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS14/198/OHK3/3T/13.

978-83-60810-58-3/$25.00

Since it is possible to capture UI concerns separately at the server-side, they can be also delivered individually to the client. The benefit at the server-side is the increase of concern reuse and thus UI code reduction. The delegation of the UI component assembly to the client-side could considerably decrease the amount of transferred information.

Efficient client-side caching of "tangled" HTML is complicated or even impossible. When only a single concern changes, the entire fragment must be transmitted again. Individual delivery of concerns to clients addresses reduction of replicated information in UI descriptions and makes it possible to cache certain concerns at the client. In this paper, we apply aspect-oriented UI (AOUI) design and research the impact of split concern streaming on the UI load time, transmission size and content caching. Our empirical results show considerable reduction of the UI description size, and the ability to cache individual concerns, which significantly reduces page load times for repeated visits. We evaluate our work by comparing the proposed approach with the conventional approach with respect to transmission content size, load time and caching.

The remainder of this paper is organized as follows: Section 2 describes the background of user interface designs. Section 3 provides an overview of existing work. AOUI approach is presented in detail in Section 4. Section 5 introduces distributed version of the approach. Its evaluation is discussed in Section 6. The final section presents our conclusion and future work.

II. Background

Enterprise system architecture [15], [22], [13] divides responsibilities into layers. For example the Java Enterprise Edition (Java EE) specification [10] divides the application into persistence, business logic and presentation layers. Developers implement each layer using a General Purpose Language (GPL). It is common practice for the presentation layer to use component-based UI approaches [2], which may involve a DSL to better describe a UI; unfortunately, such languages have limited type safety. Each layer has well defined responsibilities, and provides mechanisms to capture certain concerns. A concern [14] [20] can be understood as a set of information, which has some effect on the source code. For example, consider the concerns of data persistence, UI presentation, security, etc. Even though each layer has defined responsibilities and captures given concerns, there exist concerns that do not fit into a single layer but instead cross-cuts multiple layers. These cross-cutting concerns are responsible for tangled source code [20], and GPL languages, including OO, do not have mechanisms to effectively handle them [20], [21] so as to provide readability, maintenance or centralization. The result is that an individual concern, spreads throughout the source code and cross-cuts other concerns. Common examples of such concerns are exception handling, logging, and security as illustrated at Figure 1.

Aspect-Oriented Programming (AOP) provides an effective solution to this problem. [20], [21]. AOP suggests that, in addition to GPL components, there exists an additional concept

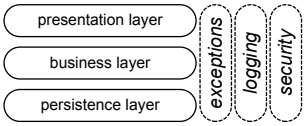


Fig. 1. Cross-cutting concerns in 3-layer enterprise system architecture

called an aspect. A particular program is then implemented using GPL and aspects. An aspect captures cross-cutting concerns separately from the GPL components. The way GPL components are connected with aspects is the main AOP contribution. An aspect consists of two parts: pointcut and advice. The pointcut specifies a situation, location or context under which the aspect is woven into the GPL component. The advice is the concern definition specified either in GPL or a custom DSL. In order to effectively address location in GPL components, AOP defines the concept of join-points. A join-point can range from code location specified by name or a wildcard, method invocation based on method name, annotation, or even a particular application context. Naturally, we can divide join-points into static and dynamic, with the difference based on whether they are activated only by location in the code or whether a runtime condition must hold to activate it. An example can be seen in enterprise systems when handing security with an AOP approach. When a user invokes an action from the UI, this action goes through an action controller [22], which has a method to implement the action. Often, this method has a security annotation determining user access, such as a user security role. Before the actual method is called, the security annotation acts as a join-point that activates a security aspect. This join-point is specified in the security aspect point-cut, and its advice looks into the application context to determine whether the actual user is logged into the system and whether he/she has sufficient permission, given by the security annotation, to be eligible to call the action. If not, the advice throws a security exception; otherwise it delegates the call to the controller method. The same security aspect applies in the UI to determine whether or not to render a given action button for a particular user, etc.

Tangled concerns can be found in the UI as well [5], [4]. Conventional design approaches do not address them separately but together in a single source code. This is directly responsible for low code reuse and readability as well as for high development and maintenance efforts. Various concerns that play a role in the UI can be considered independently as shown by Figure 2a. Unfortunately, because of limited GPL and DSL constructs, all concerns must collapse together into a single UI component, a single source code as depicted at Figure 2b. This results with tangling of all involved concerns, which makes individual concern localization difficult. Consider the concern collapse in Figure 2b at the sample code in Listing 1 while considering concerns from Figure 2a. Its graphical representation sketch is shown in Figure 3. The resulting source code is very specialized with limited reuse. At the same time, we must consider that such UI code restates information from the data definition [19], which introduces interdependencies that must be maintained. Because of limited

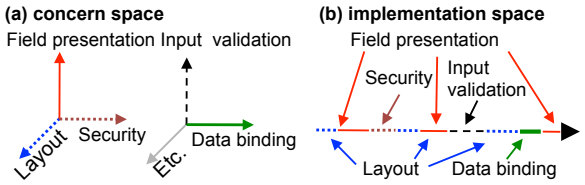


Fig. 2. (a) Concerns as orthogonal dimensions / (b) Implementation space in a single dimension with tangled concerns

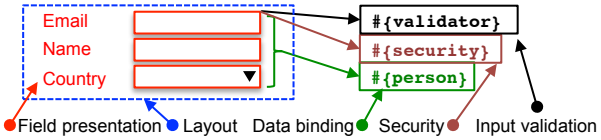


Fig. 3. Graphical sketch of Listing 1 denoting concerns from Figure 2

type safety in DSLs, it is easy to introduce an error while restating information [9]. The situation becomes worse when multiple presentations exist for given data [4]. The limited reuse forces us to maintain multiple, very similar components. Moreover, when we consider conventional approaches and aim to design adaptive or context-aware UIs, we end up with even more similar UI components that we must manage and update each time the data change [24].

The AOP approach for UIs [5] considers the data definition to be the GPL component that is being presented in the UI. In order to present data in UI, it considers individual UI concerns (e.g., Figure 2a) and weaves them together at runtime upon request to assemble the UI presentation. Such data presentation reflects the actual data definition and the particular application/user context. Data definitions, normally OO classes, define data fields and their constraints that act as static and dynamic join-points for the "data to UI" transformation. Dynamic join-points are further extended with the application context. A particular base presentation for each data field is selected based on these join-points. Subsequently, this base presentation is extended with field-level concerns through various aspects. Once all data fields have determined the presentation integrated with other concerns, the layout is woven through the fields. The resulting component reflects the data definition, context and considered concerns. Thus issues, such as information restatement or multi-component management, are eliminated. The result is that no physical UI component for data presentation exists; they are assembled on demand where each concern can be maintained and vary individually based on the user/application context. The advantage comes when we want to present novel data in the UI. All concerns are be reused and thus scaling-up the data model size does not impact the UI concern size or the UI management. What impact the size of the individual concern space are custom presentation layouts for given data, although it is possible to design generic set of layouts reusable among data.

When we consider client-server communication over HTTP, we must be aware that all the concern assembly to receive data presentation takes place at the server-side. Basically, at the server-side all concerns tangle together through an aspect-

Listing 1. Sample source code for UI form reflecting Figure 2 (b)

```
<table><tr>
 <td>Email:</td>
 <td><h:input id="email" value="#{person.email}"
     render="#{security.hasAccess('email')}"
   validate="#{validator.validate('email')}"/></td>
</tr><tr>
 <td>Name:</td>
 <td><h:input id="name" value="#{person.name}"/></td>
</tr><tr>
 <td>Country:</td>
 <td><a:smenu id="country" value="#{person.country}"/></td>
</tr></table>
```

weaver that produces UI descriptions, eventually translated to HTML. This tangling may produce large and complex HTML, containing repetitive information impacting the content size. The server transmission outcome of conventional approaches is similar, if not the same, to the outcome of AOP-based UIs. While compression as well as caching of static resources can be applied, there are two issues. First, compression, although addressing repeating patterns, is not aware of the content logical structure, and thus it addresses small repetitive fragments rather than large concerns. Second, it does not allow to cache immutable information occurring in the delivered UI description. In this work, we show that it is possible to improve the transmission content size and caching with AOP-based UIs. Since AOP has constructs to separate individual concerns, it is possible to stream separately to the client and let the client perform the assembly at the client-side. This reduces the repetitive information from the transmission and, at the same time, an individual decision on concern caching and reuse can be made. Such changes in the UI delivery impact the UI transmission and presumably reduce the delivery time.

III. Related Work

A. UI design approaches

Various approaches have been introduced to simplify development of complex UIs. These approaches can be divided into model-based, generation-based, inspection-based, and AOP-based. Each of these offers certain advantages for UI development; however, they typically fail to address UI maintenance or complex situations, such as context-based UIs adapting during the runtime. In terms of client-server communication, conventional approaches transfer large, perhaps unnecessary, amounts of data, which negatively impacts the communication and response times.

Model-driven development (MDD) [29] suggests that a model is the source of information, and the resulting source code is generated using this model together with a set of transformation rules. The main advantage should be reduction of information restatement that must be handled manually [9] for different perspectives. In [23] MDD is applied to distributed UIs, with description of a workflow that uses a task-centred approach described through the Concur Task Tree (CTT) notation [1]. This allows the description of environment and given context. MDD can handle multi-context UIs, for example, in [3] authors split the context into user, platform and environment parts. At the same time when we aim to

describe different concerns through multiple models, MDD does not provide any standard integration mechanisms to do so [31]. Sottet et al. [31], [32] provide a deep explanation of model-to-code and model-to-model transformations relevant to UI MDD. Although, MDD can be used to capture complex UIs, various contexts, and adaptive features, the model-to-code transformations may struggle in the performance perspective [24]. Transformations are usually performed at compile time as they tend to be time consuming [24]. Compile-time transformations produce source code and descriptions for all possible context states, which might not be fulfilled by the user [27]. Next, the transformation takes place at the server-side, and thus the result may contain tangled UI descriptions possibly containing repetition. The MDD-based UI design may struggle from further issues. In [27], the authors observe that it suffers during adaptation and evolution management. Such design handles base situations well, although when context variations or customizations are needed, the modification often take place in the UI code [9] rather than in the model. Manual changes are lost the next model-to-code transformation, which then leads to difficult maintenance [9]. Another issue, presented by [5], arises when the MDD applies solely to the UI but not to other parts of the system, such as persistence or business logic subsystems. In such cases, information captured by the model must match to information captured by the rest of the system. When only one part of the system changes, another part may lose compatibility and may need to address the changes manually. Such an approach is unfortunately very common in the research discipline of human-computer interaction [24].

The use of DSLs [26] for the UI model description is common, although, DSLs tend to provide weak type safety, which extends maintenance efforts, since information change propagation becomes tedious and error-prone in a manual process. DSLs are often used to directly specify UIs [26] [17]; a practical example is the Java EE standard for component-based development called JavaServer Faces (JSF) [2]. The DSL brings simplification to the UI description [5], as oppose to GPL. It is transformed to the target UI language, such as HTML. DSLs naturally fit to UI descriptions, but through their weak type-safety, it is easy to introduce errors related to restated information from lower application layers [18]. For example a DSL description may reference data, their fields and constraints that are already described in the application through GPL [12]; however, referencing a GPL component from DSL requires certain restatement with a negative impact on maintenance. Similarly to MDD, the DSL-to-native code transformation takes place at the server-side, thus the transmission streams the produced tangled code.

Another approach addresses information restatement by utilizing code-inspection and meta-programming [18], [5], [24]. It inspects data GPL definitions and from the result composes a structural model. This model is transformed to UI descriptions with all data/constraint references resolved through the model; this avoids human-errors related to inconsistencies. The output can be in the form of DSL, such as JSF. In comparison to the above approaches, this one works at runtime, although, it does not address cross-cutting UI concerns. Similarly, the product generated at the server-side is not different from the product produced through DSLs or MDD.

One possible solution that addresses tangled code and cross-cutting UI concerns is Generative Programming (GP) [11], [30], which emphasizes domain methods and integration with GPL. GP can be seen as programming that generates source code through generic code fragments or templates. The goal is to address gaps between program code and domain concepts, support reuse and adaptation, simplify management of component variants and increase efficiency. The generation, although, happens at compile time. The use of GP for UI is demonstrated at [30] through abstract UI specifications. An application that uses this concept consists of three parts: a DSL for UI description, configuration generator that automates the product UI assembly and an extensible collection of elementary components available for the assembly. The configuration generator takes the DSL specification and assembles implementation components from them and from the available components. Such an approach allows production of a large number of system variants for specific requirements. In a presented case study, a system combined two hundred features in the UI, with a resulting variability of 5×10^{17} prototypes. It is questionable whether such a large number of feature components is reasonable and could be ever used, although all states are physically generated at compile time and statically allocated. The nature of the compile-time assembly makes it hard for use with future adaptive systems that need runtime information to base its decisions on [27].

The AOP approach has been applied to extend capabilities of existing approaches. In [27] the authors apply AOP to MDD to support adaptive features at runtime. This work suggests that MDD approaches do not naturally fit into adaptive systems because they lack the runtime information, which should be considered to influence model-to-code transformation. As mentioned in [32] the MDD runtime transformation might be performance inefficient for complex situations [24]. Some suggest that MDD UI transformations may generate all possible application states and configurations for hypothetical/possible situations. In complex systems, this can grow exponentially. Also, MDD-based systems suffer and become impractical in evolution management of system adaptation. [27] thus suggests using four runtime models that represent main system data that are manipulated at runtime. These models are responsible for system runtime adaptations and generation of application components. The work describes the aspect-oriented conceptual model [33], weaving process and context very sparsely, and no performance consideration is given to the manifest approach effectivity for production systems. Alternative aspect-oriented UI design, based on conventional UI designs and enabling both code inspection and separation of concerns for data representations, is given at [5]. Similar to inspection-based approaches, meta-programming determines a structural model (join-point model). Subsequently an aspect-oriented transformation of the model to the UI enables integration of various, separately defined, UI concerns. Although, the

aspect weaving happens at runtime, it takes place at the server side and thus the UI transmission to clients is no different from the above approaches.

As shown above, existing research in UI fails to address effective UI transmission to clients or optimization of client-side caching. One of the contemporary UI frameworks, although, does address caching by compiling the GPL UI descriptions to client cacheable resources. The Google Web Toolkit (GWT) [16] suggests describing the UI in the type-safe Java language and compile it to a JavaScript (JS) UI description. Note well, that even GPL UI description consists of restated information from application lower layers, such as data field descriptions and their constraints. For example, to design a UI representation for a given data field, the developer must select an appropriate component, bind it with the field and manually restate the component constraints already defined at the field, through annotations [12]. As mentioned in [33] and [20], GPL languages do not effectively handle cross-cutting concerns; consequently, GWT tangles them together. The GWT produces a JS UI description at compile time, which is similar to the MDD approach. It produces multiple versions of those descriptions to support various end-devices. It consists of code fragments that can be cached as well as these that cannot. GWT struggles from the same disadvantages as MDD, and thus it does not fit to adaptive UIs. For instance a UI page that presents many context-based variations compiles all possible states to a single description and ships it to the client not matter whether the user uses a single UI state or multiple. This becomes obvious with complex adaptive systems [27] with combinations of states and configurations that grow exponentially. In our work, we suggest transmitting UI concerns separately to clients. This allows transmitting only the actual state needed by the client, and at the same time, each concern may change individually, avoiding exponential grow.

B. UI delivery to the client

The standard client-server communication for web systems is based on the HTTP protocol that provides the core mechanisms to improve the transmission. First, the TCP-based protocol supports connection persistence so multiple resources can be loaded from a single server. Connections are reused, rather than reopened, which requires additional overhead. Multiple connections may exist from a client to a single server. HTTP supports content compression to reduce the transmission content size. Furthermore, it supports content caching at the client-side with time-based invalidation. The caching applies mostly to static resources such as CSS, images, and JS. In [7] authors show that the average contemporary web systems consists of about 90% static and cacheable resources. To further reduce the transmission, UI developers may apply content obfuscation and resource merging [6]. To mitigate the impact of client distance, servers often apply geo-distributed caching of static resources called content-delivery networks (CDNs), such as Akamai [28].

Structured Hypertext Transfer Protocol (STTP) [34] extends HTTP to include new messages to control the resource transmission for a particular web page. A similar approach, HTTP-MPLEX [25], employs a header compression and response encoding scheme for HTTP. Similar to STTP, it multiplexes multiple responses to a single sustained stream of data to speed response times and improve application layer use of TCP. While experiments show performance improvements with these protocols, they do not consider resource distribution through CDNs, caching, or variations. Another optimization approach is brought by cooperative-web cache [7], [8]. It involves clients with cached resources in participation in an overlay peer-to-peer network, which allows clients to share these resources in the overlay. Unlike CDNs it supports natural scalability and free P2P services; however, it must deal with content invalidation in the overlay and mechanisms to disable and prevent malicious clients from sharing corrupted data.

IV. AOP-BASED UI ASSEMBLY

The development and design approach of aspect-oriented UIs (AOUI) [5] is considerably different from conventional approaches. In order to describe the AOUI design, we first illustrate the UI design with conventional approaches and describe its dependencies to data definitions. Next, we sketch the AOUI design differences and describe it genericity and relaxed dependencies.

Figure 4 demonstrates the conventional UI design with the data model at the top with individual data classes with fields and given constraints as well as application context. The bottom presents a sketch of various presentations of given data. These presentations are rendered in the application based on the context. Each presentation has physical code representation and consists of multiple elements that bind to a particular data field or its constraints. The UI presentation has to restate field names as well as the constraints in its source code, which increases UI component coupling. Each such component is specialized to display specific data, and we can hardly assume that such UI component could be used for another data class. Such coupling must be seen from the perspective of system maintenance, thus when a given data class changes, for example a new field occurs, we must manually reflect it in related UI components. When the UI component description uses DSL, limited type safety may not provide any enforcement mechanism on the compatibility with data. We can summarize the disadvantages as follows: no automated data change propagation to the UI, limited type safety does not prevent human errors, limited separation of concerns and limited reuse, data class presentation requires to design a custom UI component. To see the difference with the AOUI design, consider the case when we aim to render data in the UI. In order to do that, we need to know the physical location of a particular UI component (denoted via UI render start mark at Figure 4), and this components then uses its configuration to cooperate with the data class, and its fields based on matching names.

Next, consider the AOUI design in the same illustration of data classes and the application context at the top and data presentations at the bottom Figure 5. First, note that with AOUI

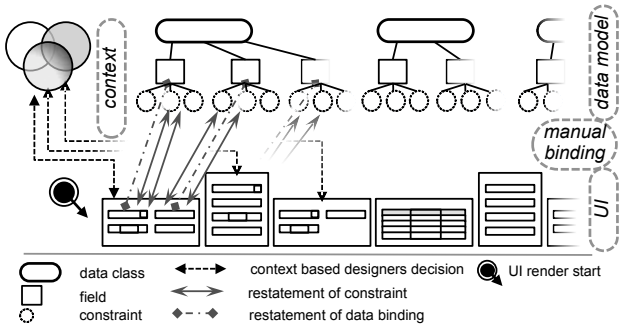


Fig. 4. UI design with conventional approach

there is no physical location of the UI presentation, because it does not exists. Instead it is generated on request by passing a data instance reference to an aspect weaver (denoted via start mark in Figure 5). The AOUI aspect weaver inspects the given instance data class, denoted by ①, and produces its structural (join point) model for given data. This model is created once upon the first use and consists of class and field information and their constraints. Upon each use, a clone model is made and modified/restricted according to user context. For example fields not relevant to given context in the UI are eliminated or restricted based on security rights. Additional elements from the application context can be exposed to this model as well. Each model element then acts as a join point for the subsequent transformation stages. In the next stage ②, the aspect weaver considers generic mapping rules that select a presentation for each particular data field based on its properties in the structural model. These rules are not specific for a given data class or a field; instead they only bind to model elements, which may occur at any data field. This brings genericity and reusability among fields and data classes. These rules become reusable among data classes or even among different systems. Each rule consists of two parts, a specification of structural model elements in a query (a point-cut) and an advice in the form of a presentation template. A rule applies to a field, which has a given constellation of specified elements. For example a query specifies a text-typed field with maximum length greater than 255 letters. The selected rule then applies an advice that selects a presentation template for the field matching the given query. This presentation template consists of a description for a basic field presentation in the target UI language. It also contains extensions in the description through which it is possible to integrate other aspects to it, providing concerns weaving. Each template aspect consists again of a point-cut that uses the same constructs as mapping rules referencing the structural model and advising how to integrate the concern. Often it embeds selected structural model element values to the output. In the stage ③, after all data fields have resolved presentation, a layout template is selected based on the context and integrated to the field presentation. This results with stage ④ that provides the presentation for given data instance and current application context and renders it to the UI. The most important benefit is that rules and presentation templates are generic and not dependent on specific data, which allows us to scale-up the

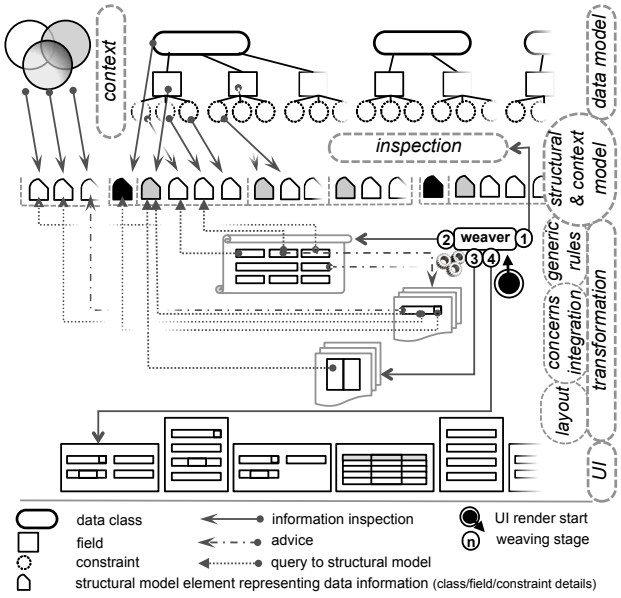


Fig. 5. UI design with AOUI

data model without an impact on the size or efforts related to the UI. In other words, a new data class passed to the AOUI aspect weaver is displayed based on the existing rules and templates. Data changes do not cause inconsistency in the UI because the change is reflected in the ad-hoc structural model that influences the selection of a given transformation rule as well as it subsequent aspect integration in the presentation template. Novel constraints apply to the UI according to the their occurrence in templates. In the [5] authors show that among 63 data classes in a production system with about 473 fields, only 28 transformation rules and presentation templates are needed. They are reused, and various concerns integrate to it based on given context. The benefits can be summarized as: Code volume reduction, constraint enforcement, separation of concerns, data independence, concern reuse, reduction of restatement. Furthermore, it is easy to integrate new concerns and thus support context-aware or adaptive UIs while not introducing complexity to the UI design.

V. Distributed AOP-based UI assembly

Conventional design approaches stream the UI description as a single block of information. The AOUI design untangles UI concerns for components reflecting data and reduces development and maintenance efforts. Such weaving takes place at the server-side and its product, with weaved-concerns, is streamed to clients. This is equivalent to conventional approaches. One may assume that content compression over HTTP solves the inefficiency, although with no doubt, it does not improve caching options. To the contrary, consider a solution where concerns (such as these from Figure 2 a) are streamed separately to clients. This might seem as an addition overhead as we need to handle multiple connections. On the other hand, this may eliminate repeating patterns in the transmitted content and enable caching for given immutable concerns.

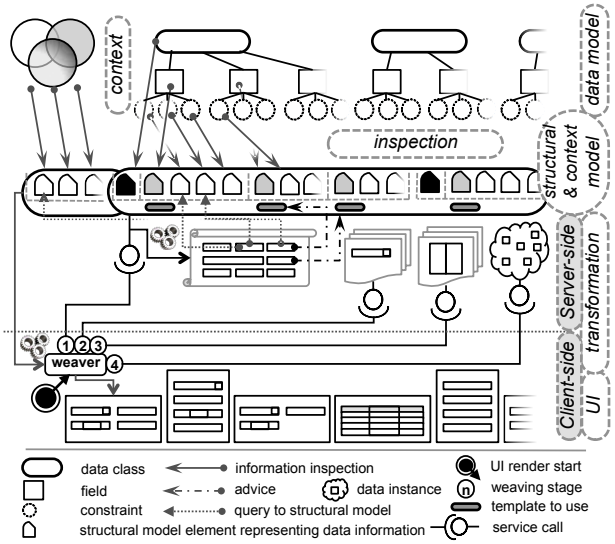


Fig. 6. UI design with dAOUI

In order to design distributed AOUI (dAOUI), the process of concern weaving should be partially pushed to the client-side. The application data model (or data transfer objects [15]) and application context are part of the server-side, and thus the inspection part takes place there. This gives a structural and context model, which could be streamed to the client. Some model elements might not be relevant to presentation or to a particular user. For example, there might be internal fields such as primary key, version field, etc., and these might not be relevant to the UI or to a particular user rights/tasks. AOUI handles this through data model constraints as well as through the context. To provide a particular user only the model elements relevant to his session and rights, the context is applied to the requested subset of structural model. This eliminates elements that would not apply for the UI composition, and this subset is streamed to the client. The selection of a particular presentation template for given data field could be done at the client-side. This increases the complexity of the client weaver since it must be aware of transformation rules and these may need to have access to internal server-side information to resolve rules to and make a decision. Thus keeping this responsibility at the server-side and providing the result to the client reduces the client weaver design complexity. Figure 6 depicts the responsibility assignments between server and client sides through service calls. Each client needs to have access to the structural and context model for the given data it represents in the UI ①. It also needs access to presentation templates ②. The decision on template selection made on the server-side is delivered together with the structural and context model ①. The client uses a particular presentation template suggested by server. Each template content is resolved towards the structural and context model of a given data field. Layouts ③, similar to presentation templates, are provided to the client-side for integration to the data UI presentation. Other concerns might be provided as separate services and integrated either at the server-side through transformation rules or via client-side presentation templates. Each client composes the UI data component based on provided concerns that are influenced by system context. The server also provides the actual data instances to the client ④. These data are displayed in the assembled UI component. The data submission uses HTTP POST or GET mechanisms or a web service.

The life-cycle for the web systems works as follows. First, the user requests navigation to a particular page or a dialog. This page consists of description elements from conventional UI design. The difference is for components representing data. Such components are replaced by custom tags interpreted at the client-side (for example a JS call). The tag indicates which data to display and what settings (context) should apply for the UI component assembly. Such content, with no data physical representation in it, is transmitted to the client-side. When client interprets the delivered content, it interprets it ordinarily. Custom tags are interpreted through a client-side weaver that requests given concerns from the server-side. Provided responses consider user rights and security. As depicted at Figure 6, the weaver assembles the UI representation of a data instance (given by the custom tag) conforms the structural model, application context and settings provided together with the data reference. The weaver may either request a particular concern from the server-side or may reuse it from its cache. For example, presentation templates will hardly change throughout a long period of time, or given data fields are immutable in given context over the time or throughout a user session/conversation.

VI. Evaluation

In order to evaluate our approach, we consider existing production level enterprise web application based on Java EE 6 platform with JSF [2] framework for the UI design. For our evaluation, we consider a subsystem for user account management. We evaluate the existing solution regarding data transmission, page load time and caching. Next, we implement the same subsystem with dAOUI design. Specifically, we design REST services at the server-side and a JS library responsible for UI component assembly, interacting with these services. These services include a service to obtain structural and context model for given data, a service to obtain the actual data from a given instance conforming system security and finally a service to handle data manipulation from the client-side. Next, we provide a JS package with presentation and layout templates and the client-side aspect-weaver. The dAOUI prototype is evaluated using the same criteria.

Figure 7 at outer left shows a sample UI subsystem considered in the evaluation. The same result can be designed with both conventional and AOUI approaches [5]. The difference relates solely to the server-side design; the content transmitted to the client-side is equivalent for both approaches. The dAOUI prototype is shown at inner right of Figure 7. The differences are that the conventional UI uses JSF, which is transformed to HTML through the framework and transmitted to the client-side all together. The second UI consists of JSF for the page without components representing data. Instead it uses a JS library that interacts with the server REST services

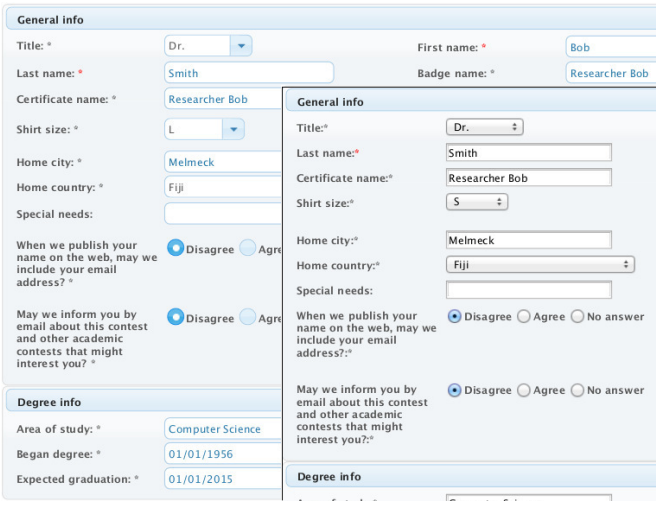


Fig. 7. Evaluated UI subsystem designed with conventional/AOUI approach in the outer left position, the dAOUI approach in the inner right position

TABLE I
BASE EVALUATION CASE, TRANSMISSION OF UIS WITH 23 FIELDS

No network throttling	No-cache		Cached	
	Size (KB)			
		Compressed		Compressed
Convent. approach	1458	329	86	11.1
Distrib. AOUI	1386	311	3.9	2.1
	Load time (using compression)			
	(sec)	(relative)	(sec)	(relative)
Convent. approach	2.3	1	1.67	1
Distrib. AOUI	1.73	0.75	0.79	0.47

TABLE II
EXTENDED EVALUATION CASE, TRANSMISSION OF UIS WITH 42 FIELDS

No network throttling	No-cache		Cached	
	Size (KB)			
		Compressed		Compressed
Convent. approach	1484	331	110	13.9
Distrib. AOUI	1394	315	6.9	3.8
	Load time (using compression)			
	(sec)	(relative)	(sec)	(relative)
Convent. approach	2.54	1	1.99	1
Distrib. AOUI	1.89	0.74	1.01	0.51

through JSON format to assemble the data presentation components and embeds them to the UI. In both cases UI panels, controllers and page navigation logic are equivalent.

A. Network transmission and page load times: The base case

To evaluate network transmission size and page load times, we use the UIs described above. In the evaluation of page loads, we consider complete page rendering. We must emphasize that both UIs utilize equivalent static web resources (JS, CSS and images). These same resources are part of the production system, so with the outcome we can consider a realistic impact on a production system. The dAOUI version has additional JS libraries, but on the other hand, the initial HTML page does not contain descriptions for data components. Instead, it consists of a JS initiation calls to assemble data components based on settings given in the HTML page. The considered data components at the page at Figure 7 consist of 23 fields of various data types given by the production system.

The conventional approach page produces 1458 KB to render the UI, although with HTTP compression the transmitted content reduces to 329 compressed KB (cKB). The main HTML document is 86 KB (11.1 cKB), and the rest 1372 KB are static resources. To download and render the UI page with compression takes 2.3 sec (average over 10 samples with standard deviation $\sigma = 0.23$). The download uses gzip compression with no network restrictions. The dAOUI page produces 1368 KB (311 cKB). The main HTML document has the size only 2.9 KB (1.2 cKB); additionally there is 10 KB (3 cKB) of JS and four calls to REST services with a total size of 13.6 KB (4.9 cKB). The page load reduces to 1.73 sec ($\sigma = 0.14$), an almost 0.6 second (or 25%) time reduction.

Next, we consider caching. All static resources are cached at the client-side. The conventional approach page with cached resources requires 86 KB (11.1 cKB) to be loaded from the server, and the page load time reduces to 1.67 sec ($\sigma = 0.24$). The dAOUI allows us to cache the weaver, presentation and layout templates, as well as the data structure, which is immutable. This reduces the transmitted content down to 3.9 KB (2.1 cKB), and the page load time needs only 0.79sec ($\sigma = 0.07$), representing reduction of almost 0.9 sec, which is less than 50% of the original wait time. The summary can be seen from Table I.

B. Increasing the UI size: The larger case

The production system on which our study is based has the person account page considerably larger than what we considered above. Next, we consider the impact related to data size extension. The extended UI has 42 fields at the page.

The conventional approach page extends to 1484 KB (331 cKB) out of which 110 KB is the HTML document (13.9 cKB). The page load time is 2.54 sec ($\sigma = 0.33$). The dAOUI size has in total 1394 KB with the HTML document 5 KB (1.4 cKB) and JSON calls 19.3 KB (8.4 cKB). The compressed transmission size has 315 cKB. The page load time for dAOUI is 1.89 sec ($\sigma = 0.17$), representing a reduction of 0.65 sec, which is similar to the previous evaluation with reduction to less than 75% compared to the conventional approach.

The cached-enable evaluation of the conventional approach design consists of the total transmitted size of 110 KB (13.9 cKB) with load time 1.99 sec ($\sigma = 0.38$). The cached dAOUI has 6.9 KB (3.8 cKB) and a load time of 1.01 sec ($\sigma = 0.15$). Similar to the previous evaluation, the reduction of wait time is almost 1 second and represents almost 50% of the load time. The summary can be seen in Table II.

C. Throttling the network: 3G/DSL users case

The next evaluation throttles the network conditions to evaluate behavior for both mobile users with a 3G network and DSL users. For the 3G evaluation, we restrict the network bandwidth to 384 kbit/s and set network delay to 20ms. Such network restrictions allow us to emulate network conditions for users with mobile devices. We evaluate the base page with 23 fields. The load time significantly grows to a barely usable system. The page using the conventional approach requires

TABLE III
3G AND DSL CASE, TRANSMISSION OF UIS WITH 23 FIELDS

	Load time (using compression)			
	(sec)	(relative)	(sec)	(relative)
384kbit/s 20ms delay	*No-cache*		*Cached*	
Convent. approach (sec)	18.89	1	2.72	1
Distrib. AOUI (relative)	15.88	0.84	1.07	0.39
768kbit/s 10ms delay	*No-cache*		*Cached*	
Convent. approach (sec)	9.45	1	1.79	1
Distrib. AOUI (relative)	8.28	0.88	1	0.49

TABLE IV
GWT COMPARISON, TRANSMISSION OF UIS WITH 23 FIELDS

	No-cache	**Size (KB)**		*Cached*
		Compressed		Compressed
GWT	379	102	11.9	5.8
Distrib. AOUI	161	41.8	3.9	2.1

18.89 sec to load ($\sigma = 2.16$). The dAOUI requires still a long time 15.88 sec ($\sigma = 0.57$). The reduction represents 3 sec, which is around 85% of the original load time. Caching becomes a "must have" for these kinds of mobile users. With caching, the conventional page loads within 2.7 sec ($\sigma = 0.42$) compare to the dAOUI with 1.07 sec ($\sigma = 0.19$). This represents a reduction of 1.6 sec, which is less than 40% of the original load time. The summary is at the top of Table III.

Next, we evaluate the situation for DSL users with bandwidth to 768kbit/s and delay 10ms. The conventional page needs 9.45 sec to load ($\sigma = 0.84$). The dAOUI reduces the load time by 1.1 sec, down to 8.28 sec ($\sigma = 0.36$). The reduction in percentage is similar to the 3G version and represents slightly less than 85% of the original load time.

Caching significantly helps for consequent page loads. The conventional page loads within 1.79 sec ($\sigma = 0.24$) compared to the dAOUI with 0.87 sec ($\sigma = 0.08$). The dAOUI is 0.9 sec faster and less than 50% of the load time. The summary can be seen at the bottom of Table III.

D. Comparison with GWT

The GWT introduced in the related work targets improvements to UI caching. To compare it with our approach, we implemented the 23-field prototype application with GWT. Note well, that the evaluated dAOUI page at Figure 7 is based on a production system, while the GWT is just a prototype. This results with differences in both prototypes regarding linked JS libraries. While the dAOUI prototype links JS resources related to given JSF component provider, the GWT prototype does not link to any generic JS library. Although, this does not impact the caching statistics, we modify the dAOUI prototype as follows: We reduce the linked JS libraries and only consider libraries related to the functionality of the dAOUI, which makes it equivalent regarding the comparison with the GWT prototype. The dAOUI prototype needs to transmit 161 KB of data to build the UI at the client (41.8 cKB). The GWT version needs 379 KB (102 cKB). The main document in GWT is converted to into JS with cacheable fractions with 141KB (50 cKB) and non-cacheable fraction with 7.2 KB (3.4 cKB).

The cached-enabled evaluation stays the same for dAOUI with transmitted content 3.9 KB (2.1 cKB). The GWT version needs to download the HTML page, displayed data and the non-cacheable JS fragment, which is in total 11.9 KB (5.8 cKB). The results are summarized in Table IV. From the results, we see that UI construction in the form of JS can considerably improve caching at the client-side. The JS presents tangled code through mixed concern, which extends its size. Separately streamed UI concerns can reduce the UI description, and improve caching.

E. Summary

Streaming various concerns separately from server to clients brings reduction to the overall transmitted content, page load times (considering complete UI rendering) and improved caching. In the evaluation, we streamed presentation and layout templates, data structure with applied security as well as the actual data. The dAOUI enables to use cache for concerns that are normally tangled together in conventional approaches. The study on a production system shows the reduction of content size in the range of tens of KBs even when compressed. Even though the dynamic content represents only around 6% (3.5% compressed) of the total content, the transmission size of uncached UI content reduced in total by 5%. With caching that strips the static content, it reduces transmission by 72-81% compare to JSF. With GWT the transmission content with cached resources improves by around 63%. The dAOUI managed to reduce the page load time to the range of 75-90% of what it takes with a conventional approach. This turns even better with caching, which gives reduction in the range of 40-50% compared to the conventional approach. The exact reduction is, although, influenced by many factors including network conditions and the UI itself.

In our evaluation, we reduced the transmitted content sometimes by an order of magnitude compare to the JSF approach. While the GWT approach compiles the UI into a JS and provides a solution with extended caching capabilities, such a solution can be further improved with dAOUI.

VII. CONCLUSION

In this paper, we suggest an alternative design approach for presentations of data in enterprise software systems. Conventional designs mix various concerns together, which results in code that is hard to maintain and reuse. We show that some of the UI concerns do not change over the time, and we may cache them at the client-side to reduce network traffic. Conventional designs fail to offer this ability, so given concern might be tangled with others without the possibly of variability and reuse. AOUI design separates concerns for UI components presenting data and thus reduces maintenance efforts as well as improves reuse of these concerns. With an extension of such design, it is possible to deliver concerns separately over the network to clients and delegate the component assembly to the client-side. This may reduce the transmitted content size that needs to be delivered to the client over the network, but mostly it enables caching and reuse of specific UI concerns at the client-side. From a case study based on a subset of a production system, we provide results showing a reduction of the total transmission size from the server to the client

and the page load time. Furthermore, we extend capabilities for caching of UI concerns at the client-side, which further reduces the data transmission as well as page load times.

Although the results from the study are promising, we must consider its limitations. The design approach fits well for data presentations; it builds on the top of other approaches that deal with interaction, page-flow, etc. AOUI easily adapts to development standards and allows integration of third parties for security, context-awareness, etc. The approach does not aim for complete design of UI pages but targets only data presentations. Existing approaches provide a large palette of various field components, suggestion boxes and data manipulations. With this approach, it is necessary to design them as no component library exists. On the other hand, it is easy to integrate HTML5 components or custom presentations for fat-clients. The interaction is not limited to only frontend and backend parts of systems, but it possible to consider middleware communication. Considering AOP's natural ability of concern separation, the design fits well to context-based UIs. At the same time, the AOP development pushes towards different development practices, and designer transition from conventional habits might be difficult. In addition there is a lack of tool and framework support as well as a missing standard for AOUI.

In future work, we aim to extend concerns with business rules integration. Our preliminary work [4] shows that it has large potential. We also look at integration of aspect-oriented design to service oriented architecture (SOA).

REFERENCES

[1] S. Berti, F. Correani, G. Mori, F. Paternò, and C. Santoro. Teresa: a transformation-based environment for designing and developing multi-device interfaces. In *CHI'04 extended abstracts on Human factors in computing systems*, pages 793–794. ACM, 2004. http://dx.doi.org/10.1145/985921.985939.

[2] E. Burns and N. Griffin. *JavaServer Faces 2.0, The Complete Reference*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 2010.

[3] G. Calvary, J. Coutaz, D. Thevenin, Q. Limbourg, L. Bouillon, and J. Vanderdonckt. A unifying reference framework for multi-target user interfaces. *Interacting with Computers*, 15(3):289–308, 2003. http://dx.doi.org/10.1016/S0953-5438(03)00010-9.

[4] K. Cemus and T. Cerny. Aspect-driven design of information systems. In *SOFSEM 2014: Theory and Practice of Computer Science, LNCS 8327*, volume 8327, pages 174–186. Springer International Publishing Switzerland, 2014. http://dx.doi.org/10.1007/978-3-319-04298-5_16.

[5] T. Cerny, K. Cemus, M. J. Donahoo, and E. Song. Aspect-driven, data-reflective and context-aware user interfaces design. *Applied Computing Review*, 13(4):53–65, 2013. http://dx.doi.org/10.1145/2513228.2513278.

[6] T. Cerny and M. J. Donahoo. Performance optimization for enterprise web applications through remote client simulation. In *Proc. of the 7th EUROSIM Congress on Modelling and Simulation, Prague, CZ*, volume 2. CTU, Prague, 2010.

[7] T. Cerny, P. Praus, S. Jaromerska, L. Matl, and J. Donahoo. Cooperative web cache. In *Systems, Signals and Image Processing (IWSSIP), 2011 18th International Conference on*, pages 1–4. IEEE, 2011.

[8] T. Černý, P. Praus, S. Jaroměřská, L. Matl, and M. Donahoo. Towards a smart, self-scaling cooperative web cache. *SOFSEM 2012: Theory and Practice of Computer Science*, pages 443–455, 2012. http://dx.doi.org/10.1007/978-3-642-27660-6_36.

[9] T. Cerny and E. Song. Model-driven Rich Form Generation. *Information: An International Interdisciplinary Journal*, 15(7, SI):2695–2714, JUL 2012.

[10] R. Chinnici and B. Shannon. JSR 316: Javatm platform, enterprise edition 6 (java ee 6) specification, Dec 2009.

[11] K. Czarnecki and U. W. Eisenecker. Components and generative programming. In *Proc. of the 7th European software engineering conference held jointly with the 7th ACM SIGSOFT intl. symposium on Foundations of software engineering*, ESEC/FSE-7, pages 2–19, London, UK, UK, 1999. Springer-Verlag. http://dx.doi.org/10.1145/318774.318779.

[12] L. DeMichiel. JSR 317: JavaTM persistence API, version 2.0, Nov 2009.

[13] L. DeMichiel and M. Keith. JSR 220: Enterprise javabeans version 3.0. java persistence API, May 2006.

[14] E. W. Dijkstra. *A Discipline of Programming*. Prentice Hall, Inc., 1976.

[15] M. Fowler. *Patterns of Enterprise Application Architecture*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2002.

[16] R. Hanson and A. Tacy. *GWT in Action: Easy Ajax with the Google Web Toolkit*. Manning Publications Co., Greenwich, CT, USA, 2007.

[17] M. Karu. A textual domain specific language for user interface modelling. In T. Sobh and K. Elleithy, editors, *Emerging Trends in Computing, Informatics, Systems Sciences, and Engineering*, volume 151 of *Lecture Notes in Electrical Engineering*, pages 985–996. Springer New York, 2013. http://dx.doi.org/10.1007/978-1-4614-3558-7_84.

[18] R. Kennard, E. Edmonds, and J. Leaney. Separation anxiety: stresses of developing a modern day separable user interface. In *Proc. of the 2nd conf. on Human System Interactions*, HSI'09, pages 225–232, Piscataway, NJ, USA, 2009. IEEE Press. http://dx.doi.org/10.1109/HSI.2009.5090983.

[19] R. Kennard and J. Leaney. Towards a general purpose architecture for ui generation. *Journal of Systems and Software*, 83(10):1896 – 1906, 2010. http://dx.doi.org/10.1016/j.jss.2010.05.079.

[20] G. Kiczales, J. Irwin, J. Lamping, J.-M. Loingtier, C. V. Lopes, C. Maeda, and A. Mendhekar. Aspect-oriented programming. In *ECOOP'97-Object-Oriented Programming, 11th European Conf.*, volume 1241, pages 220–242. Springer, June 1997. dx.doi.org/10.1007/BFb0053381.

[21] R. Laddad. *AspectJ in Action: Enterprise AOP with Spring Applications*. Manning Publications Co., Greenwich, CT, USA, 2nd edition, 2009.

[22] C. Larman. *Applying UML and Patterns: An Introduction to Object-Oriented Analysis and Design and the Unified Process*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition, 2001.

[23] K. Luyten, C. Vandervelpen, J. V. den Bergh, and K. Coninx. Context-sensitive user interfaces for ambient environments: Design, development and deployment. In *Mobile Computing and Ambient Intelligence: The Challenge of Multimedia*, Dagstuhl, Germany, 2005.

[24] M. Macik, T. Cerny, and P. Slavik. Context-sensitive, cross-platform user interface generation. *Journal on Multimodal User Interfaces*, pages 1–13, 2014. http://dx.doi.org/10.1007/s12193-013-0141-0.

[25] R. L. R. Mattson and S. Ghosh. HTTP-MPLEX: An enhanced hypertext transfer protocol and its performance evaluation. *J. Netw. Comput. Appl.*, 32(4):925–939, 2009. http://dx.doi.org/10.1016/j.jnca.2008.10.001.

[26] M. Mernik, J. Heering, and A. M. Sloane. When and how to develop domain-specific languages. *ACM Comput. Surv.*, 37(4):316–344, Dec. 2005. http://dx.doi.org/10.1145/1118890.1118892.

[27] B. Morin, O. Barais, J.-M. Jezequel, F. Fleurey, and A. Solberg. Models@ run.time to support dynamic adaptation. *Computer*, 42(10):44–51, Oct. 2009. http://dx.doi.org/10.1109/MC.2009.327.

[28] E. Nygren, R. K. Sitaraman, and J. Sun. The akamai network: A platform for high-performance internet applications. *SIGOPS Oper. Syst. Rev.*, 44(3):2–19, Aug. 2010. http://dx.doi.org/10.1145/1842733.1842736.

[29] J.-l. Perez-medina, S. Dupuy-chessa, and A. Front. A survey of model driven engineering tools for user interface design. In *Proc. of 6th Int. workshop on Task Models & Diagrams (TAMODIA'2007)*, pages 84–97, Berlin, 7-9 Nov. 2007. Springer. dx.doi.org/10.1007/978-3-540-77222-4_8.

[30] M. Schlee and J. Vanderdonckt. Generative programming of graphical user interfaces. In *Proceedings of the working conference on Advanced visual interfaces*, AVI '04, pages 403–406, New York, NY, USA, 2004. ACM. http://dx.doi.org/10.1145/989863.989936.

[31] J.-S. Sottet, G. Calvary, J. Coutaz, and J.-M. Favre. A model-driven engineering approach for the usability of plastic user interfaces. In *Engineering Interactive Systems*, pages 140–157. Springer, 2008. dx.doi.org/10.1007/978-3-540-92698-6_9.

[32] J.-S. Sottet, G. Calvary, and J.-M. Favre. Models at runtime for sustaining user interface plasticity. In *Models@ run. time workshop (in conjunction with MoDELS/UML 2006 conference)*, 2006.

[33] M. Stoerzer and S. Hanenberg. A classification of pointcut language constructs. In *Workshop on Software-engineering Properties of Languages and Aspect Technologies (SPLAT) held in conjunction with AOSD*, 2005.

[34] B. Swen. Outline of initial design of the structured hypertext transfer protocol. *J. Comput. Sci. Technol.*, 18(3):287–298, 2003. http://dx.doi.org/10.1007/BF02948898.

Author Index